DEMOGRAPHIC RESEARCH

A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 51, ARTICLE 41, PAGES 1299–1350
PUBLISHED 27 NOVEMBER 2024
https://www.demographic-research.org/Volumes/Vol51/41
DOI: 10.4054/DemRes.2024.51.41

Research Material

Using online genealogical data for demographic research: An empirical examination of the FamiLinx database

Andrea Colasurdo

Riccardo Omenti

Contents

Demographic Research: Volume 51, Article 41
Research Material

Using online genealogical data for demographic research: An empirical examination of the FamiLinx database

Andrea Colasurdo[1]

Riccardo Omenti[2]

Abstract

BACKGROUND
Online genealogies are promising data sources for demographic research, but their limitations are understudied. This paper takes a critical approach to evaluating the potential strengths and weaknesses of using online genealogical data for population studies. We focus on the FamiLinx dataset, which contains demographic information and kinship ties across multiple countries and centuries.

OBJECTIVE
We propose novel measures to assess the completeness and the quality of demographic variables in the FamiLinx data at both the individual and the familial level over the 1600–1900 period. Utilizing Sweden as a test country, we investigate how the age–sex distribution and the mortality levels of the digital population extracted from FamiLinx diverge from the registered population.

METHOD
We employ descriptive statistics, negative binomial regression modeling, and standard life table techniques for our measures of completeness and quality.

RESULTS
Missing values and accuracy in demographic information from FamiLinx are selective. When one demographic variable is available, researchers can effectively anticipate the availability of other demographic information. The completeness and quality of demographic variables within kinship networks are markedly higher for individuals with more complete and accurate demographic information. Populations from FamiLinx display lower mortality levels than the registered population and their representativeness improves towards the end of the 19th century.

[1] Kinship Inequalities Research Group, Max Planck Institute for Demographic Research, Rostock, Germany; Population Research Centre, Faculty of Spatial Sciences, University of Groningen, Groningen, the Netherlands. Email: colasurdo@demogr.mpg.de.
[2] Department of Statistical Sciences, University of Bologna, Bologna, Italy.
Email: riccardo.omenti2@unibo.it.

CONTRIBUTION
This study sheds new light on the opportunities and challenges of harnessing online genealogies for demographic research. Although this data source offers much promise, its usability in population studies is dependent on the quality and completeness of its recorded demographic information and their selectivity.

1. Introduction

The digital revolution has provided researchers with access to an unprecedented wealth of non-traditional data sources that can be used in population studies (Cesare et al. 2018; Kashyap 2021). Among these emerging sources, online genealogical data have garnered significant attention (Blanc 2021, 2024; Corti, Minardi, and Barban 2024; Cozzani et al. 2023; Gavrilova and Gavrilov 2007; Hsu et al. 2021; Minardi, Corti, and Barban 2024; Stelter and Alburez-Gutierrez 2022). These data sources present themselves as vast repositories of information from genealogical websites that enable users to reconstruct their own family trees. Online genealogical data are micro-level data that (a) are scraped from digital family tree information stored in genealogical websites, (b) link individuals not only to their parents but also to more distant relatives, and (c) provide additional detail on the demographic characteristics of individuals, such as their dates and locations of birth and death (Song and Campbell 2017).

Although online genealogical data were not primarily designed for use in social science research, they hold significant potential. First, they serve as exclusive repositories of data on extended family networks that span multiple centuries and cross-national borders. These data allow researchers to link individuals not only to their parents but also to their more distant ancestors. Additionally, the kinship structure of these data sources enables researchers to examine multi-generational processes, and thus to go beyond the traditional two-generation approach that primarily focuses on parent–offspring associations (Mare 2011; Song and Campbell 2017). Second, the large volume of demographic information in these data sources, including details about birth and death locations and dates, permits researchers to investigate long-term population dynamics in regions and historical periods for which official population data may be scarce or unavailable (Stelter and Alburez-Gutierrez 2022).

The use of genealogies in demography has emerged in response to the lack of historical data on the demographic experiences of kin groups (Post et al. 1997). Scholars have turned to genealogical data to advance the field of historical demography, to analyze historical trends in key demographic behaviors over time, and to study past mortality patterns and the influence of heredity and family dynamics (Otterstrom and Bunker 2013;

Zhao 2001). Louis Henry is recognized as the pioneer of family reconstitution. His work, which identified genealogies as rich sources for demographic research, has enabled researchers to pose a broader range of questions about family history, and to trace ancestors and more distant kin further back in time (Henry 1968; Hollingsworth 1976; Wrigley 1981). Early genealogies and existing historical studies relying on genealogical data mainly focus on the ancestors and descendants of specific social groups living in specific areas (Henry 1968; Otterstrom and Bunker 2013). Furthermore, most genealogical reconstitution efforts suffer from incomplete location specificity and family networks (Kasakoff and Adams 1995; Post et al. 1997). More recently, thanks to the digital revolution, genealogies have become powerful resources for tracing multiple generations of relatives over time using online platforms (Otterstrom and Bunker 2013).

While their inherent structure makes the use of online genealogical data in population studies appealing, we should be critical of the claims that have been made about their merits. We contend that a thorough explanation of their limitations, including issues of data quality and potential bias, is imperative to ensure the responsible use of these data in population studies. The presence of individuals in a genealogy typically hinges on genealogists' knowledge of their relatives or their decisions about whom to include in their family trees (Calderón Bernal, Alburez-Gutierrez, and Zagheni 2023). Hence, these databases generally overrepresent the family networks of individuals who experienced more favorable demographic conditions than the general population, including higher fertility, lower mortality, and higher nuptiality (Zhao 2001). Conversely, individuals with matrilineal and extinct lineages are often neglected. In genealogies, certain subpopulations are consistently underrepresented, including children who died at an early age and childless women. Online genealogies are also affected by selective remembering and the inclusion of individuals in online genealogical trees is contingent upon having a living descendant interested in tracing their family history (Chong et al. 2022; Cozzani et al. 2023; Minardi, Corti, and Barban 2023; Zhao 2001). Genealogy users are more inclined to remember ancestors with important roles in their family history (Chong et al. 2022) and may tend to downplay relatives who dishonored the family (Zhao 2001). These problems combine to create considerable demographic selectivity issues, including the underestimation of mortality and the overestimation of fertility (Calderón Bernal, Alburez-Gutierrez, and Zagheni 2023). At the same time, the underreporting of individuals dying at young ages might result in an underestimation of fertility levels (Calderón Bernal, Alburez-Gutierrez, and Zagheni 2023; Hollingsworth 1976). When genealogies exhibit inadequate coverage and representativeness, particularly when recording only a few generations or closer relatives, biases become more pronounced, consequently reducing the accuracy of estimations (Calderón Bernal, Alburez-Gutierrez, and Zagheni 2023; Zhao 2001).

Additionally, online genealogical data may suffer from a high prevalence of missing values for essential demographic variables, such as birth and death locations and dates. This is not unexpected, as users of genealogical websites are more focused on tracing their ancestors than on meticulously recording precise information about the locations and dates of their relatives' vital events. Limited familiarity with one's own relatives may also contribute to imprecise or missing information. Furthermore, certain subpopulations within genealogies are typically underrepresented and more likely to feature missing information. Examples include children who died at a young age and childless women. Genealogies often commence with a patriarch documenting his family history, with women typically recorded solely as wives or daughters, resulting in their information being less comprehensive than that of males (Zhao 2001). In light of these issues, we argue that a comprehensive examination of missing value patterns in crowdsourced genealogies is warranted.

Despite the previously mentioned limitations of crowdsourced genealogical databases (as shown by Stelter and Alburez-Gutierrez 2022; Calderón Bernal, Alburez-Gutierrez, and Zagheni 2023; Chong et al. 2022), the majority of population studies relying on these data sources have operated under the assumption that their selected individual samples accurately mirror the broader population (Blanc 2021, 2024; Cozzani et al. 2023; Hsu et al. 2021; Minardi, Corti, and Barban 2024). Prior research has attempted to illustrate the biases stemming from ascending genealogies and their impact on crucial demographic measures, such as life expectancy at birth and the total fertility rate, by means of simulations (see Calderón Bernal, Alburez-Gutierrez, and Zagheni 2023; Zhao 2001). To the best of our knowledge, our study represents the first attempt to evaluate the accuracy and the completeness of demographic variables at both the individual and the family network level in a big genealogical digital database, and to analyze the implications for the use of this database in population studies. To illustrate how the quality of the reported demographic information can vary, we look at the age structure of the population drawn from the genealogical data and a key demographic measure, life expectancy. Our aim is to offer scholars a more comprehensive understanding of the dataset's strengths and limitations, thus enabling them to make more informed decisions when utilizing the FamiLinx data for their research projects.

In this paper, our objective is to investigate the challenges associated with missing information in extensive genealogical data, and to highlight the critical issues that may hinder the usability of these data for demographic research. Specifically, we assess the accuracy and comprehensiveness of vital demographic variables in online genealogies regarding individuals and their associated family networks, including birth and death dates and locations. In our analysis we rely on the concepts of completeness and quality. Completeness refers to the quantification of the percentage of non-missing values for common demographic variables, while quality indicates the accuracy of the reported demographic information. Further details on the measurement of completeness and quality are provided in the method section. Although we focus on the FamiLinx database, we believe our findings and methods are also applicable to other genealogical databases.

In a nutshell, this paper seeks to address the following research questions:

1. What are the potential advantages and pitfalls of using online genealogies for demographic research?
2. How do the completeness and the quality of the demographic information in online genealogical data affect their usability? Are completeness and quality clustered within selected kinship networks?
3. How are the age–sex distributions and the demographic estimates derived from online genealogical populations impacted by the completeness and the quality of the reported demographic information?

2. Data

The analysis relies on the FamiLinx database, which is sourced using publicly available genealogies accessible on the geni.com website. These digital data are derived from family trees that have been constructed by a network of users from multiple countries with a common interest in tracing their own ancestral lineages. Since these genealogies have been built using a bottom-up approach, they are of the ascending type. This means that the genealogist begins the construction of their family tree from the bottom and then traces their lineage backward in time, including their parents, grandparents, great-grandparents, and so on. This process allows for the creation of a family tree that 'ascends' through the generations, illustrating the kinship ties between individuals when moving from present relatives to earlier ancestors.

Furthermore, FamiLinx has a passive registration system where only the main vital events are recorded, i.e., births and deaths, and the genealogists are not aware of the individuals' status at all the time points. This limitation hampers the applicability of FamiLinx to examine more complex demographic phenomena, such as migration trends and marriage patterns.

In recent years, scholars have increasingly turned to FamiLinx for population research. Leveraging the dataset's rich information spanning numerous centuries, FamiLinx has primarily served as a tool to investigate historical demographic trends. Existing studies have predominantly delved into historical mortality dynamics (Chong et al. 2022; Cozzani et al. 2023; Minardi, Corti, and Barban 2024; Pojman et al. 2023; Stelter and Alburez-Gutierrez 2022), scrutinizing the dataset's biases and representativeness compared to more reliable data sources (Chong et al. 2022; Stelter and

Alburez-Gutierrez 2022), or examining disparities in lifespan (Cozzani et al. 2023; Minardi, Corti, and Barban 2024; Pojman et al. 2023; Stelter and Alburez-Gutierrez 2022). Other research initiatives utilizing FamiLinx have centered on historical fertility patterns (Blanc 2021, 2024; Gay, Gobbi, and Goñi 2023) and the correlation between fertility and longevity (Hsu et al. 2021). Blanc (2024) additionally utilizes the dataset to uncover patterns of internal migration to and from urban centers. Overall, FamiLinx has emerged as a valuable resource for analyzing pivotal historical processes such as demographic transitions, shedding light on the potential of online genealogies in population research while acknowledging their inherent limitations in terms of bias and representativeness.

Our focus on FamiLinx derives from its easy accessibility, which makes it appealing to researchers. All the dataset's records are anonymized, and no formal request to access the information is needed. Additionally, FamiLinx covers more countries than other genealogies and provides quite detailed information about the location of events, surpassing the limited geographic scope of traditional genealogies. The demographic information stored in FamiLinx spans multiple generations of individuals, and thus covers a long period of time. Among the database's strengths is the ease with which the individual profiles can be linked to their family networks, thus facilitating a more comprehensive tracing of both ancestors and collateral kin.

The dataset was curated by Kaplanis et al. (2018), who gathered an extensive collection of 86 million profiles from the geni.com website. This social media platform allows users to upload their family trees and establish individual profiles for each member of their familial network. FamiLinx includes a dataset containing anonymized individual-level records for all 86 million individuals, as well as a dataset with information about the kinship ties between children and parents for approximately 43 million of these individuals. By leveraging these two types of records, researchers can identify distinct types of kin beyond parents and children, such as siblings and grandparents. Additionally, Kaplanis et al. (2018) eliminate implausible kinship ties; that is, individuals with more than two parents. The task of linking the two datasets is made easier by the fact that each individual is assigned a unique identification number. Specifically, the dataset with all the individual-level records incorporates vital demographic variables, including birth and death dates and locations, as well as gender. Each demographic variable is represented by multiple columns. For instance, the demographic information about the dates is presented in separate columns for day, month, and year. The locations of demographic events are documented through a two-digit country code representing the country name of the vital event, and through the country name itself reported as a string of text. Building on the information contained in the location-based text strings and two-digit country codes, Kaplanis et al. (2018) assigned latitude and longitude coordinates to profiles with sufficiently detailed information on the locations of vital events.

Since all the individuals who were still alive as of 2015, when the profiles were scraped from geni.com (see Kaplanis et al. (2018) for details), were omitted from the database, the demographic analysis should be restricted to individuals from extinct cohorts (see the appendix of Kaplanis et al. (2018) for details). Additionally, since the records in the database are anonymized, it is not feasible to link them to other micro-level historical data sources, such as parish records or censuses. Moreover, de-anonymization is not allowed under the terms of use of the data.

2.1 Analytical sample

We investigate how the completeness and the quality of the data are manifested within family networks.

As the individuals in genealogies are embedded in kinship networks, we believe that it is essential to investigate how the quality and the completeness of the demographic information on individuals in the genealogies are related to those of their kin. To facilitate our analysis, we define a subsample comprising approximately seven million ‘focal’ (or anchor) individuals, which we refer to as the ‘analytical sample.’

Based on our selection, we recall that individuals with identifiable kinship networks are inherently a subset of a larger population. To be included in the analytical subsample, individuals must a) have at least one parent or one child, as this ensures that the size of the kinship network of the focal individual is non-zero; and b) have at least one known place of birth or death, as determined by the following criteria.

2.2 Determination of birth and death locations

The locations of the demographic events experienced by focal individuals was determined by a three-tier method, which involved three algorithms ranked in order of preference:

a) Exact matching using the country code: Birth and death locations are inferred from the reported two-digit country code.
b) Regular expression matching: Birth and death locations are determined by a set of text strings, known as regular expressions, that specify a matching pattern for the name of the country of interest.
c) Inferred coordinates: This method leverages the latitude and longitude coordinates by Kaplanis et al. (2018) to identify the country of the vital event.

