



<http://www.diva-portal.org>

This is the published version of a paper published in *Historical Life Course Studies*.

Citation for the original published paper (version of record):

Engberg, E., Westberg, A., Edvinsson, S. (2016)

A Unique Source for Innovative Longitudinal Research: The POPLINK Database.

Historical Life Course Studies, 3: 20-31

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-118448>

HISTORICAL LIFE COURSE STUDIES

VOLUME 3
2016



MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles

This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles

This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, <http://www.esf.org>), the Scientific Research Network of Historical Demography (FWO Flanders, <http://www.historicaldemography.be>) and the International Institute of Social History Amsterdam (IISH, <http://socialhistory.org/>). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at <http://www.ehps-net.eu/journal>.

Editors: Koen Matthijs & Paul Puschmann
Family and Population Studies
KU Leuven, Belgium
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.



The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: <http://www.ehps-net.eu>.



HISTORICAL LIFE COURSE STUDIES
VOLUME 3 (2016), 20-31 published 15-03-2016

A Unique Source for Innovative Longitudinal Research: The POPLINK Database

Annika Westberg
Umeå University

Elisabeth Engberg
Umeå University

Sören Edvinsson
Umeå University

ABSTRACT

This paper presents the longitudinal database POPLINK, which has been developed at the Demographic Data Base at Umeå University, Sweden. Based on digitized Swedish population registers between c. 1700-1950, the database contains micro-data that covers the agrarian society through industrialization and further on to the Swedish welfare state and contemporary society. It is now possible to study the profound processes of the second demographic transition using individual level data with a proper size population. POPLINK allows for a large array of longitudinal studies, such as social mobility, migration, fertility, mortality, civil status, kinship relations, diseases, disability and causes of death. International standards of occupations (HISCO) and diseases (ICD-10) have been applied, facilitating comparability. POPLINK covers two large regions in Northern Sweden and is built on complete registrations. It is one of the world's most information-dense historical population databases, covering up to 15 generations and 350,000 individuals described by 300 variables, allowing the ability to monitor populations over time. POPLINK has been built to allow linkage to modern registries, clinical data and medical biobanks, which enables the study of transgenerational effects, heredity and genetic transfers in disease incidence of the population today. DDB serves as an infrastructure for research and is open to researchers of any nationality.

Keywords: POPLINK database, Longitudinal Research, Micro-data, Sweden, Historical Population Database, Record Linkage, Infrastructure, Demographic Data Base, CEDAR

e-ISSN: 2352-6343

PID article: <http://hdl.handle.net/10622/23526343-2016-0003?locatt=view:master>

The article can be downloaded from [here](#).

© 2016, Annika Westberg, Elisabeth Engberg, Sören Edvinsson.

This open-access work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/by/4.0/), which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/>

1 INTRODUCTION

Large population databases with access to longitudinal micro-data have over the years been invaluable tools in our efforts to understand the social, economic and demographic processes that changed the 18th- and 19th-century societies. The effect of industrialization, the driving forces behind the mortality decline, and long-term trends in migration are just some examples of issues which cannot be fully understood without access to large scale data on the individual and household level of society. Since the 1970s and forward, significant contributions have been made within these fields thanks to massive and groundbreaking digitization projects, such as the Demographic Data Base in Sweden, the Historical Sample of Netherlands, the Utah Population Database and the IPUMS database in the United States. When it comes to the 20th century, the situation has been different for a long time. Register-based population data is available from the 1950s and forward, but not for the first half of the century (United Nations Economics Commission for Europe 2007). Without access to substantial databases with micro-data, researchers have been required for a long time to use aggregate statistics and small-scale data to study socio-economic and demographic change during this significant period in Nordic and European history.

With a long history of national and mandatory registration, Sweden has developed population registers, which are unique in an international perspective, in terms of coverage as well as information. The registers cover the total population, from the late 17th century until today, and provide data on individuals and households. For the 18th and 19th century, there is good access to longitudinal population data on the individual level, mainly due to the efforts of the Demographic Data Base (DDB) at Umeå University and the Centre for Economic Demography at Lund University.¹

In 2008 the DDB started digitizing 20th century data from two regions in northern Sweden. The objective was to create a new asset for micro-level analysis of the processes that transformed society, through the demographic transition and beyond, by bridging the present gap between historical population databases and modern registers. Today this new resource, the population database POPLINK, provides access to micro-level data from the time period 1700-c. 1950, from the agrarian society through industrialization and further on to the Swedish welfare state and contemporary society.

The aim of this paper is to offer an overview of the POPLINK database, present issues associated with building a large infrastructure and to discuss the strategic value of increased access to 20th century microdata within a number of research areas.

