VOLUME 8 | ISSUE 4 | **OCTOBER 2017** ISSN 1757-9597

Longitudinal and Life Course Studies: International Journal

Published by

SLLS Society for Longitudinal and Life Course Studies

Inside this issue

Co-sponsors:

VU university medical center

ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

SRC SURVEY RESEARCH CENTER UNIVERSITY OF MICHIGAN

LLCS EDITORIAL BOARD

SUBSCRIPTIONS

For immediate access to current issue and preceding 2016 and 2017 issues (Volumes 7 and 8):

Annual Individual rate: **£20 (discounted rate £55 for 3 years)**.

Annual Library **rate: £200** to register any number of library readers.

Annual Society for Longitudinal and Life Course Studies (SLLS) Membership: Individual **£65,** ***Student*** **£26,** ***Corporate*** **£300**.

SLLS members have automatic free access to the journal among many other benefits. For further information about SLLS and details of how to become a member, go to http://www.slls.org.uk/#!services/ch6q

All issues are accessible free of charge 12 months after publication.

Print on demand

An attractive printed version of each Issue of the journal, including all back numbers, is available at a price of £10 (plus postage and packaging). If you would like to purchase a full set or individual copies, visit the SLLS Bookshop at http://www.slls.org.uk/#!journal-bookshop/czkc Depending on distance, you will receive the order within two weeks of submitting your order.

Please note

The reselling of personal registration subscriptions and hard copies of the journal is prohibited.

INTRODUCTION

PAPERS

RESEARCH NOTE

STUDY PROFILE

LLCS Journal can be accessed online at www.llcsjournal.org

Published by the Society for Longitudinal and Life Course Studies
info@slls.org.uk www.slls.org.uk

MEMBERSHIP OF THE SLLS

SLLS was established in 2009 and has over 350 members worldwide.
Interested in becoming a member? Then read on....

Membership Benefits

- Substantially reduced fees for attendance at the Society's annual conference
- Free, open access to read and publish in our peer reviewed journal, Longitudinal and Life Course Studies
- Access via log-in to the private members area of the SLLS website, where you will find an archive of all SLLS newsletters and be able to flick through copies of the LLCS Journal
- A regular newsletter of global news and events in longitudinal and life course research from our President
- A link with your regional Global Representative, who promotes SLLS in your region and organises local activities
- Collaborative contacts throughout the global longitudinal and life course research community
- An opportunity to be involved with the SLLS Policy Group, which promotes longitudinal research and communication between policy makers and researchers
- An opportunity to join the SLLS Cohort Network
- Access to theory and methods including our annual summer school
- The opportunity to make nominations, be nominated and vote in elections for the future SLLS President and Executive Committee

Fees
Full Annual Membership - £65/ €90/ US$115
Student Annual Membership - £26/ €35/ US$45
Corporate Annual Membership - £300/ €400/ US$540

Become a Member
To become a member or to renew your membership, visit the LLCS Journal site (www.llcsjournal.org) and follow the online instructions. Please note: You will need to log in/register first on the journal site to proceed with membership set up.

Editorial: Life course research around the world

Heather Joshi

This issue of the journal bears witness to the growing international scope for longitudinal and life course research. The work published here includes analyses of data from separate single countries, Finland and Australia; and an article by authors in the Netherlands using data from 12 European countries, New Zealand, Canada and USA. The study profile is from an international consortium, LIFEPATH, working on complex longitudinal datasets from nine developed countries. Our methodological research note (by Suzanne Walton and colleagues) reports on the development of an integrated approach to developing and scripting questionnaires for longitudinal cohort studies and surveys, which should have applications beyond the UK Life Study for which it was developed. International comparisons and collaborations are not without pitfalls, but the Society for Longitudinal and Life Course studies is well placed to recognise and address them. At the time of writing, the Society has members in 22 countries on five continents, from whom we would welcome submissions and to whose geographical coverage we would like to expand.

'The biology of inequalities in health: the LIFEPATH project' whose profile is presented by Paolo Vineis and colleagues, has a huge international and interdisciplinary scope. The consortium is investigating the biological pathways that lead to socioeconomic inequalities in healthy ageing, and in mortality. The existence of these inequalities in rich countries suggests there is room for improvement. They present a challenge for public health policy, and social and biological researchers. The European Research Council funding brings together a consortium of 55 investigators working on longitudinal datasets. They are from nine countries, in Europe, USA and Australia. Many of the datasets contain social and behavioural data as well as biological information. The latter extend to epigenomics, proteomics and other 'omics' which may be involved in the pathways through which the 'social gets under the skin'. The project has taken on a programme of collaboration, harmonisation of data and synthesis of causal modeling approaches from both social and biological sciences. Readers of this journal will be interested to read of some of the early findings presented here and to follow the project's progress in the years to come.

The article by Yu Han, Aart Liefbroer and Cees Elzinga, 'Comparing methods of classifying life courses: sequence analysis and latent class analysis' is primarily methodological. It usefully and helpfully explains two methods increasingly used for handling sequence data: Sequence Analysis and Latent Class Analysis. They demonstrate results in exemplar data on women's transitions across partnership and parenthood between the ages of 18 and 30, in a number of developed countries in the 1980s and 1990s. Although the two methods generally produce the same results, there was one way in which they differed: whether childbearing after cohabitation should be treated as a separate category to unmarried motherhood. This reflects an issue of substantive interest dividing scholars on either side of the Atlantic – whether cohabiting mothers are more like other partnered mothers than other unmarried mothers. It is notable that the USA provides a substantial subset of the observations in this study. This illustrates the sort of pitfalls to be aware of when pooling international data.

The paper by Pasi Haapakorva, Tiina Ristikari & Mika Gissler, 'The impact of parental employment trajectories on children's early adult education and employment trajectories…' uses Sequence Analysis for the trajectories of young adults in Finland. They were all born in 1987 and are followed through education and (un)employment between ages 18 and 25. The database is an administrative record that also yields histories to classify their parents' movements in and out of employment over those 25 years. Lack of employment in the parental generation was strongly associated with disadvantaged outcomes in the second generation.

The same Finnish birth cohort of 1987 also provides the data for the paper by Miia Bask, and colleagues, 'Psychiatric diagnoses as grounds for disability pension among former child welfare clients'. Rather than Sequence Analysis, they use logistic regression to link childhood circumstances with disability in early adulthood. As the dataset only extends to age 25, the term 'disability pension'

should be understood as a cash benefit awarded on the grounds of incapacity to work, rather than an old age pension. The authors are particularly interested in using the disability benefit records to identify poor mental health. Children who had been under some form of social protection (whether with or away from their parents) – the 'child welfare' cases – were at higher risk of getting the disability payment as young adults, attenuated only somewhat by other evidence on parental situation in childhood. The records also show which psychiatric diagnosis applied to the young 'pensioners', and there were some differences between genders in the relationship of diagnosis to family background.

Abuse and neglect in childhood, even if they attract official social protection, may cast a long shadow on adult wellbeing. Survey evidence about maltreatment is seldom available from administrative sources. Research tends to rely on retrospective data, since it is difficult to ask families about severe difficulties when they are happening. James Doidge and colleagues write about this in 'Adverse childhood experiences, non-response and loss to follow-up'. They use the Australian Temperament Project, which includes retrospective reports from young adults of childhood adversity. The dataset also provides prospective evidence on the earlier life course which helps their attempts to allow for the formidable bias likely to affect even retrospective reports due to missing data.

Although people sometimes mistake our field as something to do with geographical meridians, this issue shows that longitudinal studies do indeed circle the globe. The lines they trace are those that link a person's past with the present and future. Taken together they help document the diversity and inequality of human lives.