The motivation behind the implementation of this approach is that the inferred latitude and longitude coordinates by Kaplanis et al. (2018) may be affected by reporting errors due to historical changes in boundaries between countries.

To establish the definitive birth and death locations, we extract the country names from inferred coordinates harnessing a geo-parsing algorithm available in the R package *mapdata* (see Becker, Wilks, and Brownrigg (2022) for the details). We identify the 20 countries with the highest numbers of vital events.

Subsequently, we select the birth and death countries using the country codes and text strings from the 20 countries according to the methods described above. For instance, if a profile has two different birth locations, one determined by exact matching and the other based on the inferred coordinates, we assign the birth country identified by the exact matching method. Extending our analyses beyond these 20 countries would not affect our results, given the extremely low numbers of reported birth and death events in the remaining countries.

3. Methods

Our analysis consists of four steps. In the first step, we measure the completeness and the quality of the FamiLinx data. In the second step, we model the association between focal individuals and their kin in terms of the completeness and the quality of the demographic information. In the third step, we aim to generate population pyramids and age–sex distributions to evaluate the representativeness of populations drawn from online genealogical data. In the fourth step, we calculate life expectancy at age 30 based on the FamiLinx data. The methods applied in each of these steps are outlined.

3.1 Measurement of completeness and quality in FamiLinx

Our analysis relies on two pivotal concepts that determine the usefulness of FamiLinx for population studies: completeness and quality. These two concepts are investigated by considering the five main demographic variables present in the dataset: gender, and birth and death dates and locations.

Following the guidelines laid out by the United Nations (United Nations 2016), we measure completeness as the extent to which primary demographic variables (birth and death years and countries) display non-missing values. Specifically, this concept is quantified as the proportion of individuals with non-missing values for each of the aforementioned demographic variables. After the completeness of each demographic variable has been computed, we can analyze the variation in the marginal distributions of these variables when one of them is non-missing. Through this approach we are able to gain novel insights into the overall level of completeness of individual records in FamiLinx.

The concept of quality refers to the accuracy of the reported birth and death dates. Following the guidelines established by Kaplanis et al. (2018) and Minardi, Corti, and Barban (2024), we consider an individual record to exhibit higher quality if the month of the birth and/or death date is not missing. To measure the quality of the dates, we rely on the concept of year heaping. By year heaping we mean a preference for recording years with a last digit that is either 0 or 5 (see Stockwell and Wicks (1974) for a review on year heaping measurement[3]). Depending on whether we are considering births or deaths, we use the terms ‘birth year heaping’ or ‘death year heaping.’ When a sample has year heaping issues, it typically displays a non-uniform distribution of the number of births and deaths with unrealistic spikes in years ending in 0 or 5. Therefore, to examine the quality of the data in FamiLinx, we can only consider individuals with non-missing birth and death years. For this purpose, we define an indicator measuring the level of year heaping in the data. This indicator is calculated separately for the birth events and the death events in the data. Our strategy involves partitioning the selected individuals into two groups: one consisting of individuals with the non-missing month of the vital event and the other consisting of individuals for whom only the year of the vital event is available. Following this primary division, we group these individuals into 25-year intervals, and calculate the proportion whose reported year ends in 0 or 5. If the proportion in a group is close to 20%, we assume that there is no year heaping. Conversely, if it exceeds this threshold value, we conclude that there are year heaping issues in the data. In our example, if the proportion of individuals with the non-missing month of the vital event is around 20%, we can conclude that the demographic information for this group is relatively accurate (see Spoorenberg and Dutreuilh (2007) for a review of age heaping measurement). In our examination of the quality of the data, we restrict our analysis to individuals who were born or died between 1600 and 1900. We do so because the records of individuals with a birth or a death recorded before 1600 are considered unreliable (Kaplanis et al. 2018), while the cohorts born after 1900 might include individuals who were still alive as of 2015, which could result in ascertainment bias.

[3] A similar measurement concept for year heaping was employed by Cummins (2017) to assess the accuracy of the birth and death dates when analyzing the lifespan of Western European nobles from 800 to 1800.

3.2 Measurement of completeness and quality within kinship networks

After examining the completeness and the quality in the whole dataset, we explore how these concepts are applicable within the extended family networks (which include grandchildren, children, siblings, cousins, parents, aunts and uncles, and grandparents). Since researchers may be interested in examining the size and the structure of kin at any time point (Post et al. 1997) or investigating the multigenerational transmission of demographic behaviors, we believe that it is crucial to investigate the quality and the completeness of demographic information not only for focal individuals but also for their kin. This investigation could provide novel insights that are of value to researchers interested in using FamiLinx to carry out studies in the domains of historical and family demography.

To carry out this analysis, we rely on the individuals in the analytical sample and their respective kinship networks. We consider the demographic variables of birth and death countries and years. We disregard gender in the set of demographic variables due to the high percentage of non-missing values for this variable in the dataset.

To study the association between the completeness of demographic information for a focal individual and that of their kinship network, we use a negative binomial regression model. This model can be seen as a generalization of the Poisson regression model (Hilbe 2011). In both models the interpretation of the regression coefficients remains the same. However, in the negative binomial regression model the equi-dispersion assumption is not required, in that the mean of the response does not need to be equal to its variance. Hence, the modeling approach is appropriate given the over-dispersion present in the data. Concerning our model, for each combination of demographic variable *j* and type of relative *k*, we outline the following negative binomial regression model.

$$Y_{ijk}^{completeness} \,|\alpha_{ojk}\,, \alpha_{1jk}, \theta_{jk} \sim NegBinom\left(\mu_{ijk}, \theta_{jk}\right) \tag{1}$$

$$\begin{aligned} E\left(Y_{ijk}^{completeness} \mid \alpha_{oij}\,, \alpha_{1ij}, \theta_{jk}, \varphi_{jk}\right) &= \mu_{ijk} \\ &= exp\left(\alpha_{ojk} + \alpha_{1jk} z_{ij}^{completeness} + \varphi_{jk} c_{ik}\right) \end{aligned} \tag{2}$$

$$VAR\left(Y_{ijk}^{completeness} \mid \alpha_{oij}\,, \alpha_{1ij}, \theta_{jk}, \varphi_{jk}\right) = \mu_{ijk} + \frac{\mu_{ijk}^2}{\theta_{jk}} \tag{3}$$

where the dependent variable $Y_{ijk}^{completeness}$ denotes the number of relatives of type *k* of the focal individual *i* with a non-missing value in the demographic information *j* and the independent variable $z_{ij}^{completeness}$ indicates whether focal *i* has demographic information j non-missing. The parameter μ_{ijk} denotes the mean of the dependent variable and can be interpreted as the expected number of relatives of type k with non-missing values in demographic information j for a focal individual i. The parameter θ_{jk} is the reciprocal dispersion parameter and is included to account for overdispersion in the response variable. c_{ik}denotes the number of relatives of type k of the focal individual i.[4]

To evaluate the association between the focal individual and the quality of their kinship network's demographic information, we implement a negative binomial regression model for each type of relative. For the implementation of this set of negative binomial regression models, we selectively include only focal individuals and their kin conditional on possessing non-missing birth/death years. We aim to assess whether the dates of vital events reported for the relatives of a focal individual are more likely to be accurate when the month of the event for the focal individual is known.

The examination of the quality of the reported dates is based on the following multivariate negative binomial model.

$$Y_{ijk}^{quality}|\gamma_{ojk},\gamma_{1jk},\phi_{jk},\beta_{jk}\sim NegBinom(\eta_{ijk},\phi_{jk}) \tag{4}$$

$$\begin{aligned} E\left(Y_{ijk}^{quality}|\,\alpha_{ojk},\alpha_{1jk},\phi_{jk},\beta_{jk}\right) &= \eta_{ijk} \\ &= exp(\gamma_{ojk} + \gamma_{1jk}z_{ij}^{quality} + \beta'_{jk}X_i + \delta_{jk}d_{ijk}) \end{aligned} \tag{5}$$

$$VAR\left(Y_{ijk}^{quality}|\,\gamma_{ojk},\gamma_{1jk},\delta_{jk},\beta_{jk}\right) = \eta_{ijk} + \frac{\eta_{ijk}^2}{\phi_{jk}} \tag{6}$$

where the independent covariate $z_{ij}^{quality}$ denotes whether the month of the demographic event j experienced by the focal individual i is non-missing, and the dependent variable $Y_{ijk}^{quality}$ indicates the number of relatives of type k of the focal individual i with a non-missing value in the month of the date for the demographic event j. The parameter η_{ijk} is the expected value of the outcome variable and is interpretable as the expected number of relatives of type k with non-missing month in demographic information j for a focal individual i. The parameter ϕ_{jk} retains the same meaning as the parameter θ_{jk} in the previous model. X_i denotes a matrix of fixed effects made up of dummies referring to the period in which the demographic event of interest occurred. We believe that we should account for fixed effects, since the degree of heterogeneity in the quality of the reported demographic information may be higher for individuals with vital events in

[4] We included the number of relatives of type *k* as a control variable, since having a higher number of relatives of a certain type may increase the probability of having a larger number of relatives with a non-missing value in a demographic variable.

earlier historical periods. d_{ijk} indicates the number of relatives of type k of the focal individual i with the non-missing year of the demographic event j.[5]

To advance our understanding of the representativeness of digital populations drawn from online genealogies, we compare the age–sex distribution extracted from genealogical data with that of the registered population. In this analysis we identify two samples with distinct quality levels. One sample consists only of individuals with non-missing birth and death months, while the other is made up of individuals with missing birth or death months. This allows us to examine the impact of different sample selections on the age–sex distribution of the genealogical populations. To carry out this comparison we employ population pyramids, which enables us to visually investigate the extent to which the digital population drawn from online genealogies aligns with the registered population. In addition, we calculate the differences between the genealogy-based age–sex percentages and those based on census data for the same time period.

Finally, we leverage data from online genealogical populations to compute life expectancy at age 30. We aim to compute the demographic estimates from samples with distinct quality levels. This is again motivated by our interest in examining the impact of sample selection in online genealogical populations on the estimation of common demographic indicators, such as life expectancy at age 30. The previous measure is calculated using life tables with mortality rates smoothed over both ages and years. This calculation allows us to examine the ability of online genealogical data to capture historical trends in adult mortality. We smooth our estimates to avoid unrealistic shocks in life expectancy trends due to the small sample sizes. The smoothing is carried out utilizing two-dimensional P-splines implemented through the R package *mortalitysmooth* developed by Camarda (2012).[6] For more details on the mortality smoothing and its implementation in R, see the Appendix.

[5] We added the number of relatives of type *k* as a control variable, since having a higher number of relatives with a non-missing birth or death year may increase the probability of having a higher number of relatives with a non-missing birth or death month.

[6] *mortalitysmooth* allows computing smoothed mortality rates and their standard errors by age and sex via regression matrices of B-splines coupled with smoothing parameters (see Camarda (2012) and Eilers, Currie, and Durban (2006) for the methodological details). Life expectancy estimates are calculated using standard life table relationships, while the standard errors needed for the confidence intervals are computed using Monte Carlo simulations (see Mooney (1997)).

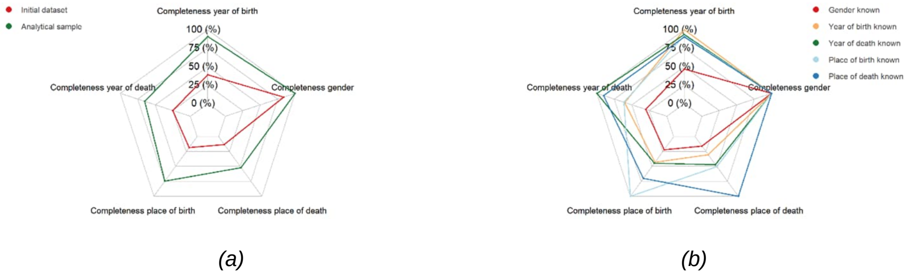
4. Results

4.1 Completeness of individual demographic information in FamiLinx data

Figure 1 presents the percentage of individuals with non-missing information for the considered demographic variables (gender, and birth and death dates and locations) to describe their availability in the initial dataset and in other subsamples (selected from the initial dataset). The characteristics of the initial full dataset and the subsamples are shown in Table A-2 in the Appendix. The radar charts (Figure 1) show that in the initial full dataset, most of the observations have missing information for the considered demographic variables, but the presence of at least one available variable considerably reduces the likelihood that other demographic variables are unavailable. The latter condition includes the analytical sample used for this study and several samples of observations, conditioned on having a specific demographic variable available. In the initial full dataset the year of birth, the year of death, the birth location, and the location of death are available for only 25% of the individuals, or even less. However, in the analytical sample the percentage of observations with available demographic information is larger. In particular, while the availability of gender information does not guarantee that other demographic information is available, knowing an individual’s place of death increases the probability of having non-missing information for the other variables. Thus, when information on one variable is available, researchers can effectively expect that other demographic information is also available, which contributes to a more comprehensive understanding of individual profiles in FamiLinx.

When we look at the completeness of the demographic information for individuals born in specific countries (Canada, Germany, Sweden, United Kingdom, United States of America) (see Figure A-1 in the Appendix), we see that the percentage of individuals with non-missing information on the selected demographic variables is much higher than that observed in the initial full dataset. In general, the individuals in the genealogies who were born in the United Kingdom have more incomplete demographic information, and indeed have the highest percentage of missing information for all the considered variables. While the percentages are quite similar for the other analyzed countries, individuals born in the United States seem to have a larger share of non-missing values for the demographic variables, especially those concerning the date and place of death.

Figure 1: (a) Percentage of non-missing values for five demographic variables in the initial full dataset (N = 86,124,644) and in the analytical subsample (N = 7,618,651). (b) Percentage of non-missing values for five demographic variables in different samples, identified by the availability of specific information



(a) (b)

Note: Each color indicates a different sample, and each colored line connects the percentages of non-missing information in each of the five variables considered. 'Analytical Sample' refers to the subsample of about seven million observations on which we perform the analysis. 'Gender Known' indicates the sample of individuals with non-missing gender information. Similarly, 'Place of Birth Known' indicates the sample of individuals with non-missing place-of-birth information, and so on.

4.2 Completeness of demographic information within kinship networks

The negative binomial models reveal a positive association between the completeness of the demographic variables for the focal individuals and their kin, independent of the size of the kinship network (see Table A-3 in the Appendix). This means that the presence of more complete variables for a focal individual is associated with having a higher number of relatives with more complete demographic information. These associations are found across distinct types of relatives and all the considered demographic variables, albeit with heterogeneous degrees of magnitude. Among all the demographic variables, the strongest association is observed for the birth year. As a robustness check, we ran a logistic regression model using as the response a binary variable equal to 1 if at least one of the relatives of a given type for a focal individual has a non-missing value in a demographic variable. As an additional sensitivity check we implemented two other regression models: a negative binomial regression model, where the number of relatives is treated as offset, and a binomial regression model. The results of the alternative models, included in the Appendix (Table A-5, Table A-7, Table A-9), are consistent with those of the negative binomial model presented in the main text.