2 DESIGN

Building a new large database is a substantial undertaking, requiring careful consideration from an infrastructural perspective, along with ample time and resources. Researchers, on the other hand, whose requests might have initiated the project, eagerly wait to get hold of the data. Trying to reduce the time between data entry and release of the data, the first version of POPLINK was built by adding new data to an already existing database, in this case the database POPUM at the DDB. Supplementary data from five geographical areas included in the historical database POPUM at the DDB (Brändström, Vikström & Edvinsson 2006; Edvinsson 2000) was digitized for the period 1900-1950, and later linked to previous data from the 18th and 19th centuries. This resulted in a database spanning a 300-year period covering up to 15 generations. The linkage between old and new data worked almost seamlessly (Engberg 2008), which proves that databases built according to the internationally recognized principles of best practice actually can be extended while maintaining their quality and consistency (Mandemakers & Dillon 2004).

Since POPLINK is explicitly designed to serve as a bridge between historical and modern registers, there are excellent preconditions for a high quality linkage to the numerous modern official registers that have been developed since the 1950s and onward. A linkage of Swedish population registers at Statistics Sweden makes it possible to extend the data even further, including also the present population. With a similar linkage of the registers at the Swedish National Board of Health and Welfare, health and lifestyle data can be added to the datasets. The combination of a major increase in

1 Demographic Data Base: <http://www.cedar.umu.se/english/?languageId=1>, The Scanian Economic Demographic Database: <http://www.ed.lu.se/databases/sedd>

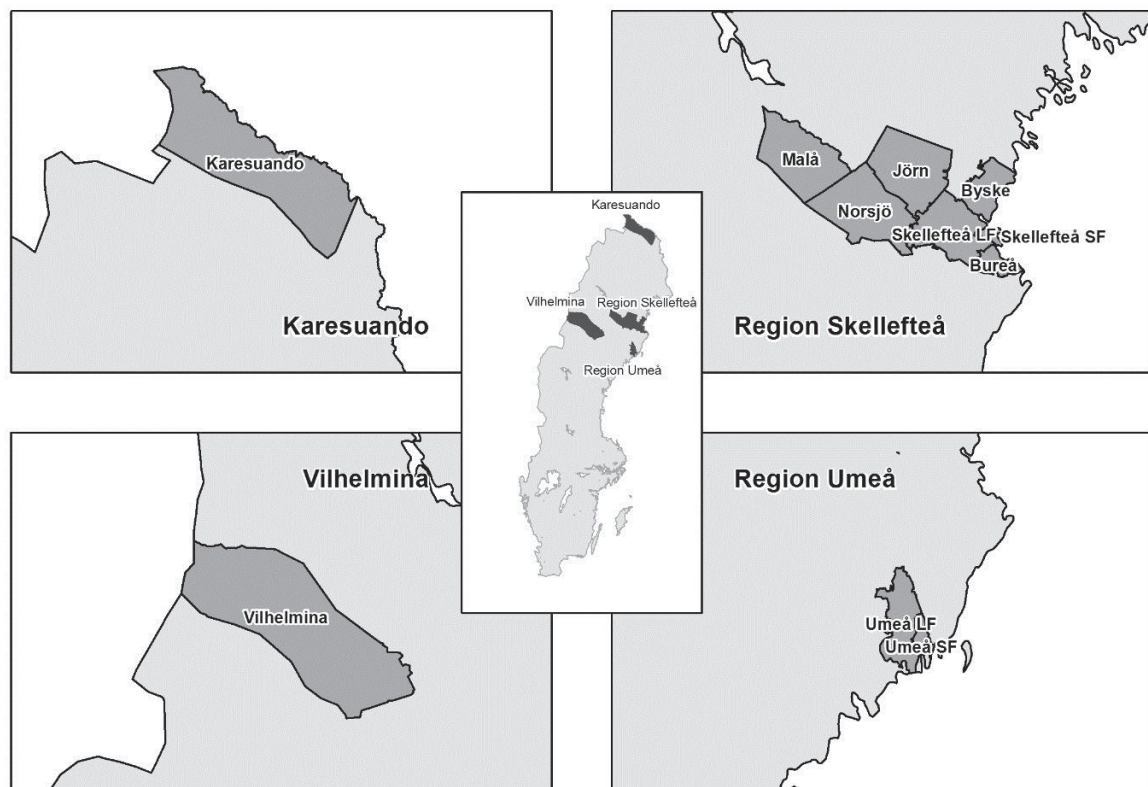
access to 20th century micro-level data, high-quality linkage between long genealogies for a substantial population, population registries and health data is what makes POPLINK such a powerful resource for the scientific community.

2.1 TARGET POPULATION

In a high-quality longitudinal database, the number of complete genealogies over generations is a critical element. This makes geographical regions with a low population turnover particularly interesting for projects of this character. The province of Västerbotten in northern Sweden, which has been chosen as the target population in POPLINK, is one such region. Västerbotten is characterized by geographically large parishes and by a population with very low mobility, which was mainly manifested in short-distance migration. An additional strong, and scientifically significant argument for focusing on this particular part of Sweden is that large parts of the population in the area are already mapped in a number of other existing high-class registries and the biobanks at the Department of Biobank Research in Umeå². One of these population-based resources is the Västerbotten Intervention Programme, one of the world's largest ongoing population-based health surveys, which includes samples, health data and lifestyle variables on the individual level for almost all 40, 50 and 60-year olds in the province (Norberg, Wall, Boman & Weinehall 2010).