Longitudinal and Life Course Studies 2017 Volume 8 Issue 4 Pp 319 – 341 *ISSN 1757-9597*

Comparing methods of classifying life courses: Sequence Analysis and Latent Class Analysis

Sapphire Y. Han Netherlands Interdisciplinary Demographic Institute (NIDI/KNAW), The Hague & University of Groningen, Groningen, The Netherlands

han@nidi.nl

Aart C. Liefbroer Netherlands Interdisciplinary Demographic Institute (NIDI/KNAW), The Hague, University of Groningen, Groningen & Vrije Universiteit, The Netherlands

Cees H. Elzinga Netherlands Interdisciplinary Demographic Institute (NIDI/KNAW), The Hague & Vrije Universiteit, The Netherlands

(Received February 2016 Revised May 2017) http://dx.doi.org/10.14301/llcs.v8i4.409

Abstract

We compare life course typology solutions generated by sequence analysis (SA) and latent class analysis (LCA). First, we construct an analytic protocol to arrive at typology solutions for both methodologies and present methods to compare the empirical quality of alternative typologies. We apply this protocol to develop and compare SA- and LCA-derived family-life typologies for women born between 1960 and 1964 in 15 European countries, using data from the Family and Fertility Survey. This paper contributes to the use of these classification techniques in four different ways. First, we present guidelines on how to establish the number of classes or clusters to use. Second, we show how to evaluate the stability of these clusters. Third, we provide a way to evaluate the validity of these clusters and finally, we provide for a formal heuristic to relate the stochastically defined latent classes to the distance-based clusters found with SA.

Keywords

Life course, sequence analysis, latent class analysis, typology comparison

Introduction

A prominent approach within life course research is to analyse life courses as sequences of states or state-transitions (see e.g. Buchmann & Kriesi, 2011). In this approach, two main methodological paradigms have been widely used: Event History Analysis (see e.g. Mills, 2011) focuses on describing or explaining the time to occurrence of specific events. The second approach takes a holistic perspective and utilises the life course itself as the unit of analysis and usually aims for a typology of life course trajectories. The typology itself may reveal substantive patterns and the resulting class-membership is often used as a dependent or independent variable in further analyses. Brzinsky-Fay and Kohler (2010) argue that these two types of approaches can be viewed as complementary rather than as competing.

In this paper, we compare two strategies to construct such holistic life course typologies. The first strategy, called Sequence Analysis (SA) (Abbott & Forrest, 1986; Cornwell, 2015), starts by calculating a distance measure over the set of sequences and then tries to partition the resulting distance matrix into clusters of trajectories. Sequence analysis and its related typology techniques have been widely applied in studying life course trajectories in the social sciences (e.g. Kleinepier, de Valk & Gaalen, (2015); Helske Steele,

Kokko, Räikkönen & Eerola, (2014)). The second strategy uses a probabilistic model that describes an observed life course sequence of categorical values as resulting from the conditional probabilities that define membership of a latent class and is called Latent Class Analysis (LCA) (Hagenaars & McCutcheon, 2002). Barban and Billari (2012) suggested that LCA can be used as an alternative to SA to derive meaningful classifications of life course patterns. Barban and Billari (2012) demonstrated (see their Table 1) that SA and LCA could generate quite different typologies from the same data. This does not come as a surprise as SA and LCA use very different methodologies. Using SA implies selecting a distance measure followed by a clustering method to partition the distance matrix. Using LCA implies that a category or class is defined by a probability distribution function over a set of categorical observations like 'living single' or 'getting married': different classes are defined by different probability distribution functions over the same states. Thus, both methods imply quite different steps to generate a life course typology.

The main aim of this paper is to discuss how typologies derived from SA and LCA can be compared and their quality assessed. The tools introduced to allow this assessment are useful in a more general sense as well. They can also be used to decide between typologies generated by different distance metrics or different clustering methods in SA, and thus are of interest to all users of holistic life course methods. This paper also offers guidance to researchers who want to use SA and/or LCA in their research, by outlining the steps to be taken and the decisions to be made in performing an SA and/or LCA analysis and by discussing practical bottlenecks that often pop up.

The paper is structured as follows. First, we offer a description of the main steps to be taken in developing a typology using SA and LCA. Next, we discuss methods to compare the SA and LCA typology solutions and to decide on which particular solution is to be preferred. We illustrate these procedures by analysing data from the Family and Fertility Survey as presented in the Methods and Data section. Next, we present the results of our illustrative example and in our final section we draw conclusions about the more general implications of the suggested procedures. Three appendices are added. In Appendix 1, practical issues are discussed. In Appendix 2, we present a heuristic explanation of sequence generation that bridges the gap between SA and LCA. In Appendix 3, R-based commands are provided that can be used as a code-model for the analyses presented.

Sequence Analysis

Sequence analysis (SA) has become the key holistic method to study life course trajectories since Abbott (1983) introduced it in the social sciences. This section briefly outlines the necessary steps and decisions to arrive at an SA-based typology. A sequence dataset has to be constructed from life course data. In this paper, we organised the sequence data as a state-sequence dataset. Other methods have been discussed in Ritschard, Gabadinho, Studer and Müller (2009). The main idea behind SA is to express the dissimilarity between pairs of sequences as a distance. The larger the distance between two sequences, the more dissimilar they are (but see Elzinga & Studer, forthcoming). Therefore, the first decision in sequence analysis is about choosing an appropriate distance metric. The two main classes of metrics available are edit-based metrics and subsequence-based metrics. Edit-based metrics measure the distance between two sequences by counting the minimum number of (weighted) edit-operations required to turn one sequence into a perfect copy of the other. In the social sciences, these metrics (and their numerous variants) are known as 'Optimal Matching' (OM) (Abbott & Forrest, 1986).

Edit-based metrics are rather insensitive to differences in the ordering of states (Elzinga & Studer, 2015). This motivated the development of so-called subsequence-based metrics (Elzinga, 2005; Elzinga & Wang, 2013). These metrics measure the distance between sequences by counting the number of (weighted) common subsequences. For a detailed review of distance metrics for SA, we refer to Robette and Bry (2012), Studer (2012) and Studer and Ritschard (2016). In our illustration, we only present the SA approach with an OM-metric. The reason for this choice is twofold. First, family-formation patterns in modern Western societies vary relatively little in the ordering of events (Billari & Liefbroer, 2010). This makes OM a quite natural choice. Today, OM is the most commonly used metric in studies on the transition to adulthood (Aassve, Billari & Piccarreta, 2007; Brzinsky-Fay, 2007; Robette, 2010). Second,

preliminary analyses showed that in this particular illustration, the OM metric clearly outperformed other metrics. The code in Appendix 3 is easily adaptable to any of the other metrics offered through R-based software.

The computation of distances between all sequences results in a distance matrix. The second step in SA uses this distance matrix to partition sequences into more or less homogeneous groups. Various clustering methods are suitable for this purpose, including hierarchical clustering (Maimon & Rokach, 2005), partitioning around medoids (PAM) (Kaufman & Rousseeuw, 2009), and self-organising maps (SOM) (Massoni, Olteanu & Rousset, 2009). Among them, Ward's method is most widely used (Aassve et al. 2007; Billari & Piccarreta, 2005; Pailhé, Robette & Solaz, 2013).[i]

Ward's method (Ward Jr, 1963) iteratively merges ever-bigger clusters of sequences such that, in each iteration, the increase of the total within-cluster distance is minimised. Critical in this step is to determine at what level of agglomeration, i.e. at which number of clusters, to stop the merging process, as this number is not determined by the method itself. The number of clusters has to be decided upon by applying a combination of substantive theory and measures of statistical cluster quality. Substantive theory alone may not adequately summarise the observed heterogeneity of life course patterns and the clustering algorithm may not lead to a parsimonious set of internally homogeneous and well-separated clusters.