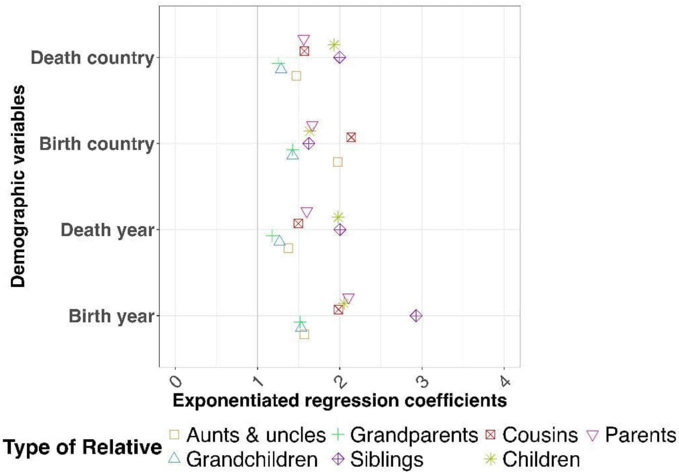
Regarding the specific types of relatives, horizontal kin, namely cousins and siblings, tend to exhibit stronger associations for all demographic variables. The associations are weaker for more distant kin such as grandparents. For instance, the expected number of siblings with a non-missing birth year for a focal individual with non-missing birth year is over three times bigger than that of a focal individual with a missing birth year. The expected number of children, cousins, and parents with non-missing birth year for the same focal individual is approximately twice higher than that of a focal individual with missing birth year. If we focus on the number of grandparents and aunts and uncles with non-missing birth year for a focal with non-missing birth year, their expected number is more than 50% higher than that of a focal individual with missing birth year.

The expected number of siblings, parents, children, and cousins with non-missing values in the death year and birth and death countries increases by at least 50% for a focal individual with non-missing values in the same demographic variables. For more distant kin, such as grandparents, grandchildren, and aunts and uncles, these estimates are still above the unit but are smaller in magnitude.

The observed differences in magnitude can be attributed to the greater proximity between the year of demographic events experienced by focal individuals and those experienced by their horizontal kin. When considering more distant kin, the temporal gap between the demographic events widens. Hence, for genealogists willing to reconstruct their own family trees, knowing the year of a demographic event experienced by the focal individual increases the likelihood of recollecting the same piece of demographic information for relatives who lived in the same temporal period; e.g., by searching in parish records. Conversely, gathering demographic information for more distant kin proves challenging due not only to the higher temporal distance between the demographic events but also to a more substantial effort to link the focal individual to their more distant relatives.

Overall, these results underscore how the completeness of demographic information tends to be shared among relatives. A focal individual with more complete demographic information has a higher likelihood of being embedded in a kinship network whose members have more complete demographic variables. This finding highlights the potential for studying demographic outcomes (fertility, longevity, etc.) within extended kinship networks beyond the classic parents–focal or children–focal relationships. Consequently, it opens up new opportunities for the exploration of demographic dynamics in the context of extended kinship networks.

Figure 2: Exponentiated coefficients from negative binomial regression measuring the association between a focal individual and their relatives in terms of the completeness of the reported demographic variables



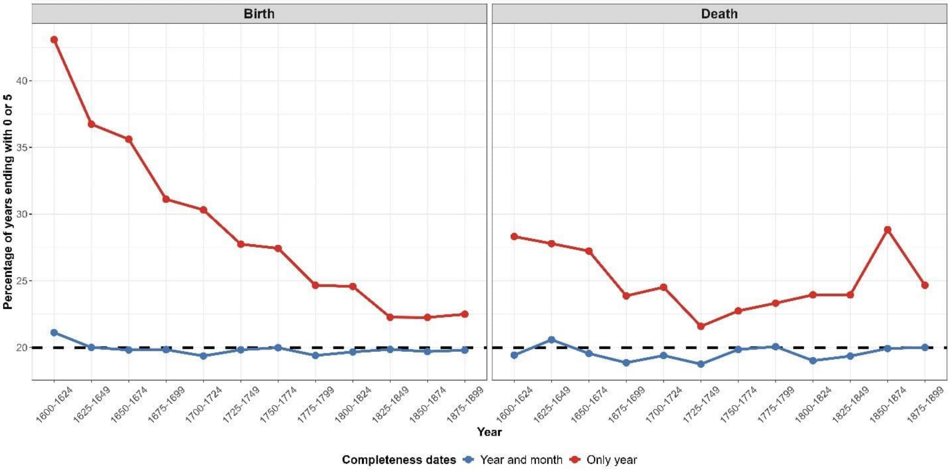
Note: The shapes indicate the ratio of the expected number of relatives with non-missing value for a demographic variable for a focal individual with a non-missing value to the same expected number for a focal individual with a missing value in the same demographic variable. Estimates larger than 1 denote that the expected number of relatives with a non-missing value for a demographic variable is higher for a focal individual with a non-missing value in the same demographic variable compared to one with a missing value. The results are reported by kin type and demographic variable. The distinct shapes and colors in the plot refer to different types of relatives. Confidence intervals are not included, as the large sample sizes result in extremely narrow confidence intervals.

4.3 Quality of individual demographic information in FamiLinx data

Figure 3 indicates that observations with complete dates of birth and death (i.e., that specify the year and the month of birth and death) do not seem to show a preference for those years. Individuals for whom only information on the years is available are more prone to year heaping issues. Thus, observations with complete dates of birth and death are of higher quality. We can see an increase in quality over time for birth year heaping. Indeed, in the 19th century the percentages for individuals with complete dates are closer to the percentages for individuals with incomplete dates. Overall, the prevalence of death year heaping is lower than the prevalence of birth year heaping, which suggests that when the year of death is available it is more likely to be correct and precise.

When we look at the occurrence of birth and death year heaping across different countries of birth (see Figures A-2 and A-3 in the Appendix) we note similar trends, but with different magnitudes. In general, among all the considered countries, observations with complete birth dates do not seem to be affected by birth year heaping. Moreover, among those with missing birth months the proportion of birth years ending in 0 or 5 decreases over time. There is no evident improvement in the quality of the reported death dates over time. However, observations with complete death dates exhibit fewer instances of death year heaping than those with incomplete death dates across all the considered countries.

Figure 3: Percentages of years of birth and years of death ending with 0 or 5 by completeness of the dates of birth and death, and by historical period



Notes: Each color indicates different completeness of dates of birth and death. The blue line refers to dates with a non-missing month. The red line indicates dates with a missing month.

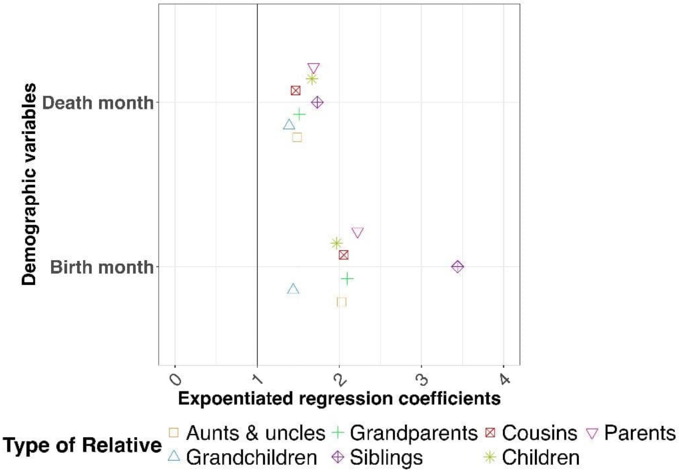
4.4 Quality of dates within kinship networks

We find a positive association between the quality of the birth and death dates for the focal individuals and those for their kin, net of the size of the kinship network (see Table A-4 in the Appendix). This implies that possessing more accurate demographic information is associated with a higher number of relatives with demographic information of higher quality. These associations are observed across distinct types of relatives for both birth and death dates, with the former showing the strongest association. As a robustness check we ran a logistic regression model using as the response a binary variable equal to 1 if at least one of the relatives of a given type for a focal individual has a non-missing month in the birth or death dates. As an additional sensitivity check we tested two other modeling approaches, namely a negative binomial regression model, where the number of relatives is treated as offset, and a binomial regression model. The results, included in the Appendix (Table A-6, Table A-8, Table A-10), are consistent with those of the negative binomial model presented in the main text.

Horizontal kin, especially siblings, tend to exhibit stronger associations for the variable ‘birth month.’ The expected number of siblings with a non-missing month in the birth date is almost four times higher for a focal individual with a non-missing month in the birth date compared to a focal individual with a missing month. Focusing on the death month, slightly higher associations are observed for siblings and children. The expected number of parents, siblings, and children with a non-missing month in the death date is at least 50% higher for a focal individual with a non-missing death month than for a focal individual with a missing month. Concerning the other relatives, the number of relatives with a non-missing month in death/birth date for a focal individual with a non-missing month in death/birth increases by over 20% compared to a focal with a missing month in the birth/death date.

Table A-4 in the Appendix displays all the regression coefficients, including the effects of the distinct birth and death cohorts on the number of relatives without a non-missing month in the birth/death date. In general, an increase in the magnitude of these cohort effects is observed, implying an improvement in the quality of the reported demographic information. Nonetheless, if we focus on grandchildren and children of focal individuals from more recent birth/death cohorts the associations are slightly lower, due to the fact that FamiLinx excludes individuals that were still alive in 2015.

Figure 4: Exponentiated coefficients from negative binomial regression measuring the association between a focal individual and their relatives in terms of the quality of the reported demographic variables



Note: The shapes denote the ratio of the expected number of relatives with a non-missing month in the birth/death dates for a focal individual with a non-missing month in the birth/death dates compared to that of a focal individual with a missing month in the birth/death dates. Estimates larger than 1 indicate that the expected number of relatives with a non-missing month in the birth/death dates is larger for a focal individual with a non-missing month in the birth/death dates than for a focal individual with a missing value. The results are reported by kin type and demographic variable. The distinct shapes and colors in the plot refer to different types of relatives. Confidence intervals are not included, as the large sample sizes result in extremely narrow confidence intervals.

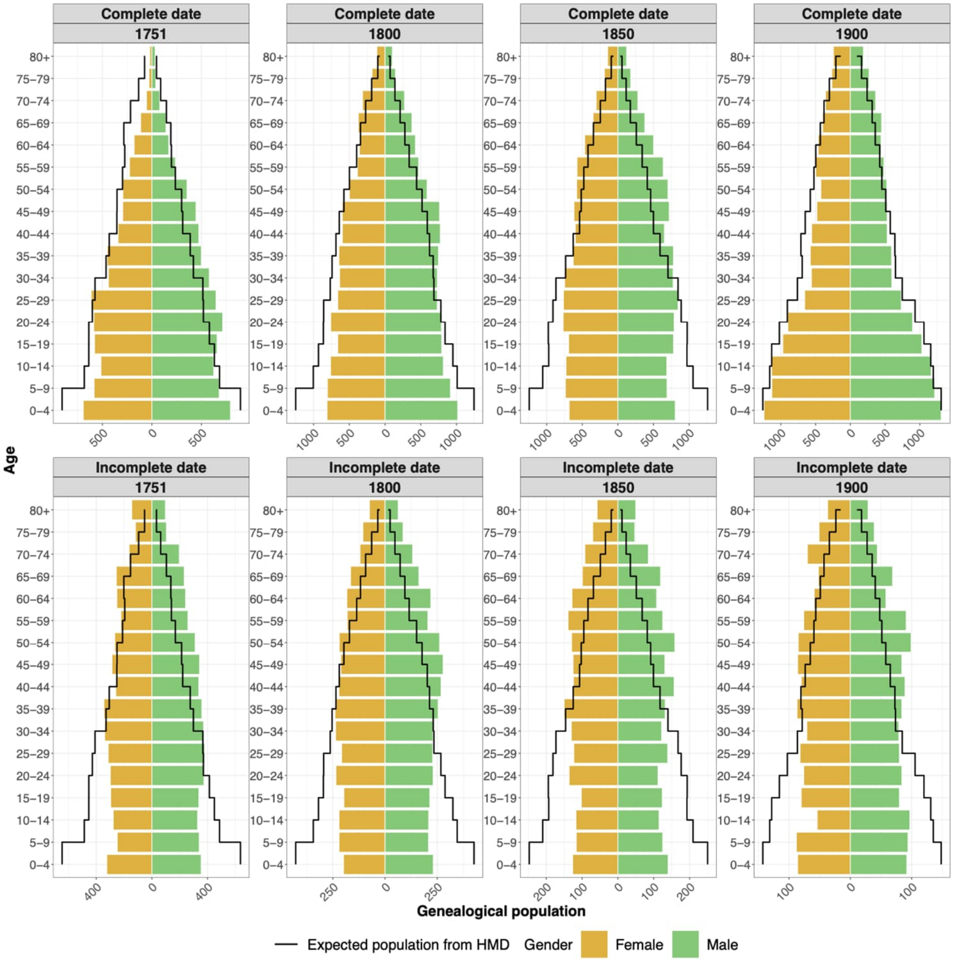
4.5 Discrepancies between the age–sex distribution in FamiLinx and in the registered population

We now compare the age–sex distribution of the digital population derived from online genealogies with that of the registered population. As an illustrative example, we concentrate on the Swedish genealogical population over the historical period of 1751–1900. Compared to other countries, Sweden stands out for its rich wealth of demographic data starting from the year 1751, including detailed population counts disaggregated by sex and age, which are available from population registers.

Figure 5 shows the percentage differences in age–sex proportions between the Swedish genealogical population extracted from FamiLinx and the registered Swedish population, over four calendar years: 1751, 1800, 1850, and 1900. These differences are computed for two distinct quality levels, one comprising individuals with precise birth and death dates (non-missing birth and death months), and the other comprising individuals with at least one less-precise date (the birth or the death month is missing). Notably, these disparities seem to be more modest for the genealogical group with higher information quality throughout the historical period under scrutiny. If we focus on the sample of Swedish individuals with precise birth and death dates, the age–sex distribution derived from this subsample mirrors the estimates for the total Swedish population toward the end of the 19th century from the Human Mortality Database. Nonetheless, regardless of the quality of the data used, a consistent pattern is observed for the Swedish genealogical population before the end of the 19th century. Individuals at younger ages and women tend to be underrepresented, whereas more longevous male individuals are overrepresented.

Figure A-4 in the Appendix shows that the underestimation of the proportions of individuals in the 0–14 age group with more accurate dates increases until the mid-19th century but then declines rapidly toward bias levels that are close to zero. Among adult individuals (aged 15–64) with higher quality information, males exhibit an upward bias that decreases toward the end of the 19th century. Conversely, females in the same age group are underrepresented in the second half of the 18th century (1751–1799) and of the 19th century (1851–1900), whereas they seem to be well-represented in the first part of the 19th century (1800–1850). Turning our attention to individuals aged 65 or older, we observe a consistent upward bias in the proportions for both men and women, which decreases slightly starting in the second half of the 19th century.