The target population in POPLINK covers two large regions in the province of Västerbotten: Umeå and Skellefteå with adjacent parishes, as well as the parishes Vilhelmina and Karesuando in inland and Northern Sweden. Longitudinal and multigenerational data is available for 226,000 individuals in the Skellefteå region between ca. 1680 – 1950. The Umeå region covers data for 114,000 individuals between c. 1900 – 1950. In total, there are linked detailed life course data for c. 392,000 individuals and around 2.2 million records. A third region with inland and coastal parishes, which will eventually link the Umeå and Skellefteå regions together, is currently under construction. New data from this region is continually added to the database, as data production is ongoing. The long-term objective is that POPLINK will include longitudinal population data for the whole province of Västerbotten, from 1700-1950.

Figure 1 Map of Sweden and of regions included in the POPLINK database



2 Department of Biobank Research, Umeå University <http://www.biobank.umu.se/biobank/?languageld=1>

Table 1 *Timespan and geographical setting of POPLINK (November 2015)*

	Period	Individuals	Records
Skellefteå region	1680-1950	225,874	1,414,720
Umeå region	1900-1950	113,943	499,771
Karesuando	1700-1950	12,853	48,046
Vilhelmina	1700-1950	40,236	238,896
N		392,906	2,201,433

Source: *Demographic Data Base, Umeå University*

2.2 COMPLETE AND INCOMPLETE LIFE COURSES

The individuals in POPLINK are observed and followed as long as they stay within the region. If they move to geographical areas not covered in the DDB databases, they are out of observation until they return again. An advantage of this approach is that there is complete coverage of the population that stays within the boundaries of a region, making it possible to define and study the population at risk in the area. The disadvantage, from a life course perspective, is that there are incomplete life courses. Working with large regions instead of individual parishes is one way of reducing this problem. Covering a cluster of neighboring parishes usually captures a large number of short-distance migrations and does, in most cases, also account for administrative changes of parish borders, affecting the population at risk. The implementation of an automatic linkage system, making it possible to link individuals over geographical and administrative borders, has significantly contributed to improvements in the quality of the data. Another advantage is that the digitizing process starts from the beginning of registration. All individuals in the registers are recorded, not only those with ancestors or relatives. In some population databases, digitization starts from the present population and builds genealogies moving backward in time. This means that there is a built-in attrition of individuals that never had children or who died before the age of reproduction.

There are at least six different types of life biographies in longitudinal databases, which are characterized by different degrees of completeness in terms of observation and records of birth and death:

- a) Complete life courses: Records of birth and death and continuous observation in between these events.
- b) Complete life courses with gaps in observation: There are records of both birth and death, but with one or more gaps in observation in between.
- c) Right-censored: Record of birth and a period of continuous observation, followed by out-migration from the region. No record of death.
- d) Left-censored: No record of birth. In-migration to the region, a period of continuous observation and a record of death.
- e) Right- and left- censored, no gaps: There are no records of birth or death but the individual is continuously observed over a large part of the life-course
- f) Right- and left- censored with gaps: There are no records of birth or death and gaps in observation over the life-course.

In the table below, a compressed overview of the distribution of the different kinds of biographies is presented for four birth cohorts: 1751-1800, 1801-1850, 1851-1900 and 1901-1950. Since 1950 represents the end of registration of the data in POPLINK, the number of right-censored life courses, with no record of death, will be significantly over-represented in the last cohort.

Table 2 Overview of life biographies in the Skellefteå region, distributed on birth cohorts

Type of biography	1751-1800		1801- 1850		1851-1900		1901 – c. 1950*	
	n	%	n	%	n	%	n	%
Complete (a)	4,543	38	13,090	51	22,564	33	9,019	8.5
Complete with gaps in presence (b)	1,379	12	2,775	11	2,935	4	236	0.2
Right-censored (c)	1,010	8.5	2,913	11	22,446	33	61,961	57.5
Left-censored (d)	1,494	12.5	2,473	10	3,238	5	588	0.5
Right- and left censored (e and f)	532	29	4,359	17	17,263	25	36,120	33.5

* End of registration

The highest proportion of complete life courses is found in the birth cohorts 1750-1800 and 1801-1850, where 50 and 62 percent of the individuals, respectively, can be followed from the cradle to the grave, although for some with minor gaps in their life spans. The subsequent decline has not yet been analyzed in detail, but is most likely explained by an increased mobility in the late 19th and early 20th centuries. The 1901-1950 cohort is, as already noted, biased by the end of registration around 1950. In the 1851-1900 cohort there is also a substantial increase in the number of individuals without death records, and individuals lacking both birth and death records, which can be interpreted as individuals moving in and moving out again. The proportion in category d, in-migrants that can be observed in the region until their death, on the other hand, appears to be more stable over time. Also these changes require a more detailed analysis to be fully explained.