We use three statistics (e.g. Table 2 in Studer, 2013) to empirically determine cluster quality: Average Silhouette Width (ASW), Hubert's C index and the Point Bi-serial Correlation (PBC). ASW (Rousseeuw, 1987), compares the average packing of points within clusters to the average distance of points to the closest cluster to which these points do not belong. A high ASW-value implies that clusters are homogeneous and well separated from each other. The HC index (Hubert & Levin, 1976) shows the gap between the partition obtained and the best partition theoretically possible with this number of groups. A low value of HC indicates good clustering. Finally, PBC (Milligan & Cooper 1985) measures the capacity of the cluster solution to reproduce the original distance matrix. A high PBC value is preferred.

Ideally, one would not only want to know what the optimal number of clusters is, but also their stability. To evaluate stability, most statistical analyses involve not only the estimation of model parameters, but also the estimation of their standard errors. In SA, this is not possible. However, once a theoretically and numerically acceptable cluster solution is obtained, one can examine its stability to data sampling fluctuations by using bootstrap methods. Such bootstrapping (e.g. James, Witten & Tibshirani, 2013) allows examining whether the clustering algorithm returns the same solution across several sub-samples. The clusterwise Jaccard Bootstrap Mean (CJBM) (Hennig, 2007), is a measure that uses the bootstrap to re-sample the data and to compute the Jaccard similarities of the original clusters to the most similar clusters in the re-sampled data. As proposed by Hennig (2008), when CJBM is below 0.6, the cluster solution should not be trusted. If CJBM is above 0.85, the classification technique generates highly stable clusters. A CJBM between 0.6 and 0.85 suggests some structure, but exact cluster membership is uncertain.

Cluster quality and stability measures have to be combined with theoretical interpretation for sound typology decisions. Therefore, the last step in creating an SA typology is to provide a substantively meaningful interpretation of the clusters. Visualisation tools such as sequence index plots (Scherer, 2001) and sequence medoid plots (Gabadinho, Ritschard, Müller & Studer, 2011) facilitate interpretation of cluster solutions. In sequence index plots each sequence is represented by a line composed of differently colored segments, with colors representing states and the length of the segments being proportional to the time spent in a state. A sequence index plot summarises large amounts of information in a single graph: order, prevalence and timing of states and overall variability within and between sequences. The medoid sequence is an observed sequence whose average distance to all the other sequences in a cluster is minimal.

The SA-cluster solution is affected by many factors: the sequence encoding, the choice of a metric, and the choice of a clustering technique. Here, we present no sensitivity analyses of our results since our purpose is mainly to elaborate on the global methodologies of applying and comparing SA and LCA.

Latent Class Analysis

Latent Class Analysis (LCA)[ii] is a statistical technique for the analysis of multivariate categorical data (see e.g. Hagenaars & McCutcheon, 2002). To concisely explain the LC-model, we need some concepts and notation. First, we write $y = y_1 \dots y_n$ to denote an observed sequence of length n. Second, we denote the latent class model Θ as a set of R conditional probability distributions $T = \{\theta_1, \dots \theta_R\}$ over the observable states, each of these characterising precisely one of the R latent classes. Furthermore, the model needs a specification of the probability that a sequence is generated from any of these latent classes, the vector $P = (\pi_1, \dots \pi_R)$ wherein π_j denotes the probability that a sequence is generated from θ_j. Thus a complete LC-model can be specified as $\Theta = (R, T, P)$. The LC-model states that the probability of observing a particular sequence, given the model, equals

$$Prob(y|T) = \sum_{r=1}^{R} \pi_r \prod_{i=1}^{n} Prob(y_i|\theta_r).$$

So, the model states that, given a fixed latent class r, the consecutive observed states are statistically independent and this assumption is known as 'local independence'. This mixture model (e.g. McLachlan & Peel, 2000) is closely related to supervised Naive Bayes classifiers (Hand & Yu, 2001; Vermunt & Magidson, 2003). Despite the highly implausible assumption of local independence (Rennie, Shih & Karger, 2003), such models often perform quite well for classification tasks because dependencies often are equal across classes or cancel out (Zhang, 2005). Of course, some sequences may be extremely (un-)likely to be generated from some of the latent classes. If the number of classes is well chosen, each observed sequence is relatively (much) more likely to have been generated from one particular latent class than from any of the other classes. Therefore, class membership of each specific sequence is often decided by assigning the sequence to the class with the highest probability of generating that sequence.

Local independence implies that, for each latent class, observing the sequence $aaabbb$ is precisely as likely as observing the sequences $bbbaaa$ or $ababab$ or any other of the 20 discernable permutations of these six observations. So, local independence is a counter-intuitive assumption in the context of modelling life course sequences. Indeed, as can be observed from Figure 2, life courses mainly differ in the timing and selection of states, not in the orderings of states. The assumption of equal probability of observing any ordering of a given collection of states arises from the assumption that, given a conditional distribution θ_i, the observable states are generated by just sampling the alphabet of states according to θ_i. The observation that order-differences are rare (Figure 2) in fact constitutes a statistical test of this assumption: the assumption should be rejected because of observing so little variation of state-orderings. Therefore, applying the LC-model to describe life courses requires an interpretation of that model that, on the one hand, includes the assumption of local independence but that, on the other hand, makes variation in the ordering of states implausible. Such an interpretation is amply discussed in Appendix 2: it is assumed that different classes arise by different "template sequences" that are edited, state by state, such that

1. successive edits are statistically independent and
2. edits resulting in an actual change of a template-state are implausible.

So, in this interpretation, sampling observable states from class-conditional distributions is replaced by sampling of edits and applying them to class-conditional templates. If one additionally assumes that there is a unique, most likely path of edits, OM-distances between pairs of sequences are monotone with the probability that these pairs result from (editing) the same class-specific template. If this assumption is valid, the observed scarcity of order differences within classes is explained. In Appendix 2, we detail and formalise this interpretation and the unifying SA-LC assumption of one dominant edit path.

Just like in cluster analysis, when using LCA, one has to decide on the optimal number of classes. The number of latent classes to a large extent determines the fit of the model: the more latent classes, the "easier" it becomes to accommodate the diversity of the observed sequences. As the number of classes increases, the likelihood of the model generating the sequences increases, but at the risk of fitting to noise and at the expense of estimating more model parameters. Although the LCA model itself does not automatically determine the number of latent classes, a variety of goodness of fit statistics are available (Lin & Dayton, 1997).

Using statistics such as BIC and relative entropy (Vermunt & Magidson, 2013), one can gain information about model fit against the number of latent classes. One usually looks for a minimum in the BIC-curve. Relative entropy is between zero and one, with values near one indicating high certainty in classification and values near zero indicating low certainty. Here, we will use information on both BIC and relative entropy.

Like in SA, visualisation tools can be used to facilitate the interpretation of the latent class solution. By estimating, separately for each latent class, the sequence state with the highest frequency at each time point, we construct a model state sequence for each latent class. The resulting sequence model state plot in LCA is comparable to the medoid plot in SA in the sense that it aims at summarizing the key features of a cluster. However, whereas in SA the medoid is an actually observed sequence, the model state sequence in LCA might be a non-existing sequence. An interpretation of the latent class can also be obtained through visual inspection of the sequence index plot. Thus, the key decision in LCA to be made is on the number of latent classes and their interpretation can be aided by sequence index plots and sequence model state plots.