Figure 5: Population pyramids for the Swedish population from FamiLinx for the calendar years 1751, 1800, 1850, and 1900, by quality level



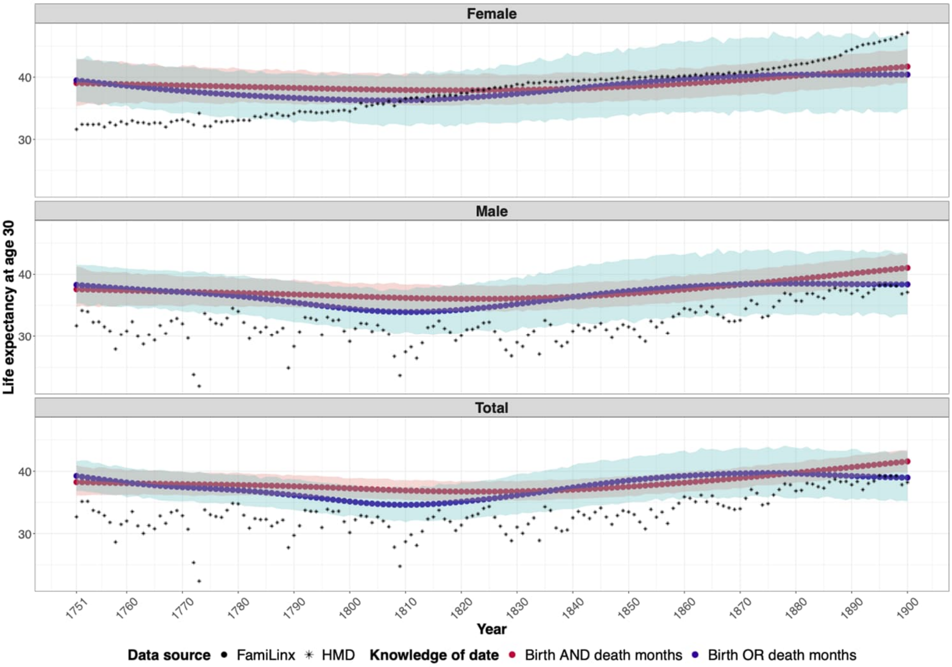
Note: The solid black lines refer to the age–sex distribution available in the Human Mortality Database. Yellow bars refer to female individuals and green bars refer to male individuals.

4.6 Discrepancies between life expectancy in FamiLinx and in the registered population

We now focus on investigating life expectancy at age 30 in Sweden over the period 1751–1900, specifically considering the two quality levels defined above. Our decision to evaluate life expectancy at age 30, as opposed to at birth, is motivated by the underestimation of child mortality inherent in the online genealogies (see Figure A-6 in the Appendix), and the recommendation of Stelter and Alburez-Gutierrez (2022). Again, we focus on Sweden due to its long time series of national demographic estimates. However, we acknowledge that our results for Sweden may not extend to populations from other countries.

Figure 6 presents the estimates of life expectancy at age 30 stratified by quality level and sex. To provide a benchmark, we incorporate life expectancy estimates from the Human Mortality Database. It is essential to note that our analysis is limited to individuals with non-missing birth and death years who were born and died in Sweden; i.e., to a sample of highly selected individuals. The results show a pronounced survivorship bias within the genealogical Swedish male population. In line with Stelter and Alburez-Gutierrez (2022) for Germany and the Netherlands, we find that the male life expectancy at age 30 estimated from genealogical data toward the end of the 19th century seems to be slightly closer to the life expectancy derived from Swedish register data. In contrast to our analysis of male longevity, our investigation of female longevity reveals unexpected trends in life expectancy at age 30. Throughout the 18th century this demographic indicator is consistently overestimated for the Swedish female population in FamiLinx. For the first half of the 19th century the estimates of life expectancy at age 30 based on genealogical data align with those from the Human Mortality Database. Nonetheless, a noteworthy shift can be observed toward the end of the 19th century, as the genealogical data consistently underestimate life expectancy at age 30 for women. In general, our analysis highlights that the observed trends in life expectancy at age 30 hold true across the quality groups under comparison. Nonetheless, our results also suggest that the bias in life expectancy at age 30 differs by gender. A possible explanation is suggested by Figure A-4 in the Appendix, in which the percentage of women in the age range 15–64 in Sweden is closer to the actual percentage from population registers during the period 1800–1870 than the share of men, which is more severely overestimated. On the contrary, in the last part of the 19th century, women aged 15–64 become more and more underrepresented, whereas the representation of men in the same age range improves. As a consequence, after 1870 we see a continuous increase in the underestimation of life expectancy at age 30 for women and a decrease in the overestimation for men. While this is an intriguing result, which would need further investigation, we lack sufficient tools to provide a robust explanation for the observed gender differences.

Figure 6: Life expectancy at age 30 in Sweden for the period 1751–1900, by sex and quality level (precise birth and death dates against at least one non-precise date) in FamiLinx, and Swedish life expectancy at age 30 from the Human Mortality Database



Note: Red lines refer to the estimates of life expectancy at age 30 calculated for Swedish individuals with non-missing birth and death months. Blue lines denote the estimates of life expectancy at age 30 among Swedish individuals whose birth or death month is missing. Star-shaped points denote the life expectancy estimates from the Human Mortality Database. Shaded regions refer to 95% confidence intervals, of which upper and lower bounds are obtained via Monte Carlo simulations.

5. Discussion

The extensive sample size and the availability of cross-border kinship networks render FamiLinx an asset for social scientists interested in exploring past population dynamics (Stelter and Alburez-Gutierrez 2022; Hsu et al. 2021; Cozzani et al. 2023) and the intergenerational transmission of demographic behaviors (Blanc 2024; Minardi, Corti, and Barban 2024). The availability of kinship ties and demographic information enables researchers to explore how demographic outcomes have changed within family networks. A noteworthy aspect is the extensive time period covered by the FamiLinx data, which facilitates the examination of long-term demographic processes. By drawing online digital trees, FamiLinx opens up new avenues for understanding the demographic behaviors of past populations through the lens of digital data, which in the field of Historical Demography are less common than other non-conventional data sources (i.e., parish records, obituaries, military records, wills). The coverage of various countries over the past four centuries provides researchers with the unique opportunity of analyzing the composition of transnational kinship networks.

In this study we have showed that when information on one demographic variable is known it is more likely that information on other demographic variables will also be known. Individuals with non-missing months in birth and death dates tend to have more precise demographic information whose quality improves over time. Furthermore, our analysis reveals that individuals with higher-quality demographic information are likely to have relatives with more complete and accurate demographic information. Additionally, using Sweden as an example, we observe that individuals with non-missing demographic information tend to experience higher life expectancy than the registered population throughout the considered historical period.

Most previous studies portray FamiLinx in a positive light and underline its potential for demographic research, leading to significant contributions, especially in the domain of Historical and Family Demography. However, we advise a cautious approach and provide some recommendations for scholars who want to utilize FamiLinx for their own research.

First, as outlined in Table A-1, the overrepresentation of individuals with vital events (births and deaths) in Western countries markedly restricts the geographical scope of the possible population studies. Previous studies that rely on this data source (Blanc 2021, 2024; Chong et al. 2022; Cozzani et al. 2023; Gay, Gobbi, and Goñi 2023; Hsu et al. 2021; Minardi, Corti, and Barban 2024; Pojman et al. 2023; Rawlik, Canela-Xandri, and Tenesa 2019; Stelter and Alburez-Gutierrez 2022) predominantly concentrate on the United States of America or Western European countries. Unfortunately, this is also almost inevitable in our study, as the vast majority of the individuals in the dataset and their family networks lived in those countries. Inevitably, when assessing the quality of the demographic information and comparing it with the population recorded in a given historical period it is necessary to limit the analysis to countries where such information is accessible.

Second, the high share of missing values in vital demographic variables, namely birth and death locations and dates, leads to a substantial reduction in the initial sample size of the data. This limitation was anticipated, as this data source was not primarily designed for population studies. Additionally, the omission of individuals who were alive in 2015 only permits the analysis of extinct birth cohorts. Hence, careful sample selection will enhance the robustness of research using FamiLinx and foster increased confidence in the completeness and quality of the chosen kinship network, enabling researchers to conduct population studies with a more solid foundation. Nonetheless, the restriction to individuals with demographic information of higher completeness and quality results in non-negligible selectivity issues. Potential FamiLinx users should maintain a critical approach to the available information in the dataset. Most individuals have missing demographic information, and even when some of their relatives can be identified the available information may be scarce.

Third, the age–sex distribution of online genealogical populations tends to diverge systematically from that observed in the general population. These observed divergences are a direct consequence of the under-representation of women and of individuals dying at young ages. Overlaying the age–sex distribution derived from the population register on the genealogy-based distribution overrepresents male individuals in older age groups, whereas women and individuals in younger age groups are generally underrepresented. Hence, scholars who are interested in examining the evolution of demographic processes in populations originating from FamiLinx are encouraged to implement bias-correcting methods to take into account the representation issues of this data source. The implementation of such methods allows researchers to obtain more accurate measures of common demographic processes; i.e., fertility, mortality, and migration. A Bayesian modeling framework can enable researchers to calibrate genealogy-based demographic indicators with more accurate estimates originating from more traditional data such as censuses and parish records, while accounting for the uncertainty of each source. For instance, Chong et al. 2022 propose a Bayesian modeling framework to correct age-specific mortality rates by combining online genealogical data with more precise estimates from the Human Mortality Database. Future research could employ a similar modeling framework to examine other demographic processes such as fertility by integrating information from multiple data streams.

Fourth, using Sweden as a test country, our results suggest that regardless of the quality of the demographic information, individuals from online genealogies are characterized by a persistently higher survival than the general population. Hence, researchers intending to harness this data source to gauge demographic outcomes should exercise caution. In general, demographic trajectories observed among individuals with non-missing birth and death years in FamiLinx are not representative of those of the broader population.

Another key consideration concerns the availability of relatives in the dataset and the completeness and quality of demographic information for the entire kinship network. FamiLinx's strength lies in its ability to provide information about relatives, facilitating the identification of kinship networks spanning multiple generations. Notably, our regression analyses underline that completeness and quality are clustered at the family level. In this regard, careful sample selection will allow researchers to conduct family-level demographic analysis. Specifically, researchers can employ the FamiLinx database to examine how fertility and longevity spread among different types of relatives beyond parents and children. Nonetheless, while the latter analysis can provide new knowledge about the transmission of demographic behaviors over time, the results should be interpreted with caution. It should be acknowledged that genealogical populations are highly selected under a set of favorable conditions, including higher survival and higher SES. This finding aligns with the existing literature about bias and selectivity in genealogies (Calderón Bernal et al. 2023; Zhao 2001; Hollingsworth 1976) and in FamiLinx (Stelter and Alburez-Gutierrez 2022; Minardi, Corti, and Barban 2024).

In conclusion, we encourage researchers to employ the FamiLinx data with caution. This data source provides great opportunities for demographic research, especially in the field of historical demography, due to its rich wealth of demographic information about individuals from various historical populations and its recorded kinship ties. Nonetheless, the inherent limitations of online genealogical data need to be addressed through the implementation of appropriate bias-correcting methods and through careful sample selection.

The findings and implications derived from this study are not automatically applicable to all (online) genealogical datasets. Specifically, the presented investigation is tailored to the unique attributes of the FamiLinx dataset, characterized by the availability of its demographic information and linked relatives. It is essential to acknowledge that different datasets may exhibit completely distinct temporal and geographical scopes, affecting the missingness and representativeness of the data. Nevertheless, we are confident that the methodologies and approaches employed in this study can be replicated to assess the completeness of the demographic information of other genealogies and to explore the association of these concepts within family networks.

6. Note on Reproducibility

To facilitate reproducibility of this research, we provide access to FamiLinx data as well as to the R codes needed to reproduce the tables and figures provided in the paper at the following Open Science Framework (OSF) repository: https://osf.io/ydzfq/.

7. Acknowledgments

Riccardo Omenti acknowledges funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No 865356). Andrea Colasurdo gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS). The authors thank the anonymous reviewers for their constructive feedback. The authors are also grateful to Diego Alburez-Gutierrez and Clara Mulder for valuable feedback on earlier versions of the manuscript.

Andrea Colasurdo and Riccardo Omenti contributed equally to this work and should both be considered first authors.

References

Becker, R.A., Wilks, A.R., and Brownrigg, R. (2022). Mapdata: Extra map databases. R package version 2.3.1. https://cran.r-project.org/web/packages/mapdata/index.html.

Blanc, G. (2021). Modernization before industrialization: Cultural roots of the demographic transition in France. (HAL science working papers). https://hal.science/hal-02318180v9.

Blanc, G. (2024). Demographic transitions, rural flight, and intergenerational persistence: Evidence from crowdsourced genealogies. (HAL working paper series hal-02922398).

Calderón Bernal, L.P., Alburez-Gutierrez, D., and Zagheni, E. (2023). Analyzing biases in genealogies using demographic microsimulation. (MPIDR working paper series WP-2023-034). Rostock: Max Planck Institute for Demographic Research. doi:10.4054/MPIDR-WP-2023-034.

Camarda, C.G. (2012). MortalitySmooth: An *R* package for smoothing Poisson counts with p-splines. *Journal of Statistical Software* 50(1). doi:10.18637/jss.v050.i01.

Cesare, N., Lee, H., McCormick, Z., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography* 55(5): 1979–1999. doi:10.1007/s13524-018-0715-2.

Chong, M., Alburez-Gutierrez, D., Del Fava, E., Alexander, M., and Zagheni, E. (2022). Identifying and correcting bias in big crowd-sourced online genealogies. (MPIDR working paper series WP-2022-005). Rostock: Max Planck Institute for Demographic Research. doi:10.4054/MPIDR-WP-2022-005.

Corti, G., Minardi, S., and Barban, N. (2024). Trends in assortative mating in the United States, 1700–1910. Evidence from FamiLinx data. *The History of the Family* 1–21. doi:10.1080/1081602X.2024.2352539.

Cozzani, M., Minardi, S., Corti, G., and Barban, N. (2023). Birth month and adult lifespan: A within-family, cohort, and spatial examination using FamiLinx data in the United States (1700–1899). *Demographic Research* 49(9): 201–218. doi:10.4054/DemRes.2023.49.9.

Cummins, N. (2017). Lifespans of the European Elite, 800–1800. *The Journal of Economic History* 77(2): 406–439. doi:10.1017/S0022050717000468.

Eilers, P.H.C., Currie, I.D., and Durban, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis* 50(1): 61–76. doi:10.1016/j.csda.2004.07.008.

Gavrilova, N.S. and Gavrilov, L.A. (2007). Search for predictors of exceptional human longevity: Using computerized genealogies and internet resources for human longevity studies. *North American Actuarial Journal* 11(1): 49–67. doi:10.1080/10920277.2007.10597437.