2.3 CONTENTS

The POPLINK database is based on the rich information in the Swedish parish registers, which have been kept since the late 17th century, and until 1990 served as the official system of civil registration (Nilsson 1993). The registers were kept locally, with the parish as the administrative unit. Separate records were maintained for births, marriages, death and migration. In addition to these event registers, longitudinal parish registers, *husförhörslängder*, were also kept, accounting for all individuals present within a parish over time. They were continually updated with events and attendance and hence a dynamic source. The individual's attendance at communion, as well as his or her progress in literacy and knowledge in the catechesis was annually recorded, which makes it possible to determine the population at risk at a given point in time.

The composition of the sources makes it easy to follow individuals through the different phases of life: from the family of origin through the start-up of occupational careers in adolescence, and eventually, the establishment of new family units. Events such as births, marriages, deaths and migrations are registered in separate registers, but are also noted in the family-based longitudinal parish registers. When a new individual is born, the name and date is carefully added to the list of family members in the register. And in reverse, when individuals were leaving the family group, by death or by migration, their names were crossed out in the register. The presence of this dynamic source is particularly valuable for life-course studies. The historically important circulation of working adolescents and servant migration can be studied in detailed, following individuals between households (Dribe & Lundh 2005; Kok 2007). Occupational and residential careers can be reconstructed, tracing the presence and impact of family ties, as well as intergenerational co-residence (Bras et al 2010, Fusé 2008, Janssens 1993).

Defining families is rather straightforward, thanks to the detailed kinship information in the sources. Defining co-resident kin is also quite unproblematic since spouses and children are recorded together in the parish registers. The main principles when defining a family are (1) that they are written on the same page in the source on subsequent rows and (2) they are related by partnership or (3) being a child (biological or non-biological) to someone on a row immediately before.

However, there is a shortcoming to the parish registers when it comes to determining members of households. Typically, households are defined as individuals who live in the same dwelling, sharing meals or living accommodation. A household may also include individuals outside immediate kin, such as servants or lodgers. Since servants usually are registered at the bottom lines on the pages of the parish registers, it cannot always be determined to which household each individual belonged,

particularly in cases when more than one family are registered on the same page.

The records are often very detailed. There is information about place and date of birth and baptism, marriage, death and burial, legitimacy and civil status. Information is provided about midwife's attendance at childbirth, vaccination, disease and disabilities and causes of death. Other things that can be studied are occupations, literacy, delinquency, socio-economic status, and kinship relations. Relations are defined by their biological and non-biological nature. The non-biological relations give information on foster or step relations as well as adoption and orphanhood.

Transferring the information from parish registers to a database is however not without problems. One aspect of this is that the contents and structure of parish registers changed over time. 17th and 18th centuries sources have less precise information, for example not presenting exact dates but only a year. During the 19th century, the parish registers became more and more standardized and the differences between individual ministers on how to keep the books diminished. Still, important information can be missing, either due to how the books were kept or sometimes due to completely missing sources, thereby making some types of studies impossible. Earlier studies have, for example, showed how missing birth and death records cannot be compensated by data in parish registers for studies of fertility and infant mortality (Edvinsson 1992). Furthermore, incomplete and inconsistent information complicates record linkage, as well as construction of variables that require a combination of information. These problems have not always been completely solved, but the use of the database has been facilitated by constructing specific tables for some of the vital variables, for example residence periods, marital status (with changes), with detailed rules for how the variables are constructed.

Although the Swedish records and databases are exceptional in many ways, the aspiration is to make them comparable with other similar databases around the world. For this reason, international standards become important. Occupations are coded into HISCO (Historical International Standard Classification of Occupations, cf. van Leeuwen, Maas & Miles (2002, 2004)), to facilitate comparable research on occupations and social mobility, and cause of death is translated into the WHO standard ICD-10 (WHO International Statistical Classification of Diseases and Related Health Problems). Finally, DDB has implemented the standardized IDS (Intermediate Data Structure), cf. Alter & Mandemakers (2014) for parts of the pre-1900 data. The plan is that the IDS database will increase substantially within a few years' time.

2.4 SECURE LINKAGE

In order to produce a longitudinal multigenerational database, a high quality record and relational-linkage system is essential. DDB is linking individual data, such as birth, death, migrations and occupations, to construct life biographies, and relations – spouses, parents and children – to construct genealogies. Establishing biological relations is vital in issues related to heritability, and this has rendered extra attention in the linkage process.

In the international community of database owners and researchers, there has been an ongoing debate on what strategy is most effective and accurate when conducting linkage: automatic or semi-automatic. The DDB has evaluated both methods, and given the character of the sources, the robust conclusion is that a combination of both produces a very high quality. The automated data-linking process at the DDB manages to link 95 % of all records, according to strict rules. To handle the rest, a semi-automatic step follows, taking care of records that either were not handled by the automatic linkage application, or were flagged by the system as needing manual attention. This concerns, for example, when the automation has detected gaps in life biographies that are not associated with matching migration records (Larsson & Engberg 2015; Wisselgren, Edvinsson, Berggren & Larsson 2014).