Typology Comparison

Both SA and LCA can assist researchers to detect structures in sequence data by segmenting the life course sequences into clusters or classes. However, the most basic distinction between SA and LCA is the way in which the classes or types are defined. An SA typology is based on a distance measure, a clustering procedure and a set of statistics that determine the quality of the cluster solution, while an LCA typology is obtained via the maximisation of a likelihood function that derives from a probabilistic model. Two questions arise in this context: (i) how similar are both typologies and (ii) can we decide on which solution has to be preferred?

A number of tools are available to judge how similar the two solutions are. The simplest tool is a cross tabulation of both typologies. Let S be a set of N data-items and let U denote an SA-typology over S: $U = \{U_1, U_2, \dots, U_R\}$ with $U_i \cap U_j = \emptyset$; and $\cup_i U_i = S$. Similarly, let V denote an LCA-typology over S: $V = \{V_1, V_2, \dots, V_C\}$ with $V_i \cap V_j = \emptyset$; and $\cup_i V_i = S$. An $R \times C$ cross-tabulation summarises the overlap between the two typologies by listing the numbers $n_{ij} = |U_i \cap V_j|$. A quantification of the overlap or agreement between two classifications can be achieved through the Rand index. The Rand index (Rand, 1971) is built upon counting pairs of items on which two typologies agree or disagree. As used above, the N item pairs in S can be classified into one of four types N_{11}: the number of pairs of which the members are in the same cluster in both U and V; N_{00}: the number of pairs of which the members are in different clusters in both U and V; N_{01} : the number of pairs of which the members are in the same cluster in U but in different clusters in V; and N_{10}: the number of pairs of which the members are in different clusters in U but in the same cluster in V. These numbers can be calculated using the n_{ij}. N_{11} and N_{00} can be used as indicators of agreement between U and V. The Rand index (R) is defined as: $R = (N_{00} + N_{11})/(N_{00} + N_{01} + N_{10} + N_{11})$, which ranges from 0 (no pair classified in the same way under both typologies) to 1 (identical typologies). In our illustration, we will adopt the adjusted Rand index (Hubert & Arabie, 1985) used by Barban & Billari (2012) to compare typologies.

The adjusted Rand index provides us with information about how different the two typologies are, but does not provide any clue about whether one or the other typology is superior. We suggest that the concept of construct validity as developed in psychometrics (Cronbach & Meehl, 1955; Ross, Wright & Anderson, 2013) can be fruitfully applied to decide which one of a set of alternative typologies is preferable. A typology can be viewed as a theoretical construct that is always part of a larger nomological network, i.e. a set of relationships between the concept of interest (the typology) and other concepts. If one has expectations about the statistical relationship between a typology and other concepts, one can measure the strength of these statistical relationships for each of the typologies. The stronger these relationships, the more likely it is that the measure of a construct is valid. This approach is often used in psychology to assess the validity of psychological constructs (e.g. Leary, Kelly, Cottrell & Schreindorfer, 2013). The nomological network approach is illustrated in Figure 1. A life course typology is part of a nomological network that specifies how this typology is related to other variables, either determinants ($x_1 \dots x_i$) or

consequences ($y_1 \ldots y_i$). Given our knowledge about the expected relationships in this network, one can assess how strongly alternative typologies are related to the rest of the nomological network. In general, one would prefer the typology that shows the strongest associations with other variables in the network. Given that we can use substantive information about the expected relationship between a life course typology and related concepts (e.g. levels of education, religiosity and well-being), this approach offers an elegant, theory-driven solution to the dilemma of deciding between alternative life course typologies.

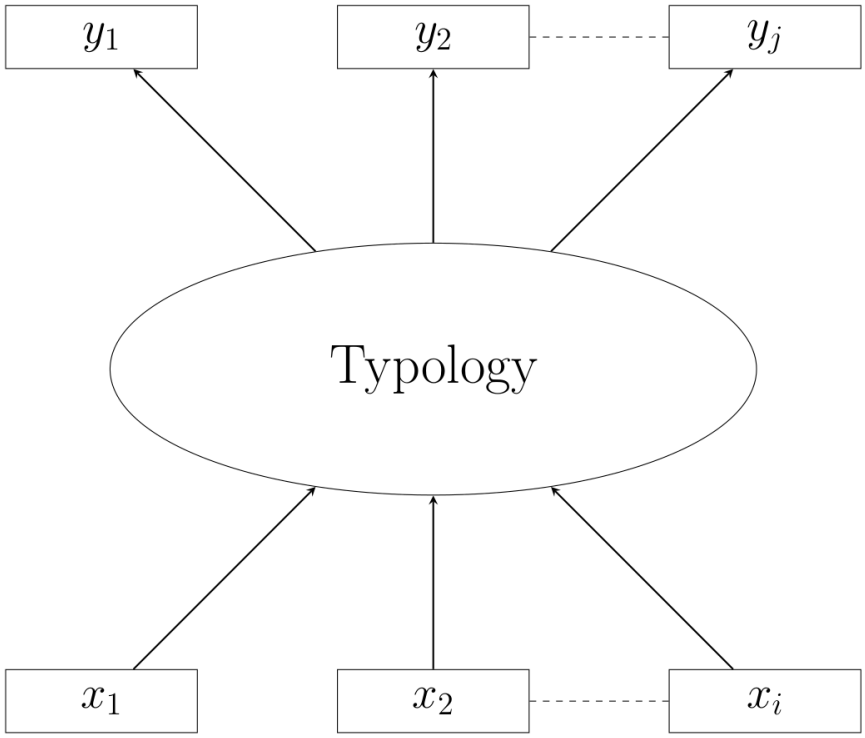


Figure 1. A representation of a nomological network surrounding a life course typology

The statistical application of this idea of construct validity depends on the kind of relationships within the nomological network. If one examines relationships between the typology and consequences, one can use linear or logistic regression, depending on the measurement level of the dependent variables. If one examines the relationship between typologies and their determinants, the easiest test of construct validity would be to estimate separate multinomial logit models with each of the competing typologies as the dependent variable, and a joint set of predictors. What complicates matters here, is that one cannot simply compare the fit-statistics of such models, as the dependent variable (class membership) differs between typologies, and thus one cannot use standard indicators of model fit. However, an alternative procedure is possible by swapping the dependent and independent variables, and predicting the available background variables from the SA- and LCA-generated typologies. Given that the dependent variables are the same now, one can use BIC and other fit-indices to judge which typology is more strongly related to the background variable of interest. One can repeat this procedure for multiple background variables, and decide upon the best typology by comparing the sets of BIC values for the alternative typologies. Alternatively, one can use a MANOVA-approach to compare the quality of the different typologies. Here, we do not use such a multivariate MANOVA-approach because the underlying distributional assumptions are hard to test thoroughly. However, such multivariate tests and the instruments for

analysing the power of such tests are readily available (e.g. Faul, Erdfelder, Lang & Buchman, 2007) to the interested reader.