Gay, V., Gobbi, P., and Goñi, M. (2023). Revolutionary transition: Inheritance change and fertility decline. (HAL working paper series hal-04285818).

Henry, L. (1968). Historical demography. *Daedalus* 97(2): 385–396.

Hilbe, J.M. (2011). *Negative binomial regression*. 2nd ed. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511973420.

Hollingsworth, T.H. (1976). Genealogy and historical demography. *Annales de Démographie Historique* 167–170. doi:10.3406/adh.1976.1310.

Hsu, C.-H., Posegga, O., Fischbach, K., and Engelhardt, H. (2021). Examining the trade-offs between human fertility and longevity over three centuries using crowdsourced genealogy data. *PLoS One* 16(8): e0255528. doi:10.1371/journal.pone.0255528.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D.G., Price, A.L., and Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science* 360(6385): 171–175. doi:10.1126/science.aam9309.

Kasakoff, A.B. and Adams, J.W. (1995). The effect of migration on ages at vital events: A critique of family reconstitution in historical demography. *European Journal of Population / Revue Européenne de Démographie* 11(3): 199–242. doi:10.1007/BF01264948.

Kashyap, R. (2021). Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem. *Population Studies* 75(sup1): 47–75. doi:10.1080/00324728.2021.1969031.

Mare, R.D. (2011). A multigenerational view of inequality. *Demography* 48(1): 1–23. doi:10.1007/s13524-011-0014-7.

Minardi, S., Corti, G., and Barban, N. (2024). Historical patterns in the intergenerational transmission of lifespan and longevity: A research note on US cohorts born between 1700 and 1900. *Demography* 61(4): 979–994. doi:10.1215/00703370-11458359.

Mooney, C.Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: SAGE Publications. doi:10.4135/9781412985116.

Otterstrom, S.M. and Bunker, B.E. (2013). Genealogy, migration, and the intertwined geographies of personal pasts. *Annals of the Association of American Geographers* 103(3): 544–569. doi:10.1080/00045608.2012.700607.

Pojman, E., Mwedzi, D.E., Bucaro, O.O., Zhang, S., Chong, M., Alexander, M., and Alburez-Gutierrez, D. (2023). Leaving for life: Using online crowd-sourced genealogies to estimate the migrant mortality advantage for the United Kingdom and Ireland during the 18th and 19th centuries. (MPIDR working paper series WP-2023-050). Rostock: Max Planck Institute for Demographic Research. doi:10.4054/MPIDR-WP-2023-050.

Post, W., van Poppel, F., van Imhoff, E., and Kruse, E. (1997). Reconstructing the extended kin-network in the Netherlands with genealogical data: Methods, problems, and results. *Population Studies* 51(3): 263–278. doi:10.1080/00324 72031000150046.

Rawlik, K., Canela-Xandri, O., and Tenesa, A. (2019). Indirect assortative mating for human disease and longevity. *Heredity* 123(2): 106–116. doi:10.1038/s41437-019-0185-3.

Song, X. and Campbell, C.D. (2017). Genealogical microdata and their significance for social science. *Annual Review of Sociology* 43(1): 75–99. doi:10.1146/annurev-soc-073014-112157.

Spoorenberg, T. and Dutreuilh, C. (2007). Quality of age reporting: Extension and application of the modified Whipple's index. *Population (English Edition)* 62(4): 729–741. doi:10.3917/popu.704.0847.

Stelter, R. and Alburez-Gutierrez, D. (2022). Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics. *Proceedings of the National Academy of Sciences* 119(10): e2120455119. doi:10.1073/pnas.2120455119.

Stockwell, E.G. and Wicks, J.W. (1974). Age heaping in recent national censuses. *Social Biology* 21(2): 163–167. doi:10.1080/19485565.1974.9988102.

United Nations (2016). A review of key concepts: Coverage and completeness. (UN Expert Group Meeting, 3–4 November 2016). New York: United Nations.

Wrigley, E.A. (1981). The prospects for population history. *The Journal of Interdisciplinary History* 12(2): 207–226. doi:10.2307/203025.

Zhao, Z. (2001). Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies* 55(2): 181–193. doi:10.1080/00324720127690.

Appendix

Table A-1: Absolute number of births and deaths for the top 20 countries in terms of number of births and deaths

Country	Number of births	Number of deaths
USA	2,479,761	2,122,063
UK	936,188	324,630
NORWAY	468,391	281,471
SWEDEN	359,999	222,005
NETHERLANDS	301,079	184,430
GERMANY	298,271	137,164
ESTONIA	267,137	121,194
CANADA	248,248	185,322
DENMARK	180,569	97,780
FRANCE	177,715	112,167
POLAND	112,382	58,575
FINLAND	111,272	73,401
AUSTRALIA	94,687	90,788
SPAIN	81,812	24,752
IRELAND	69,739	17,991
BELGIUM	67,638	46,338
INDIA	67,132	52,773
SWITZERLAND	55,116	17,388
SOUTH AFRICA	50,815	44,364
RUSSIA	49,605	24,145
ITALY	36,962	16,487
CZECH REPUBLIC	24,237	13,971
NEW ZEALAND	20,368	23,738
ISRAEL	10,890	35,030

Table A-2: Absolute frequency and percentage of missing and non-missing values in relevant demographic variables in the complete sample and in the analytical sample

Variable	Complete sample	Analytical sample
Sample size	86,124,644	7,618,651
Gender		
Missing	14,925,928 (17.33%)	4708 (0.06%)
Male	37,997,466 (44.12%)	4,108,522 (53.93 %)
Female	33,201,250 (38.55%)	3,505,421 (46.01 %)
Birth date information		
Missing	52,405,914 (60.85%)	677,215 (8.89 %)
Only year	13,692,092 (15.90%)	2,389,882 (31.37 %)
Year and month	849,377 (0.99%)	195,874 (2.57 %)
Complete date	19,177,261 (22.27%)	4,355,680 (57.17 %)
Death date information		
Missing	64,383,957 (74.77%)	2,656,270 (34.87 %)
Only year	6,736,492 (7.82%)	1,143,731 (15.01 %)
Year and month	853,888 (0.99%)	217,310 (2.85 %)
Complete date	14,150,307 (16.43%)	3,601,340 (47.27 %)
Birth location information		
Missing	70,464,808 (81.82%)	1,048,638 (13.76 %)
Reported	15,659,836 (18.18%)	6,570,013 (86.24 %)
Death location information		
Missing	74,861,173 (86.92%)	3,290,684 (43.19 %)
Reported	11,263,471 (13.08%)	4,327,967 (56.81 %)
Parent/child linkage		
Missing	47,172,309 (54.77%)	
At least one link	38,952,335 (45.23%)	7,618,651 (100.00%)

Details of the estimation of smoothed mortality rates by age and sex

To estimate life expectancy at age 30 from online genealogies, we rely on the R package developed by Camarda (2012), which allows smoothing mortality rates over years and ages.

We consider mortality experienced by individuals who were born and died in Sweden during the historical period 1751–1900. To obtain smoothed estimates of mortality rates by year and age, we model the number of deaths in Sweden in a year t at an age x, $Y_{x,t}$, as a Poisson distribution.

$$Y_{x,t} \sim Pois(E_{x,t} \cdot \mu_{x,t})$$
$$x = 30, \dots, 80$$
$$t = 1751, \dots, 1900$$

where $E_{x,t}$ indicates the number of exposed Swedish individuals in year *t* and age *x* and $\mu_{x,t}$ denotes the risk of death for Swedish individuals aged *x* in year *t*.

For the performance of the mortality analysis, death counts, exposure, and mortality risks by year and age are arranged in rectangular matrices, called Y, M, and E, in which rows represent ages and columns refer to years. The smoothness is achieved by incorporating two-dimensional P-splines. Specifically, we model the mean of the Poisson distribution of the number of deaths as follows.

$$\ln(E(Y)) = ln\, ln\,(E) + ln(M) = ln\, ln\,(E) + B_y A B_a^T$$

In the model, the B-splines spaced over the ages are stored in the regression matrix B_a of dimension $k_a \times k_a$. The B-splines spaced over the years are stored in the regression B_y of dimension $k_y \times k_y$. Both B_a and B_y have an associated set of regression coefficients. Note that the numbers k_a and k_y indicate the number of B-splines chosen over the ages (k_a) and years (k_y). Following the guidelines by Camarda (2012), we chose B-splines that are equally spaced over the years and the ages. The rows of the matrix *A* of dimension $k_a \times k_y$ denote the regression coefficients for B_a, whereas its columns indicate the regression coefficients for B_y. The estimation of the regression parameters is performed via Iterative Regression Weighted Least Squares (IRWLS). We set the diagonal matrix of weights required for this estimation procedure to be equal to the identity.

To reduce the number of parameters in the model, we can choose the number of B-splines with an additional two-dimensional penalty *P* on the regression coefficients.

$$P = \lambda_a \left(I_{k_y} \otimes D_a^T D_a\right) + \lambda_y \left(I_{k_a} \otimes D_y^T D_y\right)$$

where λ_a and λ_y are the smoothing parameters used for the ages and the years. D_a and D_y are the difference matrices. I_{k_y} and I_{k_a} are identity matrices of dimension k_y and k_a respectively. The symbol $\otimes$ stands for the Kronecker product. The optimal values for λ_a and λ_y are chosen so that either the Bayesian Information Criterion (BIC) or Akaike Information Criterion are minimized.

To smooth mortality rates in R, we employed the function *mort2Dsmooth(x,y,Z,offset)* from the R package *mortalitysmooth* by Camarda (2012). The function requires the following arguments:

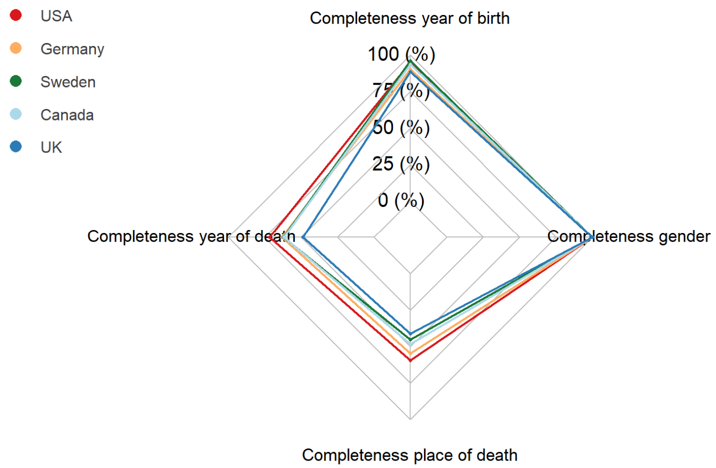
- A vector of ages named x (in our application x= 30,…,80)
- A vector of years named y (in our application y=1751,…,1900)
- A matrix of death counts over ages (rows) and years (columns) named Z (matrix Y in the model notation)
- A matrix of logged population counts over ages (rows) and years (columns) named offset (matrix E with log-transformed entries in the model notations)

Concerning the remaining arguments, we opted for the default options.
Optional arguments include:

- The degree of the polynomials for the construction of B-splines (q), whose default option is set to be *q*=3 (necessary for the construction of matrix B)
- The order of the differences for the penalty matrix (*d*), whose default option is set to be *d*=2 (necessary for the specification of penalty matrices D_a and D_y)
- A matrix of weights over the ages and years (*W*) which is set by default to be equal to the identity matrix (necessary for the specification of the diagonal matrix of weights to be used in the estimation of the regression coefficients)
- The selection of the smoothing parameters (λ_a and λ_y in the model notation) is carried out by default using Bayesian Information Criterion (BIC). Alternative selection criteria can be specified via the option *method.*

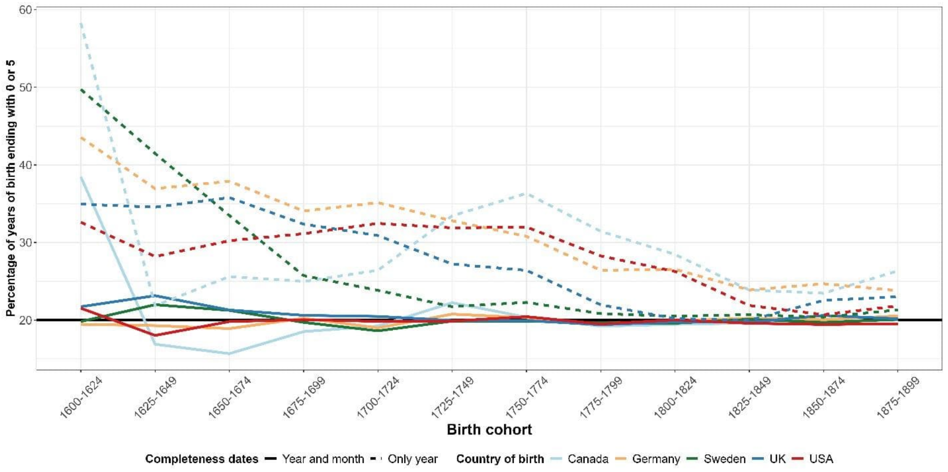
As part of the output, the function *mort2Dsmooth* provides a matrix of smoothed mortality rates over the ages and the years. Exploiting standard life table relationships, these smoothed mortality rates by age and year are then used to obtain smoothed estimates of life expectancy at birth and at age 30.

Figure A-1: Percentage of non-missing values for 4 relevant demographic variables in the dataset, by country of birth of the focal individual



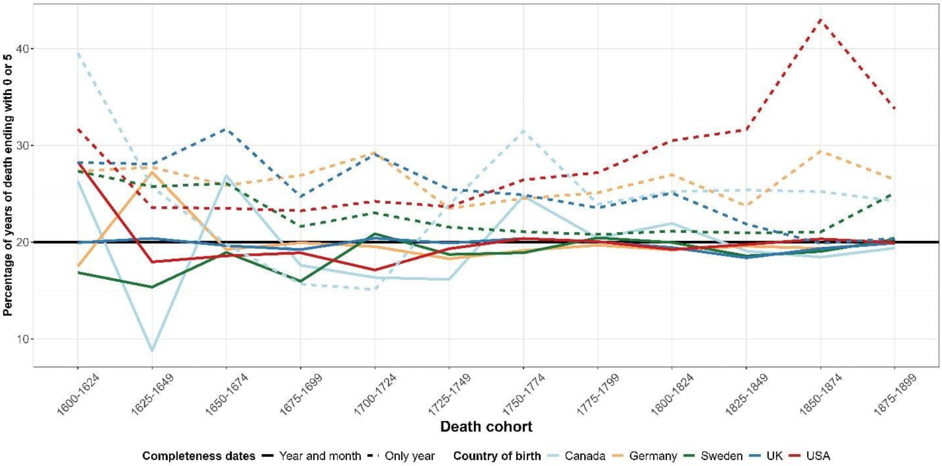
Note: Each color indicates a different sample composed of individuals born in selected countries, and each colored line connects the percentages of non-missing information in each of the four variables considered.