Access to mid-20th century data has made it possible to compare the kinship links in the DDB data with the official Swedish registers, with excellent results (Engberg 2008; SCB 2007). Out of a test sample of 2555 individuals, 2093 (82%) were successfully linked to the Multi-Generation Register (MGR) at Statistics Sweden on the basis of their civil registration number. The large majority of the remaining 18% of the sample did not belong to the target population of the MGR, and were for that reason not possible to link. Only three individuals could not be linked, due to problems with the data. The quality of the links was also very high; the correspondence of kinship information between the two registers reached 97 percent. The high correspondence rate between the registers is an important

quality stamp, vouching for a secure linkage between datasets from the POPLINK database and other civil registration based research registers. Similar methods have also been successfully used for linkage to Swedish census data (Wisselgren, Edvinsson, Berggren & Larsson 2014).

Methods for handling linkage between data in POPLINK and other registers have been developed in collaboration with Statistics Sweden, within the Swedish legal framework concerning protection of privacy (Engberg 2008). Linkage to national level registries can include, for example, the National Board of Health and Welfare that keeps the Cause of death register, Cancer register, In-patient register and the Prescribed drug register, or the many different registries held at Statistics Sweden. Large synergies can be created with the combination of POPLINK and the numerous research databases at Umeå University, covering large parts of the target population in POPLINK (Malmberg, Nilsson & Weinehall 2010).

3 NEW POSSIBILITIES FOR RESEARCH

Data from the DDB has been used in research during more than 40 years (Brändström 1984; Edvinsson & Lindkvist 2011; Engberg 2005; Junkka & Edvinsson 2015; Maas & van Leeuwen 2002; Vikström 2003). The main fields of research has been within historical demography, but the data has proven to be of great value for many other disciplines and for a diversity of research questions. The continuous extensions and improvements have further provided the possibility for new research questions and methods, all of course in a reciprocal development with trends and new directions in the research society. In the following we highlight some research fields that we believe are either under-researched or have been made possible, or substantially facilitated, by the recent improvements and additions in the database.

On a general level, the strengths and possibilities of POPLINK for research and the application of research questions can be described in different dimensions, as will be apparent from the discussion below. Perhaps the most rewarding new opportunities relate to studies in medicine, for example within (historical) epidemiology and genetics, which is not a completely new area for historical population data, but is now becoming more relevant with POPLINK.

Another strength of the database is that it presents people in their social context, allowing network studies on the impact of family circumstances and the role of different types of social networks. Furthermore, the potential for spatial analyses is good since the registers include geographical information on residence and migration. There are also excellent preconditions to study trans-generational transfers and processes, such as fertility and mortality patterns through many generations (Alter 2013).

With the new data in POPLINK, the first half of the 20th century, a time of vital scientific interest in Swedish history, is for the first time opened up for large-scale research. Data on the individual level has been scarce for this period. It has until now only been possible to study fundamental processes, such as the fertility and mortality decline using aggregate statistics. Such figures are, at best, available on a provincial level, but frequently only as national figures. The period is characterized by profound social, demographic, economic and political change and dramatic improvements in health. People received democratic rights, and Sweden changed from a mainly agrarian society into a modern industrialized welfare state with radically improved health and living conditions. When it comes to demography, the Swedish 20th-century development is of particular interest. The demographic transition resembled the European pattern in many ways, but Sweden and the Nordic countries have in many cases been in the forefront of the development (Edvinsson, Gardarsdottir & Thorvaldsen 2008). It is probably no coincidence that later in the 20th century Sweden became the most salient example of a society characterized by the behaviors, attitudes and values of the second demographic transition, for example by the increased practice of cohabitation as well as changes in family formation and fertility (Lesthaeghe 2010; Van de Kaa 1986). The connection to the Nordic welfare state is of special interest in this case. Swedish mortality rates fell continuously from the early 19th century, making Sweden have among the lowest rates in the world. The early 20th century was a crucial period in this development, and for infant and early childhood mortality the shift was never as dramatic as from 1940 onward.

Household size and family structures changed, fertility rates declined and women's position in society was markedly altered. With POPLINK, it will be possible to conduct in-depth studies of these large-scale transformations in a longitudinal perspective, and to achieve a better understanding of the finer mechanisms behind the 20th century population changes and present challenges, such as the ageing population.

With the possibilities of linkage to other Swedish registry resources as well as clinical data, POPLINK will be a useful resource for the study of major epidemiological issues, such as interactions between lifestyle, socioeconomic and cultural conditions in past generations. Longitudinal data in POPLINK can, for example, be used to improve our understanding of the still present epidemiological transitions within populations, which today is a global issue of high relevance. Although theoretical constructs for these processes have been considered over the past 40 years, there is still a lack of evidence-based understanding, particularly as to how past transitions may illuminate future processes in less-developed settings. Historical population data from Sweden and more recent INDEPTH data from developing countries can be brought together for comparisons and analyses, which can elucidate the process of the epidemiological transition from an evidence-based perspective.