Methods and Data

Data

We used a subset of the Family and Fertility Survey (Festy & Prioux, 2002). This subset (Elzinga & Liefbroer, 2007) includes 10,301 female respondents from 15 countries born between 1960 and 1964. Full monthly event history information was available regarding respondents' fertility and partnerships between ages 18 and 30, their country of birth, years of education after age 15, religion and parental divorce, as shown in Tables 1 and 2.[iii]

Table 1. Definition of social background variables used in the typology comparison

Abbreviation	Meaning
Edu1	No education after age 15
Edu2	0-3 years of education after age 15
Edu3	3-5 years of education after age 15
Edu4	5+ years of education after age 15
Pardiv0	Parents not divorced
Pardiv1	Parents divorced
Pardiv3	Parents' divorce not known
Reli0	Not religious
Reli1	Catholic
Reli2	Protestant
Reli3	Other religion
Reli4	Religion unknown

Table 2. Number of respondents per country and percentage of the respondents per category of the social background variables

	Region	edu1	edu2	edu3	edu4	pardiv0	pardiv1	pardiv3	reli0	reli1	reli2	reli3	reli4	Nr. Resp.
1	Estonia	9.51	48.59	41.20	0.70	66.20	22.18	11.62	47.54	0.00	41.20	11.27	0.00	284
2	Czech Republic	5.10	23.81	50.34	20.75	85.37	14.29	0.34	--	--	--	--	--	294
3	France	8.02	44.32	22.94	24.72	87.53	10.91	1.56	--	--	--	--	--	449
4	New Zealand	--	--	--	--	--	--	--	--	--	--	--	--	460
5	Hungary	21.58	30.56	25.85	22.01	85.04	14.53	0.43	39.10	47.65	8.12	3.21	1.92	468
6	Latvia	0.85	22.67	39.41	37.08	77.12	19.07	3.81	31.78	20.55	17.37	26.69	3.60	472
7	Lithuania	1.17	15.37	33.46	50.00	80.54	18.09	1.36	8.17	80.93	0.78	8.56	1.56	514
8	Slovenia	18.02	20.14	32.69	29.15	92.40	7.42	0.18	21.02	68.90	0.18	8.66	1.24	566
9	Netherlands	4.08	35.85	22.84	37.22	85.02	9.98	4.99	39.94	34.80	19.21	5.90	0.15	661
10	Spain	36.95	22.36	10.84	29.85	97.32	2.68	0.00	17.40	78.05	0.54	3.08	0.94	747
11	Austria	20.88	30.05	33.51	15.56	90.03	9.44	0.53	30.72	59.18	3.59	6.38	0.13	752
12	Canada	--	--	--	--	82.72	15.31	1.96	3.80	46.73	33.77	15.71	0.00	764
13	Italy	30.47	15.48	18.43	35.63	97.79	2.21	0.00	8.72	90.05	0.49	0.61	0.12	814
14	Portugal	--	--	--	--	94.38	5.18	0.44	--	--	--	--	--	908
15	U.S.A.	50.65	0.28	8.47	40.60	75.84	24.07	0.09	8.47	29.05	49.67	12.71	0.09	2148

The table is ordered by the total number of respondents per region
"–": Data not available

We organised the sequence data as a state sequence (STS) data set (Ritschard, Gabadinho, Studer & Müller, 2009). STS is a chronologically ordered list of the states based on the survey information. We distinguish six family formation states: living single (S), unmarried cohabitation (U), marriage (M), living single with a child/children (SC), cohabitation with a child/children (UC), and marriage with a child/children (MC). Therefore, the sequence data consist of 144 monthly family-life statuses; an example from one respondent is shown below.

$$\overbrace{S \ldots S}^{87}\ \overbrace{M \ldots M}^{56}\ \overbrace{MC \ldots MC}^{11}$$

This person has first spent 87 months in the Single state, followed by 56 months in the Married state and 11 months in the Married with Children state.

Methods

All methods introduced in the previous sections were applied to data set. All analyses were performed in the R software environment for statistical computing and graphics on a 3.2 GHz CPU, 32GB RAM and 64-bit PC, using the R packages TraMineR (OM), stats (hierarchical clustering), WeightedCluster (cluster decision), fpc (bootstrapping), poLCA (LCA), flexclust (Rand index), and nnet (multinomial logistic regression).

Results

SA Typology

Following the steps outlined earlier, we first compute a distance matrix using the TraMineR software (Gabadinho et al., 2011). There are many different metrics to construct distances between sequences. The choice for either of these metrics may affect the nature of the resulting typology. So, one should be aware of the differences between these metrics (see Robette & Bry (2012), Elzinga & Studer (2015) and Studer & Ritschard (2016) for a detailed discussion of these issues). Here, we experimented with a variety of distance measures: OM with various cost settings and a sequence-based vector representation (Elzinga & Studer, 2015) with various parameter settings. OM with indel-cost of 4 and substitution-cost of 2 (the default setting of TraMineR) generates the best solution. Here, we present the results for the selected OM-metric only. The cost setting used implies that all substitutions were equally costly and that mere deletion or insertion never occurred. The solutions for other distance measures can be obtained from the first author upon request.

In our example, we use, for reasons already explained, hierarchical clustering (Ward's method). In Table 3, values of ASW, PBC, and HC are presented for solutions with two to eight clusters. The values in Table 3 show that a solution with six clusters is favored, as this solution combines the highest values of ASW (0.63) and PBC (0.35) with the lowest value of HC (0.10). The next step is to test the stability of the cluster solution by using the CJBM statistic. This statistic indicates to what extent sequences are likely to be assigned to the same cluster over a large number of random draws from the sample. The higher this likelihood, the more stable the cluster solution is. The CJBM statistic for the six clusters of the SA-6 solution varies between 0.45 and 0.76, falling short of the 0.85-level, which is considered to indicate high cluster stability. This suggests that quite a few clusters are not very stable. Table 4 shows that for the SA-7 solution, the CJBM statistics vary between 0.42 and 0.76.

Table 3. Values of cluster quality statistics

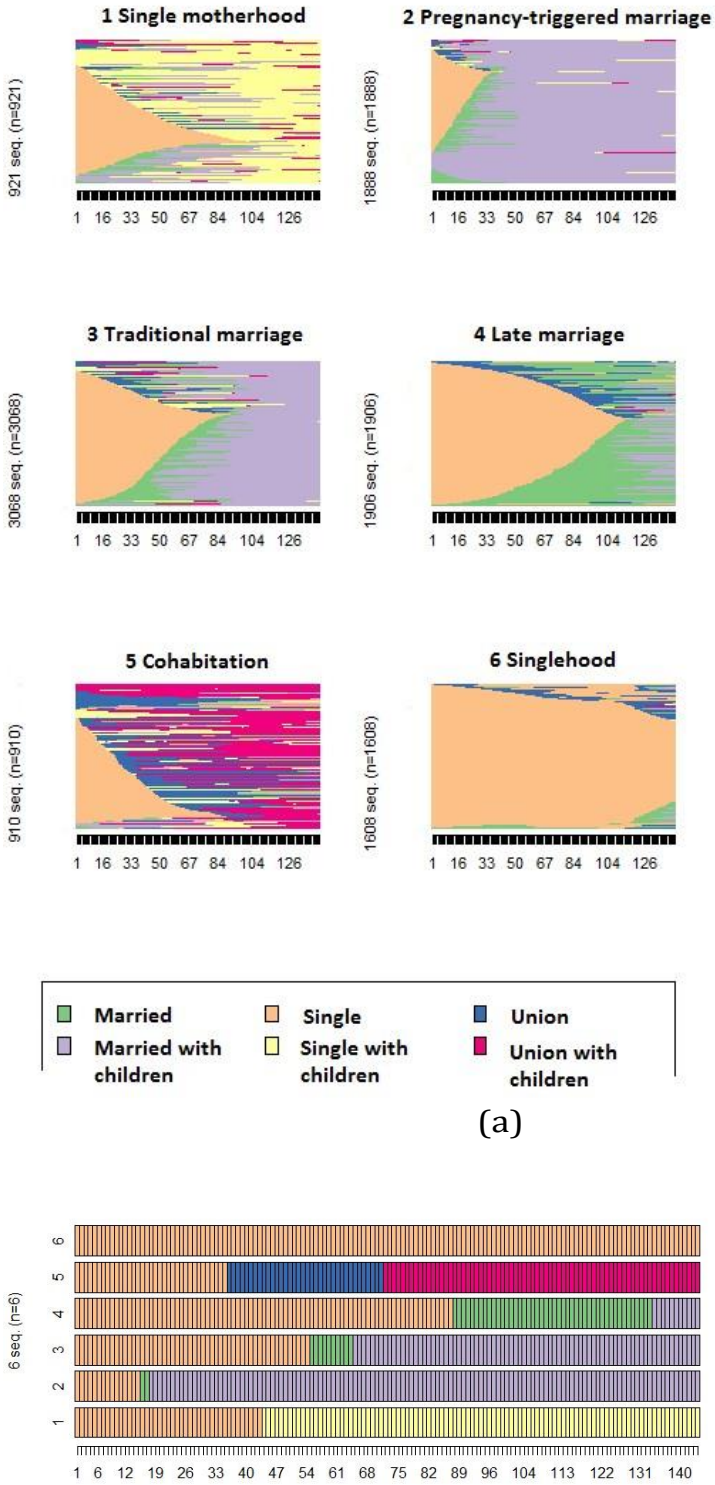
Number of clusters	PBC	ASW	HC
2	0.48	0.34	0.21
3	0.49	0.29	0.22
4	0.62	0.34	0.13
5	0.57	0.31	0.14
6	0.63	0.35	0.10
7	0.59	0.33	0.11
8	0.58	0.33	0.10