Figure A-2: Percentage of years of birth ending with 0 or 5, by country of birth and birth cohort



Note: Each color indicates a different country of birth. The solid line refers to individuals with a non-missing birth month. The dotted line indicates individuals with a missing birth month.

Figure A-3: Percentage of years of death ending with 0 or 5, by country of birth and death cohort



Note: Each color indicates a different country of birth. The solid line refers to individuals with a non-missing death month. The dotted line indicates individuals with a missing death month.

Table A-3: Coefficients of the negative binomial regression models to test association in terms of completeness, by type of relative and demographic variable

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth year	Intercept	–0.9420 (0.0015)	–2.1268 (0.0021)	–0.6295 (0.0011)	–0.3781 (0.0013)	0.4618 (0.0011)	0.7929 (0.0016)	–0.3481 (0.0020)
	Yes	0.7193 (0.0015)	0.7448 (0.0009)	0.4147 (0.0010)	1.0749 (0.0013)	0.4516 (0.0011)	0.6863 (0.0016)	0.4259 (0.0020)
	No. of relatives	0.2361 (0.0001)	0.9680 (0.0010)	0.3507 (0.0002)	0.1412 (0.0000)	0.1139 (0.0000)	0.0510 (0.0000)	0.1334 (0.0000)
Death year	Intercept	–1.1837 (0.0012)	–2.0984 (0.0031)	–0.5676 (0.0008)	–0.5448 (0.0008)	0.0237 (0.0008)	0.5744 (0.0010)	–0.4926 (0.0015)
	Yes	0.6840 (0.0012)	0.4690 (0.0006)	0.1647 (0.0005)	0.6939 (0.0007)	0.3192 (0.0008)	0.4019 (0.0010)	0.2377 (0.0016)
	No. of relatives	0.2147 (0.0001)	1.0318 (0.0016)	0.0909 (0.0002)	0.1683 (0.0001)	0.1373 (0.0001)	0.0536 (0.0000)	0.1185 (0.0000)
Birth country	Intercept	–1.1167 (0.0015)	–2.2153 (0.0036)	–0.9520 (0.0016)	–0.8279 (0.0010)	–0.3143 (0.0013)	0.3158 (0.0014)	–0.6374 (0.0016)
	Yes	0.4913 (0.0015)	0.5097 (0.0012)	0.3544 (0.0013)	0.4845 (0.0010)	0.6805 (0.0011)	0.7612 (0.0014)	0.3553 (0.0017)
	No. of relatives	0.2311 (0.0001)	0.9417 (0.0017)	0.3389 (0.0003)	0.0244 (0.0001)	0.1387 (0.0001)	0.0521 (0.0000)	0.1237 (0.0000)
Death country	Intercept	–1.4895 (0.0014)	–2.2700 (0.0043)	–0.8599 (0.0014)	–0.9690 (0.0011)	–0.5005 (0.0012)	0.0795 (0.0012)	–0.8895 (0.0015)
	Yes	0.6586 (0.0014)	0.4457 (0.0008)	0.2254 (0.0009)	0.6926 (0.0010)	0.3858 (0.0011)	0.4506 (0.0013)	0.2517 (0.0017)
	No. of relatives	0.1982 (0.0002)	0.9804 (0.0022)	0.3275 (0.0004)	0.1747 (0.0001)	0.1437 (0.0001)	0.0536 (0.0000)	0.1067 (0.0001)
No. of relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
No. of focal individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Note: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the negative binomial regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals in the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative, they are omitted from the regression model. Also included are the sample size of focal individuals employed for each model and the size of their kinship network.

Table A-4: Coefficients of the negative binomial regression models to test association in terms of quality, by type of relative and demographic variable

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth date	Intercept	–1.0791 (0.0041)	–2.6315 (0.0085)	–2.6170 (0.0113)	–0.8550 (0.0045)	–0.7846 (0.0082)	0.1840 (0.0098)	–0.2631 (0.0038)
	Non-missing month	0.6772 (0.0010)	0.7982 (0.0011)	0.6793 (0.0013)	1.2354 (0.0009)	0.7056 (0.0011)	0.7204 (0.0012)	0.3632 (0.0013)
	No. of relatives	0.2129 (0.0001)	0.6265 (0.0026)	0.3616 (0.0005)	0.1468 (0.0001)	0.1471 (0.0001)	0.0538 (0.0000)	0.1150 (0.0001)
	Birth period							
	1625–1649	0.1674 (0.0051)	0.1987 (0.0084)	0.3613 (0.0136)	0.0559 (0.0056)	0.1251 (0.0103)	0.1090 (0.0123)	0.0671 (0.0050)
	1650–1674	0.2303 (0.0047)	0.3442 (0.0077)	0.5792 (0.0123)	0.1015 (0.0050)	0.1822 (0.0093)	0.2555 (0.0110)	0.1307 (0.0046)
	1675–1699	0.2768 (0.0045)	0.5427 (0.0073)	0.7764 (0.0118)	0.1509 (0.0048)	0.3203 (0.0088)	0.3463 (0.0104)	0.1650 (0.0045)
	1700–1724	0.2841 (0.0044)	0.6479 (0.0071)	0.9946 (0.0115)	0.1791 (0.0047)	0.3709 (0.0085)	0.3629 (0.0102)	0.1991 (0.0044)
	1725–1749	0.2984 (0.0044)	0.7310 (0.0070)	1.1281 (0.0114)	0.1947 (0.0046)	0.4343 (0.0084)	0.3556 (0.0101)	0.2859 (0.0043)
	1750–1774	0.3455 (0.0043)	0.7460 (0.0070)	1.1999 (0.0114)	0.1885 (0.0046)	0.4569 (0.0084)	0.3605 (0.0100)	0.3661 (0.0042)
	1775–1799	0.4151 (0.0042)	0.7807 (0.0070)	1.2028 (0.0114)	0.2064 (0.0045)	0.4426 (0.0083)	0.3902 (0.0100)	0.3896 (0.0042)
	1800–1824	0.3879 (0.0042)	0.8312 (0.0069)	1.1996 (0.0113)	0.2306 (0.0045)	0.4398 (0.0083)	0.4193 (0.0099)	0.4447 (0.0042)
	1825–1849	0.4260 (0.0042)	0.9354 (0.0069)	1.2760 (0.0113)	0.2224 (0.0045)	0.5136 (0.0083)	0.4255 (0.0099)	0.5089 (0.0042)
	1850–1874	0.5380 (0.0042)	1.0499 (0.0069)	1.3887 (0.0113)	0.2335 (0.0045)	0.5537 (0.0082)	0.4420 (0.0099)	0.4438 (0.0043)
	1875–1900	0.5432 (0.0042)	1.1066 (0.0069)	1.5278 (0.0113)	0.2715 (0.0045)	0.5739 (0.0082)	0.4877 (0.0099)	0.2283 (0.0048)
	No. of relatives	9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
	No. of focal individuals	2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717

Table A-4: (Continued)

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Death date	Intercept	–0.8865 (0.0060)	–1.7394 (0.0081)	–1.2475 (0.0085)	–0.3347 (0.0068)	–0.2626 (0.0106)	0.3442 (0.0136)	–0.3524 (0.0066)
	Non-missing month	0.5103 (0.0013)	0.5223 (0.0013)	0.4158 (0.0015)	0.5491 (0.0012)	0.3973 (0.0015)	0.3857 (0.0018)	0.3287 (0.0015)
	No. of relatives	0.2164 (0.0001)	0.6759 (0.0026)	0.3893 (0.0006)	0.1616 (0.0001)	0.1534 (0.0001)	0.0719 (0.0001)	0.1338 (0.0001)
	Death period							
	1625–1649	0.0752 (0.0076)	0.0221 (0.0085)	0.0497 (0.0110)	0.0205 (0.0088)	0.0109 (0.0138)	0.0514 (0.0177)	0.0698 (0.0084)
	1650–1674	0.0831 (0.0070)	0.0281 (0.0079)	0.0713 (0.0102)	0.0456 (0.0080)	0.0496 (0.0127)	0.0918 (0.0162)	0.0654 (0.0077)
	1675–1699	0.1300 (0.0065)	0.0994 (0.0072)	0.1252 (0.0094)	0.0903 (0.0074)	0.1206 (0.0116)	0.1773 (0.0148)	0.0675 (0.0072)
	1700–1724	0.1470 (0.0064)	0.1654 (0.0070)	0.1964 (0.0090)	0.1194 (0.0071)	0.1491 (0.0112)	0.2028 (0.0143)	0.1038 (0.0071)
	1725–1749	0.1690 (0.0063)	0.1938 (0.0068)	0.2516 (0.0087)	0.1193 (0.0070)	0.1684 (0.0109)	0.2177 (0.0140)	0.1666 (0.0069)
	1750–1774	0.2124 (0.0062)	0.2047 (0.0067)	0.2877 (0.0086)	0.1383 (0.0069)	0.1852 (0.0107)	0.2357 (0.0139)	0.2310 (0.0069)
	1775–1799	0.2637 (0.0061)	0.1993 (0.0067)	0.2893 (0.0086)	0.1509 (0.0069)	0.1926 (0.0107)	0.2494 (0.0138)	0.2857 (0.0068)
	1800–1824	0.3176 (0.0061)	0.2000 (0.0066)	0.2848 (0.0085)	0.1792 (0.0068)	0.2064 (0.0107)	0.2811 (0.0137)	0.3521 (0.0068)
	1825–1849	0.3530 (0.0061)	0.2144 (0.0066)	0.2745 (0.0085)	0.2040 (0.0068)	0.2377 (0.0106)	0.3103 (0.0137)	0.3930 (0.0068)
	1850–1874	0.3905 (0.0060)	0.2607 (0.0066)	0.2980 (0.0085)	0.2345 (0.0068)	0.2791 (0.0106)	0.3326 (0.0137)	0.4222 (0.0067)
	1875–1900	0.4089 (0.0060)	0.2947 (0.0066)	0.3477 (0.0085)	0.2530 (0.0068)	0.3074 (0.0106)	0.3414 (0.0136)	0.4433 (0.0067)
	No. of relatives	3,420,063	7,531,416	3,008,568	5,514,632	4,085,603	6,918,263	4,440,454
	No.. of focal individuals	1,379,573	3,877,896	1,567,008	1,173,828	816,322	615,066	1,051,907

Note: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the negative binomial regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals in the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative they are is omitted from the regression model. Also included are the sample size of focal individuals employed for each model and the size of their kinship network.

Table A-5: Coefficients of the logistic regression models to test association in terms of completeness, by type of relative and demographic variable

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth year	Intercept	–0.6938 (0.0033)	–3.3657 (0.0076)	–1.0098 (0.0061)	–1.7600 (0.0057)	0.7014 (0.0071)	0.6504 (0.0076)	–0.3283 (0.0039)
	Yes	1.5491 (0.0033)	2.4178 (0.0036)	1.7874 (0.0045)	3.1204 (0.0051)	2.4317 (0.0066)	2.3151 (0.0073)	0.8848 (0.0040)
	No. of relatives	0.1427 (0.0004)	1.7515 (0.0036)	0.6921 (0.0021)	0.4011 (0.0010)	0.3556 (0.0012)	0.1751 (0.0007)	0.0666 (0.0002)
Death year	Intercept	–0.7980 (0.0023)	–2.4095 (0.0065)	–0.4066 (0.0038)	–1,2217 (0.0030)	–0.5456 (0.0036)	–0.4848 (0.0038)	-0.2924 (0.0023)
	Yes	1.0737 (0.0023)	1.3920 (0.0021)	0.8335 (0.0026)	1.6112 (0.0027)	1.0649 (0.0034)	1.1940 (0.0041)	0.2956 (0.0029)
	No. of relatives	0.1621 (0.0004)	1.4078 (0.0033)	0.5480 (0.0014)	0.3391 (0.0005)	0.3035 (0.0006)	0.1576 (0.0004)	0.0855 (0.0003)
Birth country	Intercept	–0.7375 (0.0027)	–2.3430 (0.0066)	–0.8112 (0.0043)	–1.0073 (0.0038)	–0.7838 (0.0037)	–0.5777 (0.0041)	–0.6096 (0.0030)
	Yes	0.7711 (0.0027)	1.0423 (0.0027)	0.7506 (0.0033)	1.5841 (0.0033)	1.4575 (0.0032)	1.3856 (0.0040)	0.5212 (0.0027)
	No, of relatives	0.1219 (0.0004)	1.1878 (0.0031)	0.4415 (0.0012)	0.1857 (0.0004)	0.1722 (0.0004)	0.0934 (0.0002)	0.0744 (0.0002)
Death country	Intercept	–1.2219 (0.0022)	–2.1237 (0.0064)	–0.6300 (0.0031)	–1.3860 (0.0023)	–0.8676 (0.0028)	–0.7655 (0.0030)	–0.8117 (0.0023)
	Yes	0.8803 (0,0022)	0.8904 (0.0018)	0.4692 (0.0021)	1.1758 (0.0023)	0.7714 (0.0027)	0.8570 (0.0034)	0.2280 (0.0025)
	No, of relatives	0.1540 (0,0003)	1.0258 (0.0033)	0.3681 (0.0011)	0.2622 (0.0004)	0.2153 (0.0004)	0.1147 (0.0002)	0.0905 (0.0002)
No. of relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
No. of focal individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Note: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Here, we used a binary response indicating whether a focal individual has at least one relative with a non-missing value in a demographic variable. Coefficients of the logistic regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals in the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative they are omitted from the regression model. Also included are the sample size of focal individuals employed for each model and the size of their kinship network.