Regarding research approaches and analytical methods, the longitudinal population data in POPLINK is ideal for life course studies (Kuh, Cooper, Hardy, Richards & Ben-Schlomo 2014). The linked information at the individual level provides a large amount of full or partial life biographies, which allows large-scale studies of individuals and families from a life course perspective. The life course approach has developed during the last decades, significantly improving our understanding of how health, occupational careers, social life and so on are constructed over the life span. In order to understand the lives of people, we need to consider not only the present conditions, but also previous experiences throughout life. This is obvious, for example, when it comes to early life experiences and their effects on health and survival later in life. Concomitant with the interest in life course approaches, several methods have been developed for this type of analysis, for example event history analysis where transitions between states are measured (Broström 2012). Another type of statistical method is sequence analysis, where the complete life courses and individual trajectories are analyzed as separate threads, providing information on what you might call life patterns based on similarities in chosen dimensions between sequences, when it comes to timing and order of states (Svensson, Lundholm, de Luna & Malmberg 2015).

Table 3 *Generational structure in the POPLINK database*

Approximate birth year of known ancestor	Generational depth (<i>N</i> generations)	Individuals (<i>N</i>)	Proportion of total cohort (%)
1875-1900	2	15,290	15.0
1825-1850	3	10,699	10.5
1825-1850	4	7,526	7.4
1800-1825	5	9,020	8.8
1780-1800	6	7,922	7.8
1760-1780	7	7,970	7.8
1740-1760	8	11,204	11.0
1720-1740	9	16,474	16.1
1700-1720	10	14,531	14.2
1680-1700	11	1,540	1.5
1660-1680	12	1	0

The time dimension does in some cases also include the period before birth. During recent years, interest in influences that extend the individual life course has increased. A mediating approach in this case is the study of possible effects of conditions during gestation, for example that nutrition during certain periods in utero can lead to increased risks of cardiovascular diseases in adulthood (Barker 1995). In such cases the health of the unborn child is strongly connected to the mother. The pathways can however be extended to previous generations. Fertility, family size, age of marriage and infant mortality are just some of the family factors which exhibit distinct traits of social inheritance,

and we often find that these patterns can follow families over time (Kolk 2011; Vandezande 2012). Intergenerational social mobility is another interesting field in which POPLINK data can be used. The large generational scope of 12 generations, in POPLINK allows the study of social inheritance, causes of transmissions of demographic and socioeconomic patterns between generations, for example intergenerational social mobility, which is difficult to study with panel and census data.

When it comes to transgenerational effects, they have mainly been studied from a health perspective, focusing on heredity and genetic transfers which identify genetic disorders that cause poor health and premature death. But there might also be other traits that can be transferred between generations. For a long time, there has been a wide gap between the social science and the genetic approach, or between nature and nurture. This conflict has however diminished during recent years as genetic research has increasingly come to acknowledge the strong interdependence between environmental influences and genes and, on the other side, many social scientists being more open to the new developments within biology and genetics. This is particularly the case when it comes to the new advances in epigenetics and the possible transfers across generations (Meloni 2014; Pembrey, Saffery, Bygren, *Network in Epigenetic Epidemiology* 2014; Rose 2013).

The multigenerational scope and the detailed information on kinship in POPLINK will allow researchers to develop these new perspectives on heritability, intergenerational patterns in disease incidence, and the interaction between different risk factors over long time periods. Observations of large prospective cohorts can for example be supplemented by retrospective studies, since the historical registers enable the inclusion of lifestyle factors in earlier generations in the analysis. This is considered particularly useful within the field of genetic epidemiology, where the availability of long family histories, access to biobank data and modern sequencing techniques gives rise to groundbreaking studies of gene-lifestyle interactions and disease outcomes, for instance the development of type 2 diabetes. Swedish population data from the 18th and 19th centuries have already successfully contributed to top-quality research within this area.

Studies of early life conditions and their importance for the outcomes of disease have, for example, suggested that a reduced availability of food during 19th century famines gave rise to epigenetic changes, affecting the longevity of future generations (Bygren, Kaati & Edvinsson 2001; Pembrey et al. 2006). Population data from the DDB has proved to be useful as a testbed for studying the influence of endogamy and consanguinity on genetic disorders (Egerbladh & Bittles 2008; Lundevaller & Edvinsson 2011). With an infrastructure like POPLINK, and its unique capacity to integrate high quality population data with modern registry resources, competitive studies like this can be achieved on a much larger scale.

4 ACCESS FOR RESEARCH

Ever since the start of the Demographic Data Base in 1973, the intention has been to build a research database for general purposes, rather than with specific projects in mind, as a resource for the wider research community. In 1977, this intent was confirmed in a state regulation in which the Demographic Data Base was given a national commission to *collect, process and release population data for scientific, educational and archival purposes; and to promote scientific cooperation and method development*. From the 1970s and forward the use of DDB data within the social sciences, the humanities, medicine and life sciences is reflected in more than 900 titles of published research.

As a consequence of its role as a national infrastructure, open to researchers nationally as well as internationally, the DDB has developed a well-functioning organization for comprehensive user support and service. The databases are vast and complex, but from the very first contact, the researcher receives guidance and assistance by academically experienced staff. Customized datasets, complying as closely as possible to the researchers' requirements, are produced and distributed through specified contracts, and for POPLINK only in de-identified form. In order to maintain high quality retrievals, all datasets are being tested and verified before they are delivered to the researcher. Since the variety of variables and information frequently make the retrievals complex, all data sets are accompanied by a detailed documentation in order to give the researcher the best basis for performing analyses.