Note: PBC (maximal value preferred), ASW (maximal value preferred) and HC (minimal value preferred)

Table 4. Values of CJBM statistics of all six clusters of the OM optimal solution

CJBM	Cluster 1	Cluster2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
SA-6	0.57	0.66	0.54	0.49	0.45	0.76	NA
SA-7	0.62	0.69	0.54	0.42	0.48	0.46	0.76

The final step is to interpret the cluster solution. Figure 2a presents a sequence index plot to show how individuals within each cluster move between states over time. Figure 2b shows a sequence medoid plot that presents the sequence with the smallest average distance to the other sequences in the pertaining cluster. In Figure 2a, cluster 1 mainly consists of sequences starting as 'single' (S) followed by a transition to 'single with children' (SC). This suggests that a meaningful label to this cluster could be 'single motherhood', which is confirmed by the sequence medoid plot shown in Figure 2b. Figure 2b shows that the medoid sequence of cluster 1 spent a spell of 43 months as Single (S), followed by a spell of 101 months being single with children (SC). Based on the interpretation of these plots, labels are assigned to all six clusters. Next to the 'single motherhood' cluster (8.9% of the cohort), we found a cluster that we label 'pregnancy-triggered marriage' (18.3%) as these sequences often have less than nine months between marriage and parenthood, a cluster labeled 'traditional marriage' as marriage is usually followed rather soon by motherhood (29.7%), a cluster labeled 'late marriage' (18.5%), a cluster labeled 'cohabitation' as many of these sequences are characterised by spells of cohabitation either with or without children (8.8%), and finally a cluster labeled 'singlehood' (15.6%).



(a)

(b)

Figure 2. Sequence Index plot (a) and Sequence Medoid plot (b) of the OM-6 solution

LCA Typology

The first step in a LCA is to decide on the number of classes to generate. In a stepping-stone paper, Frahley & Rafterey (1998, Figs. 3 and 4) demonstrated how BIC can be used for cluster-model comparison. Within one type of covariance-structure, these authors plotted the BIC for different numbers of clusters, expecting the BIC to first decrease and then increase again around the optimal number of clusters. Here, following the principle set out by these authors, we compare LC-models that only differ in their number of clusters. So, theoretically, one could expect BIC first to drop drastically with the increase in the number of classes, followed by a slow decrease and finally by an increase again, the latter due to the large number of parameters estimated. High relative entropy (close to 1) indicates good model fit and therefore a desirable number of latent classes. In Table 5, the observed BIC and relative entropy values of the 2- to 8-class solutions are presented. The BIC and relative entropy values of the LCA typology in our example do not exactly show the expected pattern. One observes a drastic decrease in BIC values up to about a five-class solution and a much slower decline up till an eight-class solution. Relative entropies for all solutions are close to 1, suggesting high certainty in classification. Given that all entropies are close to 1, it is hard to base model-selection on relative entropy. The decrease in BIC obviously slows down after five classes, and we decided to examine the six- and seven-class solutions in more detail.

Table 5. Values of latent class analysis model fit statistics: BIC (minimal value preferred), and relative entropy (closest to one value preferred)

Number of clusters	BIC*10^6	relative entropy
2	3.0	0.9993
3	2.6	0.9992
4	2.3	0.9982
5	2.1	0.9980
6	2.0	0.9979
7	1.9	0.9976
8	1.8	0.9975

The interpretation of the LCA typology can be facilitated by sequence index plots and sequence model state plots (Figures 3 and 4). The six-class solution (for short: LCA-6, presented in Figures 3a and 3b) partitions female respondents into classes that we interpret as 'singlehood' (16.4%), 'childbirth outside marriage' (13.6%), 'traditional marriage' (21.2%), 'late marriage' (16.8%), 'cohabitation without children' (9.2%), and 'pregnancy-triggered marriage' (22.8%). LCA-7 (Figures 4a and 4b) generates five classes that are quite comparable to LCA-6, namely 'singlehood' (16.4%), 'late marriage' (17.0%), 'traditional marriage' (21.5%), 'pregnancy-triggered marriage' (21.2%), and 'cohabitation without children' (9.2%). The main difference with LCA-6 is that instead of one class dominated by sequences with children outside marriage, there are now two classes. One of them can be interpreted as 'cohabitation with children' (6.1%), and the other as 'single motherhood' (8.6%). As it is hard to judge which class number is optimal, we decide to compare both solutions to the SA-6 solution in the next section.