Table A-6: Coefficients of the logistic regression models to test association in terms of quality, by type of relative and demographic variable

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth date	Intercept	–1.1422 (0.0086)	–2.7385 (0.0166)	–3.0844 (0.0171)	–1.9024 (0.016)	–2.1304 (0.0215)	–1.2527 (0.0284)	–0.4264 (0.0090)
	Non-missing month	1.8989 (0.0033)	1.7805 (0.0027)	1.2973 (0.0027)	3.1068 (0.0039)	1.9150 (0.0038)	2.1687 (0.0054)	1.2711 (0.0045)
	No. of relatives	0.1771 (0.0006)	0.5114 (0.0057)	0.3874 (0.0014)	0.2340 (0.006)	0.2269 (0.0006)	0.1136 (0.0004)	0.3188 (0.0016)
	Birth period							
	1625–1649	0.2587 (0.0113)	0.3021 (0.0160)	0.4816 (0.0206)	0.1350 (0.0208)	0.2380 (0.0273)	0.1241 (0.0365)	0.1241 (0.0365)
	1650–1674	0.3667 (0.0107)	0.4504 (0.0148)	0.7692 (0.0189)	0.1821 (0.0193)	0.3559 (0.0250)	0.2381 (0.0336)	0.2381 (0.0336)
	1675–1699	0.4610 (0.0104)	0.7343 (0.0141)	1.0403 (0.0181)	0.3437 (0.0184)	0.5421 (0.0249)	0.4698 (0.0320)	0.4698 (0.0320)
	1700–1724	0.4945 (0.0101)	0.9118 (0.0138)	1.2970 (0.0176)	0.3769 (0.0179)	0.6104 (0.0229)	0.5578 (0.0308)	0.5578 (0.0308)
	1725–1749	0.6056 (0.0010)	1.0753 (0.0136)	1.4905 (0.0174)	0.4457 (0.0175)	0.7636 (0.0225)	0.6310 (0.0302)	0.6307 (0.0302)
	1750–1774	0.7912 (0.0099)	1.1161 (0.0135)	1.6396 (0.0173)	0.5445 (0.0172)	0.8384 (0.0221)	0.67945 (0.0298)	0.6795 (0.0298)
	1775–1799	1.0089 (0.0099)	1.1923 (0.0133)	1.6767 (0.0172)	0.6529 (0.0170)	0.8919 (0.0220)	0.7628 (0.0294)	0.7628 (0.0295)
	1800–1824	0.8771 (0.0098)	1.2782 (0.0133)	1.6752 (0.0171)	0.8224 (0.0169)	0.9105 (0.0219)	0.8877 (0.0294)	0.8877 (0.0294)
	1825–1849	0.8083 (0.0098)	1.5914 (0.0132)	1.8173 (0.0170)	0.6775 (0.0166)	1.1986 (0.0218)	0.8772 (0.0288)	0.8772 (0.0291)
	1850–1874	1.0600 (0.0099)	2.0107 (0.0134)	2.0420 (0.0170)	0.6391 (0.0165)	1.2891 (0.0217)	0.7806 (0.0290)	0.7810 (0.0230)
	1875–1900	1.3812 (0.0110)	2.2704 (0.0134)	2.4092 (0.0170)	0.7833 (0.0166)	1.2924 (0.0216)	0.8716 (0.0289)	0.8712 (0.0289)
	No. of relatives	9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
	No. of focal individuals	2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717

Table A-6: (Continued)

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Death date	Intercept	0.8960 (0.0017)	–1.6829 (0.0222)	–1.6102 (0.0222)	–1.1530 (0.0310)	–1.0645 (0.0419)	–0.5877 (0.0563)	–0.5020 (0.0018)
	Non-missing month	1.5848 (0.0046)	1.5308 (0.0004)	1.2019 (0.0048)	2.0668 (0.0032)	1.5629 (0.0073)	1.6757 (0.0099)	1.2277 (0.0024)
	No. of relatives	0.3717 (0.0018)	0.7297 (0.0067)	0.5626 (0.0028)	0.4409 (0.0017)	0.4211 (0.0019)	0.2634 (0.0016)	0.3188 (0.0016)
	Death period							
	1625–1649	0.1781 (0.0215)	0.0198 (0.0249)	0.0993 (0.0048)	–0.0414 (0.0411)	–0.0138 (0.0548)	–0.0357 (0.0737)	0.0980 (0.0239)
	1650–1674	0.2260 (0.0198)	–0.0110 (0.0230)	0.1366 (0.0286)	–0.0723 (0.0379)	–0.0633 (0.0508)	0.0258 (0.0639)	0.1005 (0.0207)
	1675–1699	0.2360 (0.0183)	0.1700 (0.0214)	0.2205 (0.0265)	0.0922 (0.0355)	0.0576 (0.0472)	0.0877 (0.0639)	0.0987 (0.0207)
	1700–1724	0.2194 (0.0183)	0.3496 (0.0208)	0.4030 (0.0256)	0.1525 (0.0354)	0.1011 (0.0455)	0.0744 (0.0595)	0.1345 (0.0203)
	1725–1749	0.2135 (0.0179)	0.4419 (0.0203)	0.5500 (0.0248)	0.1075 (0.0333)	0.1401 (0.0440)	0.1614 (0.0586)	0.2632 (0.0199)
	1750–1774	0.3353 (0.0178)	0.4800 (0.0200)	0.6601 (0.0245)	0.1908 (0.0329)	0.1842 (0.0434)	0.2456 (0.0586)	0.4471 (0.0198)
	1775–1799	0.5170 (0.0177)	0.4740 (0.0200)	0.6883 (0.0242)	0.2737 (0.0326)	0.2532 (0.0431)	0.3522 (0.0577)	0.5624 (0.0199)
	1800–1824	0.6661 (0.0176)	0.4654 (0.0197)	0.7010 (0.0240)	0.3777 (0.0324)	0.3094 (0.0427)	0.4237 (0.0574)	0.7006 (0.0200)
	1825–1849	0.7529 (0.0177)	0.5112 (0.0195)	0.6946 (0.0238)	0.4546 (0.0320)	0.4119 (0.0424)	0.5084 (0.0572)	0.7334 (0.0202)
	1850–1874	0.8162 (0.0177)	0.6769 (0.0194)	0.7627 (0.0236)	0.5516 (0.0320)	0.5540 (0.0423)	0.5084 (0.0572)	0.7559 (0.0230)
	1875–1900	0.7939 (0.0178)	0.7846 (0.0194)	0.9077 (0.0237)	0.5512 (0.0320)	0.6370 (0.0423)	0.5024 (0.0571)	0.7791 (0.0205)
	No. of relatives	3,420,063	7,531,416	3,008,568	5,514,632	4,085,603	6,918,263	4,440,454
	No. of focal individuals	1,379,573	3,877,896	1,567,008	1,173,828	816,322	615,066	1,051,907

Note: All p-values are smaller than 0.01, standard errors are shown in parentheses and Y denotes the inclusion of controls in the logistic regression models. The models are fitted considering the individuals in the analytical sample, who were born and/or died in the historical period 1600–1900, and their kinship network. For each type of relative, if the focal individual does not have that type of relative they are omitted from the regression model. We include the birth and death years as a control, which are grouped in 25-year classes. These classes are entered in the regression as a series of dummies. Also included in the table are the number of relatives and number of focal individuals with non-missing birth (death) years for each regression model.

Table A-7: Coefficients of the negative binomial regression models with number of relatives as offset to test association in terms of completeness, by type of relative and demographic variable

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth year	Intercept	1.0412 (0.0014)	–0.9092 (0.0009)	–0.5998 (0.0009)	–1.0632 (0.0010)	–0.4737 (0.0013)	–0.7321 (0.0007)	–0.8762 (0.0022)
	Yes	0.7078 (0.0014)	0.7606 (0.0009)	0.4027 (0.0010)	0.9638 (0.0010)	0.3885 (0.0013)	0.5841 (0.0013)	0.3765 (0.0023)
Death year	Intercept	–1.4047 (0.0011)	–0.5721 (0.0005)	–1.2772 (0.0005)	–1.1397(0.0010)	–0.777 (0.0006)	–0.9973 (0.0007)	–1.1491 (0.0016)
	Yes	0.6781 (0.0012)	0.2376 (0.0006)	0.1279 (0.0006)	0.6973 (0.0010)	0.3226 (0.0007)	0.4017 (0.0009)	0.2166 (0.0018)
Birth country	Intercept	–2.0997 (0.0020)	–1.0360 (0.0012)	–0.9683 (0.0013)	–2.4134 (0.0017)	–1.1055 (0.0011)	–1.2938 (0.0012)	–1.2880 (0.0016)
	Yes	0.3901 (0.0022)	0.5110 (0.0012)	0.3536 (0.0013)	0.4499 (0.0018)	0.6743 (0.0012)	0.7649 (0.0014)	0.3754 (0.0019)
Death country	Intercept	–1.7927 (0.0013)	–1.0123 (0.0007)	–0.9112 (0.0007)	–1.5288 (0.0010)	–1.2557 (0.0008)	–1.4976 (0.0009)	–1.6462 (0.0015)
	Yes	0.6625 (0.0015)	0.4434 (0.0008)	0.2251 (0.0009)	0.7002 (0.0010)	0.3893 (0.0011)	0.4567 (0.0012)	0.2519 (0.0018)
No. of relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
No. of focal individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Note: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the negative binomial regression models with the number of relatives as offset to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals in the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative they are omitted from the regression model. Also included are the sample size of focal individuals employed for each model and the size of their kinship network.

Table A-8: Coefficients of the negative binomial regression models with number of relatives as offset to test association in terms of quality, by type of relative and demographic variable

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth date	Intercept	–1.1446 (0.0037)	–2.0640 (0.0069)	–2.5501 (0.0124)	–1.5306 (0.0039)	–1.3671 (0.0081)	–1.3671 (0.0090)	–0.7800 (0.0037)
	Non-missing month	0.6773 (0.0010)	0.7979 (0.0011)	0.6790 (0.0014)	1.2034 (0.0086)	0.6933 (0.0011)	0.6945 (0.0011)	0.3687 (0.0012)
	Birth period							
	1625–1649	0.1493 (0.0046)	0.1975 (0.0083)	0.3559 (0.0149)	0.0533 (0.0048)	0.1030 (0.0101)	0.0680 (0.0112)	0.0585 (0.0047)
	1650–1674	0.1872 (0.0043)	0.3425 (0.0077)	0.5730 (0.0135)	0.1084 (0.0043)	0.1614 (0.0092)	0.1798 (0.0100)	0.0719 (0.0044)
	1675–1699	0.2093 (0.0041)	0.5401 (0.0073)	0.7674 (0.0130)	0.1451 (0.0041)	0.2978 (0.0086)	0.2530 (0.0095)	0.0710 (0.0043)
	1700–1724	0.1986 (0.0040)	0.6443 (0.0071)	0.9832 (0.0127)	0.1508 (0.0040)	0.3500 (0.0086)	0.2707 (0.0093)	0.0793 (0.0042)
	1725–1749	0.2013 (0.0040)	0.7264 (0.0071)	1.1139 (0.0126)	0.1488 (0.0040)	0.4004 (0.0084)	0.2706 (0.0092)	0.1358 (0.0041)
	1750–1774	0.2300 (0.0039)	0.7409 (0.0071)	1.1847 (0.0125)	0.1362 (0.0040)	0.3999 (0.0082)	0.2577 (0.0092)	0.1684 (0.0040)
	1775–1799	0.2671 (0.0039)	0.7751 (0.0070)	1.1863 (0.0125)	0.1429 (0.0039)	0.3722 (0.0082)	0.2645 (0.0092)	0.1451 (0.0040)
	1800–1824	0.2035 (0.0039)	0.8253 (0.0069)	1.1819 (0.0124)	0.1583 (0.0039)	0.3578 (0.0082)	0.2754 (0.0091)	0.1690 (0.0039)
	1825–1849	0.2152 (0.0038)	0.9292 (0.0069)	1.2573 (0.0124)	0.1451 (0.0039)	0.4188 (0.0082)	0.2657 (0.0091)	0.2296 (0.0040)
	1850–1874	0.3044 (0.0038)	1.0436 (0.0069)	1.3688 (0.0124)	0.1584 (0.0039)	0.4510 (0.0081)	0.2748 (0.0091)	0.2572 (0.0041)
	1875–1900	0.3374 (0.0038)	1.1002 (0.0069)	1.5069 (0.0124)	0.2067 (0.0039)	0.4674 (0.0081)	0.3264 (0.0091)	0.2760 (0.0047)
	No. of relatives	9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
	No. of focal individuals	2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717

Table A-8: (Continued)

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Death date	Intercept	–0.9188 (0.0051)	–1.0783 (0.0065)	–1.1089 (0.0110)	–0.8673 (0.0056)	–0.7759 (0.0110)	–0.7041 (0.0107)	–0.7403 (0.0067)
	Non-missing month	0.4893 (0.0011)	0.5222 (0.0013)	0.4164 (0.0019)	0.5238 (0.0001)	0.3744 (0.0015)	0.3336 (0.0013)	0.3006 (0.0014)
	Death period							
	1625–1649	0.060 (0.0065)	0.0219 (0.0085)	0.0497 (0.0143)	0.0186 (0.0072)	0.0062 (0.0143)	0.0256 (0.0140)	0.0664 (0.0084)
	1650–1674	0.0760 (0.0060)	0.0278 (0.0079)	0.0703 (0.0133)	0.0591 (0.0065)	0.0344 (0.0131)	0.0280 (0.0126)	0.0607 (0.0077)
	1675–1699	0.1060 (0.0056)	0.0989 (0.0072)	0.1233 (0.01217)	0.0967 (0.0060)	0.0878 (0.0120)	0.0708 (0.0116)	0.0761 (0.0072)
	1700–1724	0.1061 (0.0055)	0.1646 (0.0070)	0.1933 (0.0117)	0.0997 (0.0059)	0.1038 (0.0115)	0.0833 (0.0112)	0.0801 (0.0071)
	1725–1749	0.1051 (0.0054)	0.1978 (0.0067)	0.2481 (0.0114)	0.0920 (0.0060)	0.1191 (0.0112)	0.0905 (0.0109)	0.1089 (0.0070)
	1750–1774	0.1291 (0.0053)	0.2047 (0.0067)	0.2841 (0.0112)	0.0920 (0.0057)	0.1208 (0.0111)	0.0944 (0.0109)	0.1493 (0.0069)
	1775–1799	0.1669 (0.0053)	0.1978 (0.0067)	0.2850 (0.0111)	0.0957 (0.0057)	0.1179 (0.0111)	0.1023 (0.0108)	0.1850 (0.0069)
	1800–1824	0.2027 (0.0052)	0.1983 (0.0066)	0.2793 (0.0111)	0.1156 (0.0056)	0.1136 (0.0110)	0.1226 (0.0108)	0.2234 (0.0068)
	1825–1849	0.2201 (0.0052)	0.2126 (0.0066)	0.2671 (0.01104)	0.1363 (0.0056)	0.1273 (0.0106)	0.1382 (0.0108)	0.2288 (0.0068)
	1850–1874	0.2321 (0.0052)	0.2588 (0.0066)	0.2884 (0.0110)	0.1553 (0.0056)	0.1552 (0.0.110)	0.1576 (0.0107)	0.2290 (0.0068)
	1875–1900	0.2336 (0.0052)	0.2927 (0.0065)	0.3364 (0.0110)	0.1710 (0.0056)	0.1786 (0.0110)	0.1670 (0.0107)	0.2350 (0.0068)
	No. of relatives	3,420,063	7,531,416	3,008,568	5,514,632	4,085,603	6,918,263	4,440,454
	No. of focal individuals	1,379,573	3,877,896	1,567,008	1,173,828	816,322	615,066	1,051,907

Notes: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the negative binomial regression models with the number of relatives as offset to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals in the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative they are omitted from the regression model. Also included are the sample size of focal individuals employed for each model and the size of their kinship network.