5 ETHICS

Handling individual-level data stretching as far as into the 1950's raises several ethical issues. Sweden has a strict legal framework concerning the protection of privacy, with implications for the data production process as well as for the release of data for research. One of the requirements is that personal information in the sources, particularly variables that are considered private and sensitive according to the Swedish Data Act (Personuppgiftslagen), have to be omitted from digitalization. This concerns, for example, information about race, ethnicity, political and religious beliefs and information referring to sexual issues.

The physical and technical measures for data protection are high. The entire data production process, including all stages of linkage described above, takes place in closed premises and is executed in closed computer networks, without connection to other internal or external networks. All personnel involved in the project are further obliged to sign an extended commitment of professional secrecy.

After processing, the POPLINK database is safeguarded with a double layer of physical security and access authentication. When data is released, it is in anonymized form, without names, addresses or other information that allows the identification of single individuals. All data retrievals from the POPLINK database require the approval from an ethical vetting board.

6 CONCLUSION

Longitudinal data on the individual level offer excellent possibilities to study long-term population dynamics on different levels in society, such as mobility, regional development and migration, as well as for simulations of these processes. With extensive and detailed life-course data, spanning up to 15 generations and 300 years, and a design allowing linkages to other registers, POPLINK is a powerful tool for life course studies and an asset to many areas of research.

REFERENCES

- Alter, G. (2013). Generation to Generation: Life Course, Family, and Community. *Social Science History*, (37), 1, 1-26.
- Alter, G. & Mandemakers, K. (2014). The intermediate data structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1-26.
- Barker, D.J.P. (1995). Fetal origins of coronary heart disease. *British Medical Journal*, 311, 171-174.
- Bras, H., Liefbroer, A. C. & Elzinga, C. H. (2010). Standardization of Pathways to Adulthood? An Analysis of Dutch Cohorts Born Between 1850 and 1900. *Demography* (47) 4, 1013-1034.
- Broström, G. (2012). *Event History Analysis with R*. London: Chapman & Hall.
- Brändström, A., Vikström, P. & Edvinsson, S. (2006). Longitudinal databases – sources for analyzing the life course: Characteristics, difficulties and possibilities. *History and Computing*, 14 (1 and 2), 109-128.
- Brändström, A. 1984. *'De kärlekslösa mödrarna.'* *Spädbarnsdödligheten i Sverige under 1800-talet med särskild hänsyn till Nedertorneå*. (Doctoral dissertation). Umeå: Demographic Data Base.
- Bygren, L. O., Kaati G. & Edvinsson, S. (2001). Longevity determined by paternal ancestors' nutrition during their slow growth period. *Acta Biotheoretica*, 49(1), 53-59.
- Dribe, M. & Lundh, C. (2005). Determinants of Servant Migration in Nineteenth Century Sweden. *Continuity and Change*, (20) 1, 53-91.
- Edvinsson, S. (1992). *Den osunda staden: sociala skillnader i dödlighet i 1800-talets Sundsvall* (The unhealthy town: social inequality regarding mortality in 19th century Sundsvall), Umeå, Demographic Data Base.