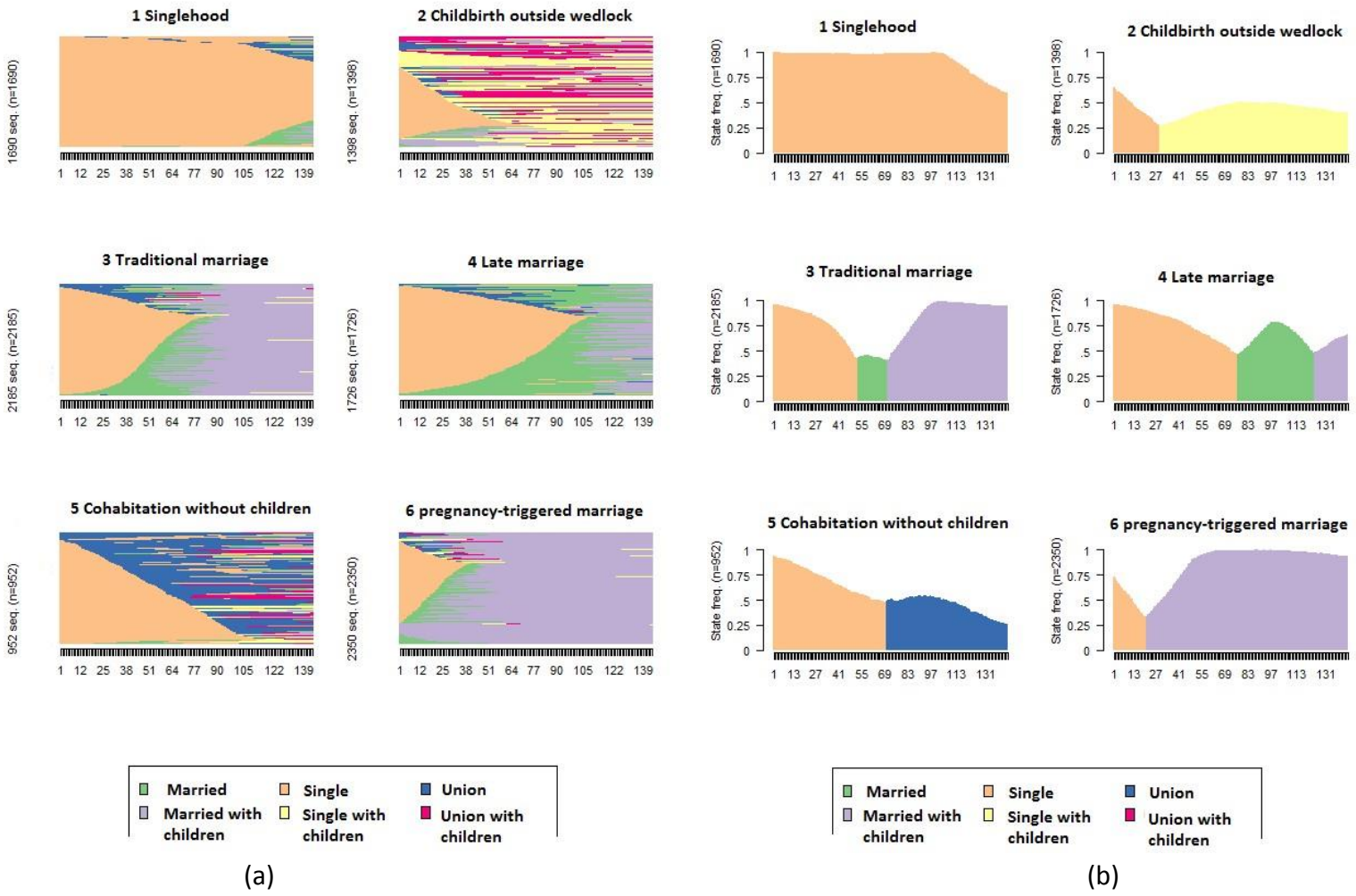
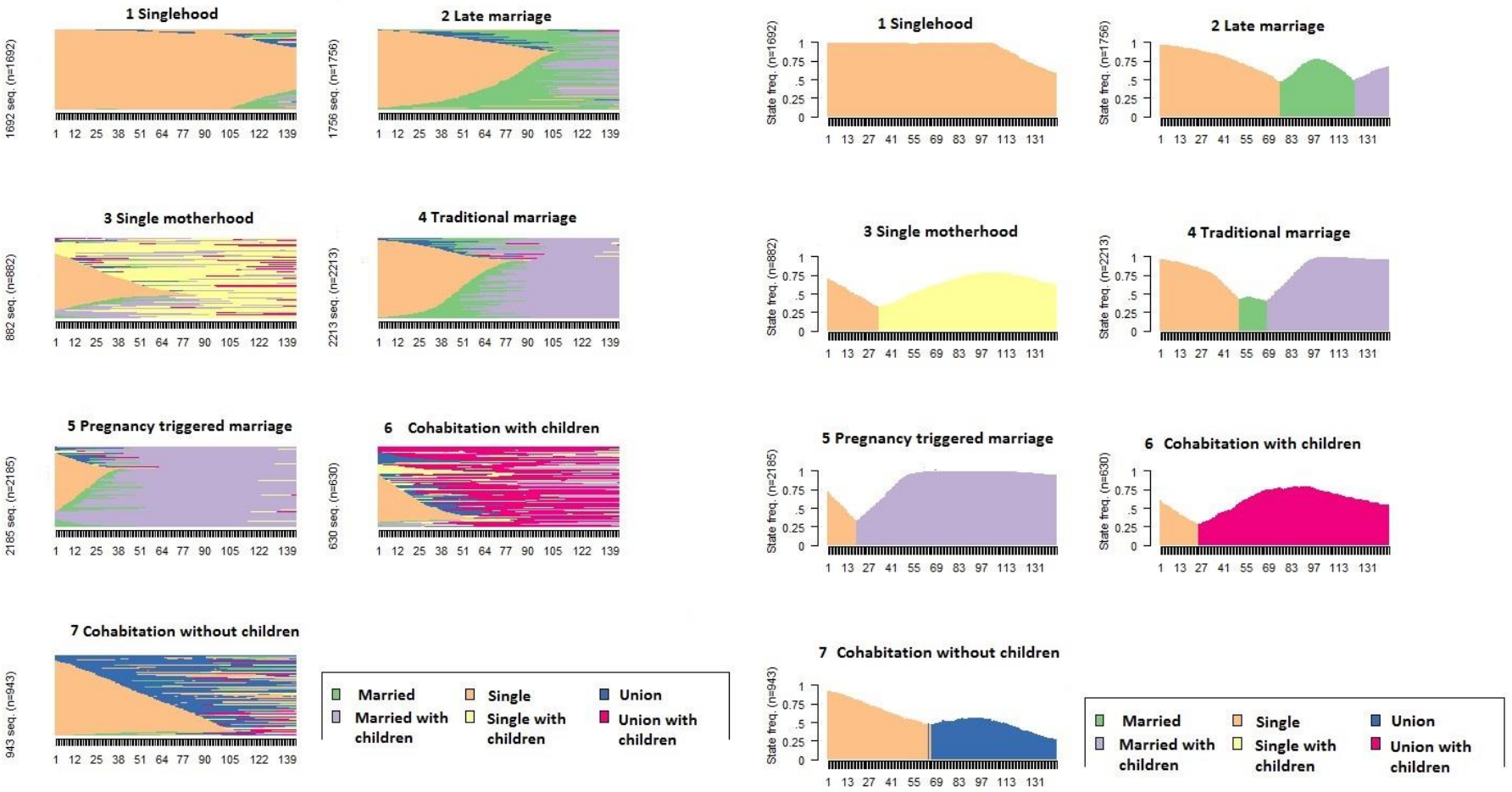


Figure 3. Sequence Index plot (a) and Sequence Model State plot (b) of the LCA-6 solution



(a) (b)

Figure 4. Sequence Index plot (a) and Sequence Model State plot (b) of the LCA-7 solution

Typology Comparison

As a first step in the typology comparison, we simply compare the frequency distributions of each cluster or class in SA and LCA solutions. SA-6 and both LCA typologies have four major clusters in common. These are 'pregnancy-triggered marriage', 'late marriage', 'traditional marriage' and 'singlehood'. As shown in Figure 5, the traditional marriage cluster is somewhat larger in the SA-6 solution (30%) than in either of the LCA solutions (around 21%). In reverse, the pregnancy-triggered marriage cluster is somewhat smaller in SA-6 (18%) than in both LCA solutions (around 22%). The late marriage cluster (17%) and the singlehood cluster (16%) roughly match each other in all three typologies. The smaller clusters differ considerably, both between the SA and LCA solutions, and between the two LCA solutions. In SA-6, the two smaller clusters are single motherhood (9%) and cohabitation (9%). In the LCA-6 solution, the two smaller clusters are interpreted as childbirth outside marriage (14%) and cohabitation without children (9%). In the LCA-7 solution, the three smaller clusters are labeled as single motherhood (9%), cohabitation with children (6%) and cohabitation without children (9%). Note that the cohabitation cluster in the SA solution includes both cohabiters with and without children. Besides, the LCA-6 solution combined single motherhood and cohabitation with children into one class. Only in the LCA-7 solution do these three groups form separate clusters. Thus, the key difference between the three classifications is the way that cohabiters with children are classified. They are classified separately in the LCA-7 solution, grouped together with single mothers in the LCA-6 solution and grouped together with cohabiters with children in the SA-6 solution.