Table A-9: Coefficients of the binomial regression models to test association in terms of completeness, by type of relative and demographic variable

Demographic variable	Effect	Type of relative Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth year	Intercept	–1.0350 (0.0014)	–0.9092 (0.0022)	–0.5997 (0.0020)	–1.0632 (0.0018)	–0.4737 (0.0016)	–0.7065 (0.0013)	–0.8168 (0.0017)
	Yes	0.7105 (0.0019)	0.7606 (0.0009)	0.4027 (0.0020)	0.9637 (0.0018)	0.3885 (0.0016)	0.5634 (0.0013)	0.3790 (0.0018)
Death year	Intercept	–1.4161 (0.0013)	–0.7407 (0.0008)	–0.5721 (0.0008)	–1.1293 (0.0008)	–0.7641 (0.0006)	–0.9384 (0.0005)	–1.1632 (0.0012)
	Yes	0.6939 (0.0013)	0.4693 (0.0010)	0.2377 (0.0006)	0.6946 (0.0008)	0.3155 (0.0008)	0.3636 (0.0006)	0.2539 (0.0013)
Birth country	Intercept	–2.0997 (0.0020)	–1.0360 (0.0018)	–0.9683 (0.0017)	–2.4134 (0.0021)	–1.0932 (0.0009)	–1.2933 (0.0007)	–1.2887 (0.0010)
	Yes	0.3901 (0.0022)	0.5111 (0.0018)	0.3536 (0.0018)	0.4499 (0.0022)	0.6657 (0.0010)	0.7693 (0.0008)	0.3814 (0.0011)
Death country	Intercept	–1.2826 (0.0014)	–1.0123 (0.0008)	–0.9112 (0.0008)	–1.5067 (0.0007)	–1.2328 (0.0006)	–1.4424 (0.0005)	–1.7035 (0.0011)
	Yes	0.5144 (0.0015)	0.4434 (0.0011)	0.2251 (0.0011)	0.6910 (0.0008)	0.3752 (0.0009)	0.4231 (0.0007)	0.2966 (0.0013)
No. of relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
No. of focal individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Note: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the binomial regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals in the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative they are omitted from the regression model. Also included are the sample size of focal individuals employed for each model and the size of their kinship network.

Table A-10: Coefficients of the binomial regression models to test association in terms of quality, by type of relative and demographic variable

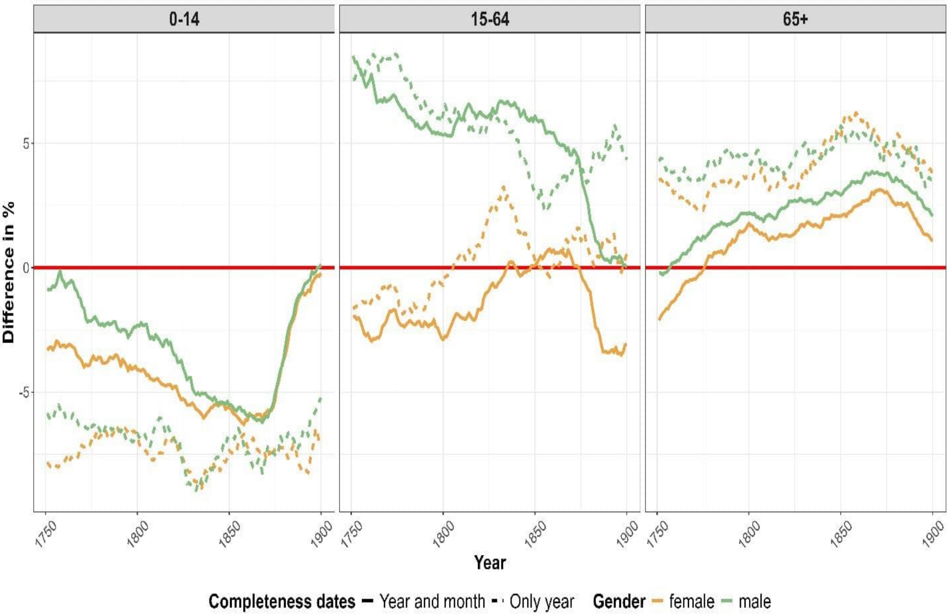
Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Birth date	Intercept	–1.1434 (0.0051)	–2.0721 (0.0010)	–2.5528 (0.0131)	–1.5421 (0.0049)	–1.3671 (0.0078)	–1.2277 (0.0075)	–0.7414 (0.0036)
	Non-missing month	0.6782 (0.0012)	0.8019 (0.0017)	0.6799 (0.0016)	1.2033 (0.0011)	0.6744 (0.0010)	0.6391 (0.00074)	0.3637 (0.0011)
	Birth period							
	1625–1649	0.1520 (0.0063)	0.1995 (0.0123)	0.3572 (0.0159)	0.0619 (0.0062)	0.0896 (0.0098)	0.0659 (0.0091)	0.0550 (0.0046)
	1650–1674	0.1891 (0.0063)	0.3393 (0.0112)	0.5728 (0.0144)	0.1145 (0.0055)	0.1478 (0.0088)	0.1708 (0.0081)	0.0570 (0.0043)
	1675–1699	0.2103 (0.0057)	0.5373 (0.0110)	0.7675 (0.0138)	0.1546 (0.0053)	0.2912 (0.0082)	0.2333 (0.0077)	0.0447 (0.0042)
	1700–1724	0.1982 (0.0055)	0.6422 (0.0104)	0.9805 (0.0135)	0.1580 (0.0052)	0.34134 (0.0081)	0.2536 (0.0076)	0.0495 (0.0040)
	1725–1749	0.2009 (0.0054)	0.7252 (0.0103)	1.1117 (0.0133)	0.1551 (0.0051)	0.3865 (0.0080)	0.2479 (0.0076)	0.1006 (0.0040)
	1750–1774	0.2315 (0.0054)	0.7415 (0.0102)	1.1851 (0.0133)	0.1468 (0.0050)	0.3812 (0.0079)	0.2262 (0.0075)	0.1216 (0.0039)
	1775–1799	0.2703 (0.0054)	0.7775 (0.0101)	1.1883 (0.0132)	0.1563 (0.0050)	0.3479 (0.0079)	0.2225 (0.0075)	0.0982 (0.0038)
	1800–1824	0.1992 (0.0053)	0.8306 (0.0101)	1.1855 (0.0132)	0.1777 (0.0050)	0.3314 (0.0079)	0.2242 (0.0075)	0.1328 (0.0038)
	1825–1849	0.2079 (0.0052)	0.9374 (0.0101)	1.2624 (0.0132)	0.1584 (0.0049)	0.3914 (0.0078)	0.2120 (0.0075)	0.1944 (0.0039)
	1850–1874	0.2999 (0.0052)	1.0533 (0.0101)	1.3724 (0.0131)	0.1687 (0.0049)	0.4224 (0.0078)	0.2321 (0.0075)	0.2196 (0.0041)
	1875–1900	0.3375 (0.0054)	1.1071 (0.0100)	1.5080 (0.0131)	0.2169 (0.0049)	0.4395 (0.0078)	0.2867 (0.0075)	0.2373 (0.0049)
	No. of relatives	9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
	No. of focal individuals	2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717

Table A-10: (Continued)

Demographic variable	Effect	Type of relative						
		Child	Parent	Grandparent	Sibling	Aunt and Uncle	Cousin	Grandchild
Death date	Intercept	–0.9210 (0.0097)	–1.0812 (0.0116)	–1.1102 (0.0131)	–0.8664 (0.0096)	–0.7776 (0.0140)	–0.6806 (0.0126)	–0.7433 (0.0083)
	Non-missing month	0.4897 (0.0021)	0.5226 (0.0023)	0.4167 (0.0023)	0.5238 (0.0017)	0.3746 (0.0019)	0.3234 (0.0015)	0.3016 (0.0018)
	Death period							
	1625–1649	0.0624 (0.0125)	0.0213 (0.0151)	0.0490 (0.0172)	0.0173 (0.0124)	0.0062 (0.0143)	0.0229 (0.0164)	0.0675 (0.0105)
	1650–1674	0.0754 (0.0114)	0.0258 (0.0079)	0.0689 (0.0160)	0.0568 (0.0113)	0.0340 (0.0167)	0.0235 (0.0147)	0.0605 (0.0096)
	1675–1699	0.1062 (0.0107)	0.0984 (0.0128)	0.1219 (0.0147)	0.0956 (0.0104)	0.0885 (0.0152)	0.0652 (0.0135)	0.0765 (0.0091)
	1700–1724	0.1061 (0.0055)	0.1646 (0.0124)	0.1927 (0.0141)	0.0983 (0.0101)	0.1047 (0.0147)	0.0751 (0.0131)	0.0801 (0.0089)
	1725–1749	0.1058 (0.0104)	0.1933 (0.0124)	0.2476 (0.0137)	0.0891 (0.0010)	0.1199 (0.0143)	0.0827 (0.0128)	0.1095 (0.0088)
	1750–1774	0.1045 (0.0102)	0.2041 (0.0120)	0.2837 (0.0135)	0.0894 (0.0098)	0.1213 (0.0141)	0.0852 (0.0127)	0.1508 (0.0087)
	1775–1799	0.1294 (0.0101)	0.1991 (0.0119)	0.2845 (0.0134)	0.0938 (0.0098)	0.1181 (0.0141)	0.0917 (0.0127)	0.1876 (0.0086)
	1800–1824	0.1696 (0.0100)	0.2005 (0.0118)	0.2795 (0.0133)	0.1148 (0.0097)	0.1144 (0.0140)	0.1108 (0.0127)	0.2267 (0.0086)
	1825–1849	0.2060 (0.0100)	0.2156 (0.0117)	0.2680 (0.0132)	0.1361 (0.0097)	0.1287 (0.0140)	0.1256 (0.0126)	0.2317 (0.0085)
	1850–1874	0.2229 (0.0010)	0.2630 (0.0117)	0.2902 (0.0132)	0.1553 (0.0096)	0.1574 (0.0139)	0.1448 (0.0126)	0.2290 (0.0068)
	1875–1900	0.2348 (0.0097)	0.2969 (0.0116)	0.3393 (0.0132)	0.1702 (0.0096)	0.1810 (0.0139)	0.1551 (0.0126)	0.2375 (0.0085)
	No. of relatives	3,420,063	7,531,416	3,008,568	5,514,632	4,085,603	6,918,263	4,440,454
	No. of focal individuals	1,379,573	3,877,896	1,567,008	1,173,828	816,322	615,066	1,051,907

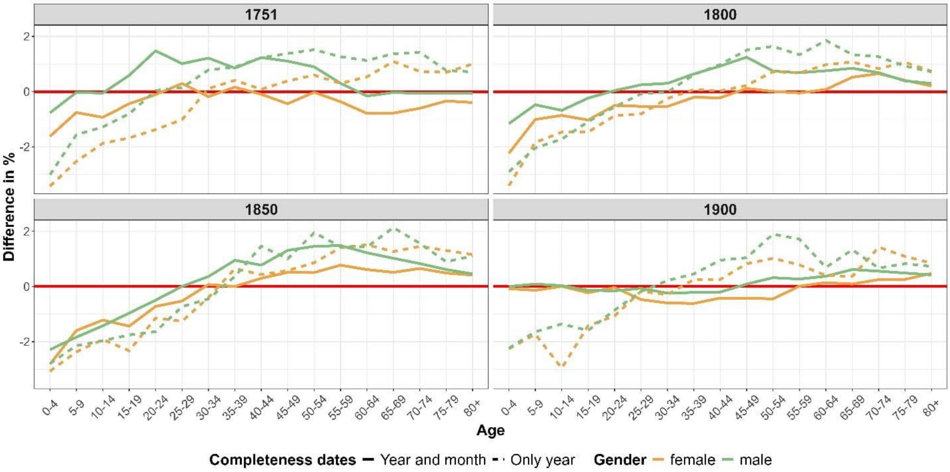
Note: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the binomial regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals in the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative they are omitted from the regression model. Also included are the sample size of focal individuals employed for each model and the size of their kinship network.

Figure A-4: Difference between the age–sex distribution in percentage between the Swedish population from FamiLinx by quality level (precise birth and death dates against at least one non-precise date) and the registered Swedish population over the historical period 1751–1900



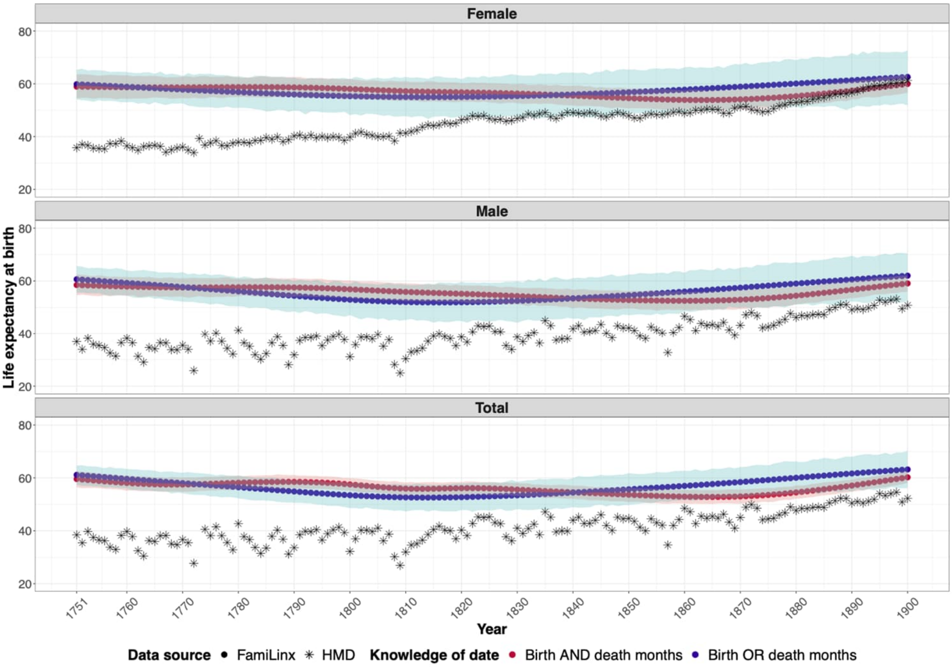
Note: Three wider age groups were considered (0–14, 15–64, 65+). The solid line refers to individuals with non-missing birth and death months. The dotted line indicates individuals having at least either a birth or death month missing. Yellow lines refer to female individuals, green lines refer to male individuals.

Figure A-5: Difference between the age–sex distribution in percentage between the Swedish population from FamiLinx by quality level (precise birth and death dates against at least one non-precise date) and the registered Swedish population over the years 1751, 1800, 1850, and 1900



Note: The solid line refers to individuals with non-missing birth and death months. The dotted line indicates individuals with at least either birth or death month missing. Yellow lines refer to female individuals, green lines refer to male individuals.

Figure A-6: Life expectancy at birth in Sweden for the period 1751–1900, by sex and quality level (precise birth and death dates against at least one non-precise date) in FamiLinx and Swedish life expectancy at birth from the HMD



Note: Red lines refer to estimates of life expectancy at birth calculated for Swedish individuals with non-missing birth and death months. Blue lines denote estimates of life expectancy at birth among Swedish individuals with missing birth or death months. Star-shaped points denote life expectancy estimates from the HMD. 95% confidence intervals obtained using Monte Carlo simulations.