- Edvinsson, S. & Lindkvist, M. (2011). Wealth and health in 19th century Sweden. A study of social differences in adult mortality in the Sundsvall region. *Explorations in Economic History*, (48)3, 376-388.
DOI: [10.1016/j.eeh.2011.05.007](https://doi.org/10.1016/j.eeh.2011.05.007).
- Edvinsson, S., Gardarsdottir, O. & Thorvaldsen, G. (2008). Infant mortality in the Nordic countries 1780-1930. *Continuity and Change*, 23(3), 457-485.
- Edvinsson, S. (2000). The Demographic Data Base at Umeå University: A resource for historical studies. In: P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of International Historical Microdata for Population Research* (pp. 231-248). Minneapolis, Minnesota: Minnesota Population Center.
- Egerbladh, I. & Bittles, A. (2008). The influence of consanguineous marriage on reproductive behavior and early mortality in northern coastal Sweden, 1780–1899. In: T. Bengtsson, G.P. Mineau (Eds.), *Kinship and Demographic Behavior in the Past* (pp. 225-224). Dordrecht: Springer.
- Engberg, E. (2008). *Karesuando 1900-talsmaterial: Rapport från ett utvecklingsprojekt i samarbete med Landsarkivet i Härnösand och Statistiska Centralbyrån*. Umeå: Demographic Data Base.
- Engberg, E. (2005). *I fattiga omständigheter. Fattigvårdens former och understödstagare i Skellefteå socken under 1800-talet*. Umeå: Demographic Data Base.
- Fusé, L. (2008). *Parents, children and their families: living arrangements of old people in the XIX century, Sundsvall region, Sweden*. Umeå: Demographic Data Base.
- Janssens, A. (1993). *Family and Social Change: The Household as a Process in an Industrializing Community*. Cambridge: Cambridge University Press.
- Junkka, J. & Edvinsson, S. (2015). Gender and fertility within the free churches in northern Sweden, 1860-1921. *History of the Family*.
DOI: [10.1080/1081602X.2015.1043929](https://doi.org/10.1080/1081602X.2015.1043929).
- Kok, J. (2007). Principles and Prospects of the Life Course Paradigm. *Annales de démographie historique* (113) 1, 203-230.
- Kolk, M. (2011). Deliberate birth spacing in nineteenth century northern Sweden. *European Journal of Population*, 27(3), 337-359.
- Kuh, D., Cooper, R., Hardy, R., Richards, M. & Ben-Schlomo, Y. (2014). *A life course approach to healthy ageing*. Oxford: Oxford University Press.
- Larsson, M. & Engberg, E. (2015). *Record linkage over administrative borders*. Paper presented at the 40th Annual Meeting of the Social Science History Association, Baltimore, MD.
- Lesthaeghe, R. (2010). The unfolding story of the second demographic transition. *Population and Development Review*, 36(2), 211-251.
- Lundevaller, E. H. & Edvinsson, S. (2012). The effect of the Rh negative disease on perinatal mortality. Evidence from Skellefteå 1840-1900. *Biodemography and Social Biology*, 58(2), 116-132.
- Maas, I. & van Leeuwen, M.H.D. (2002). Industrialization and intergenerational mobility in Sweden. *Acta Sociologica*, 45(3), 179-194.
- Malmberg, G., Nilsson, L. & Weinehall, L. (2010). Longitudinal data for interdisciplinary ageing research. Design of the Linnaeus Database. *Scandinavian Journal of Public Health*, 38(7), 761-767.
- Mandemakers, K. & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 37(1), 34-38.
- Meloni, M. (2014). How biology became social, and what it means for social theory. *The Sociological Review*, 62(3), 593-614.
DOI: [10.1111/1467-954X.12151](https://doi.org/10.1111/1467-954X.12151).
- Nilsdotter Jeub, U. (1993). *Parish Records: 19th Century Ecclesiastical Registers*. Information from the Demographic Data Base.
- Norberg, M., Wall, S., Boman, K. & Weinehall, L. (2010). The Västerbotten Intervention Programme: background, design and implication. *Global Health Action*, 3.
DOI: [10.3402/gha.v3i0.4643](https://doi.org/10.3402/gha.v3i0.4643).
- Pembrey, M., Saffery, R. & Bygren, L.O. & Network in Epigenetic Epidemiology (2014). Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research. *Journal of Medical Genetics*, 51(9), 563-572.
DOI: [10.1136/jmedgenet-2014-102577](https://doi.org/10.1136/jmedgenet-2014-102577).

- Pembrey, M. E., Bygren, L.O., Kaati, G., Edvinsson, S., Northstone, K., Sjöström, M., Golding, J. & The ALSPAC Study Team. (2006). Sex-specific, male-line transgenerational responses in humans. *European Journal of Human Genetics*, 14(2), 159-166.
DOI: [10.1038/sj.ejhg.5201538](https://doi.org/10.1038/sj.ejhg.5201538).
- Rose, N. (2013). The human sciences in a biological age. *Theory, Culture & Society*, 30(1), 3-34.
DOI: [10.1177/0263276412456569](https://doi.org/10.1177/0263276412456569).
- SCB Statistiska Centralbyrån (2007). *Multi-generation register 2006: A description of contents and quality*.
- Svensson, I., Lundholm, E., de Luna, X. & Malmberg, G. (2015). Family life course and the timing of women's retirement – A sequence analysis approach. *Population, Space and Place*, 21(8), 856-871.
DOI: [10.1002/psp.1950](https://doi.org/10.1002/psp.1950).
- United Nations Economic Commission for Europe (2007). *Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics* (pp.5). New York and Geneva: United Nations.
- Van De Kaa, D. J. (1986). Europe's Second Demographic Transition. *Population Bulletin*, 42(1), 1-59.
- Van Leeuwen, M.H.D., Maas, I. & Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.
- Van Leeuwen M. H.D., Maas, I. & Miles, A. (2004). Creating a historical international standard classification of occupations an exercise in multinational interdisciplinary cooperation. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 37(4), 186-197.
- Vandezande, M. (2012). *Born to die. Death clustering and the intergenerational transfer of infant mortality, the Antwerp district, 1846-1905*. (Doctoral dissertation). Retrieved from Faculteit Sociale Wetenschappen, Katholieke Universiteit, Leuven.
- Vikström, L. (2003). *Gendered routes and courses. The Socio-spatial mobility of migrants in nineteenth-century Sundsvall, Sweden*. (Doctoral dissertation). Umeå: Demographic Data Base.
- Wisselgren, M.J., Edvinsson, S., Berggren, M. & Larsson, M. (2014). Testing methods of record linkage on Swedish censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 47(3), 138-151.
DOI: [10.1080/01615440.2014.913967](https://doi.org/10.1080/01615440.2014.913967).
- WHO International Classification of Diseases, ICD-10.