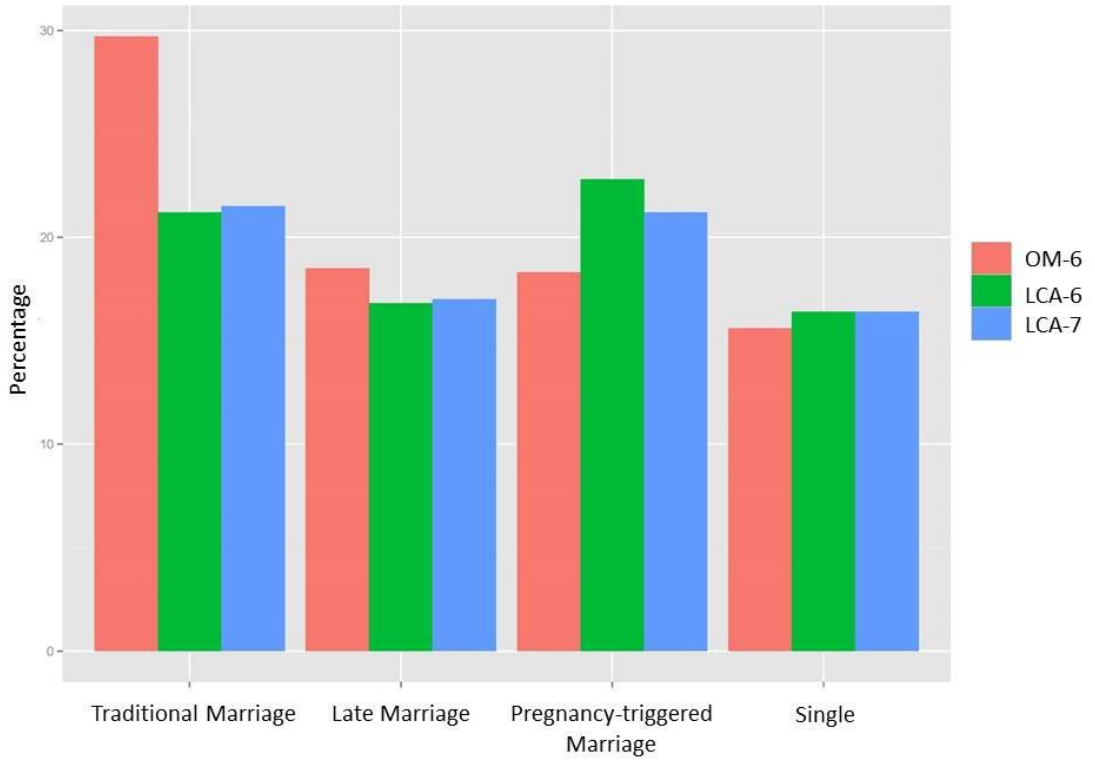


Figure 5. Percentage of respondents in the four large clusters or classes of OM-6, LCA-6 and LCA-7

Given that the substantive interpretation and labeling of SA-6 and LCA-7 were quite similar, we use these cluster solutions to illustrate the usefulness of cross-tabulation (Table 6). If the typologies would be exactly the same, there would be a permutation of its rows and columns such that only the diagonal elements would have positive values, whereas the rest of the table would be empty; when the numbers of classes in the two typologies differ, one may expect that members of one class in the one typology are distributed over a small number of classes in the other typology

(Agresti, 2001). Five classes in SA-6 and LCA-7 received the same label. In Table 6, we observe that most of the sequences that were assigned to these classes in the SA solution were assigned to the same class in LCA-7. In all, 73% of sequences are assigned to clusters with roughly the same substantive interpretation in both analyses. The major difference between SA-6 and LCA-7 is that the former only has one cluster with sequences dominated by cohabitation, whereas the latter contains two clusters, one dominated by sequences of cohabitation without children and one dominated by sequences of cohabitation with children. Table 6 shows that the sequences that are assigned to the cohabitation cluster in SA-6 are almost equally split between the two cohabitation clusters in LCA-7. Four additional differences are found between SA-6 and LCA-7. Of those sequences that are classified as traditional marriage in SA-6, 3.6% are classified as late marriage in LCA-7. Another 3.0% of the traditional marriage sequences in SA-6 are classified as pregnancy-triggered marriages in LCA-7. Thus, it seems that the traditional marriage cluster in SA-6 encompasses a broader range of marriages than the traditional marriage cluster in LCA-7. Similar results hold for the late marriage cluster in SA-6. About 3.3% of the sequences in this cluster are classified as cohabitation without children in LCA-7, and another 2.1% as single in LCA-7. A comparison of the sequence index plots in Figures 2a and 4a shows that some cohabiters who married at a relatively late age, are classified as late marriage in SA-6 but as cohabiters without children in LCA-7. Similarly, some respondents who had been single for most of the time and only married just before turning 30 are classified as late marriage in SA-6 and as single in LCA-7.

Table 6: Cross tabulation of LCA and SA typology solutions, values shown as percentages

LCA/SA	Cohab	Lmarriage	Pmarriage	Smothers	Single	Tmarriage
Cohab with c	4.14	0.00	0.04	0.45	0.00	1.50
Cohab without c	3.80	3.27	0.00	0.93	1.10	0.06
Lmarriage	0.12	13.03	0.00	0.11	0.17	3.62
Pmarriage	0.01	0.00	17.88	0.30	0.00	3.02
Smothers	0.75	0.08	0.32	6.46	0.00	0.96
Single	0.00	2.08	0.00	0.01	14.34	0.00
Tmarriage	0.03	0.05	0.09	0.69	0.00	20.63

Cohab = Cohabitation, Lmarriage = Late marriage, Pmarriage = pregnancy-triggered marriage, Smother = single mother, Tmarriage = Traditional marriage, Cohab with c = Cohabitation with children, and Cohab without c = Cohabitation without children

The adjusted Rand index for the cross-classification of SA-6 and LCA-6 is 0.59, 0.67 for the cross-classification of SA-6 and LCA-7, and 0.88 for the cross-classification of LCA-6 and LCA-7. Evidently, LCA-6 and LCA-7 have a large overlap, but also SA-6 and LCA-7 are highly comparable.

Which typology to prefer? Based on our discussion of construct validity, the strength of the statistical relationship between class membership and relevant background variables can be used to judge the quality of the typology. The stronger the relationship between class membership and other variables that are expected to be related to that typology, the better the typology performs. Therefore, we use class membership as the independent variable in multinomial logistic regression models predicting a series of relevant background variables. In this example, we use four external variables: level of education, parental divorce, level of religiosity and country. Some variables are not available for all countries, and respondents from these countries are excluded in the relevant analyses. To balance the analysis, we compare not only SA-6 with LCA-6 and LCA-7, but also add SA-7, even though the cluster quality statistics in Table 3 clearly favor SA-6 above SA-7. BIC's of all four typologies for all four dependent variables are presented in Table 7.

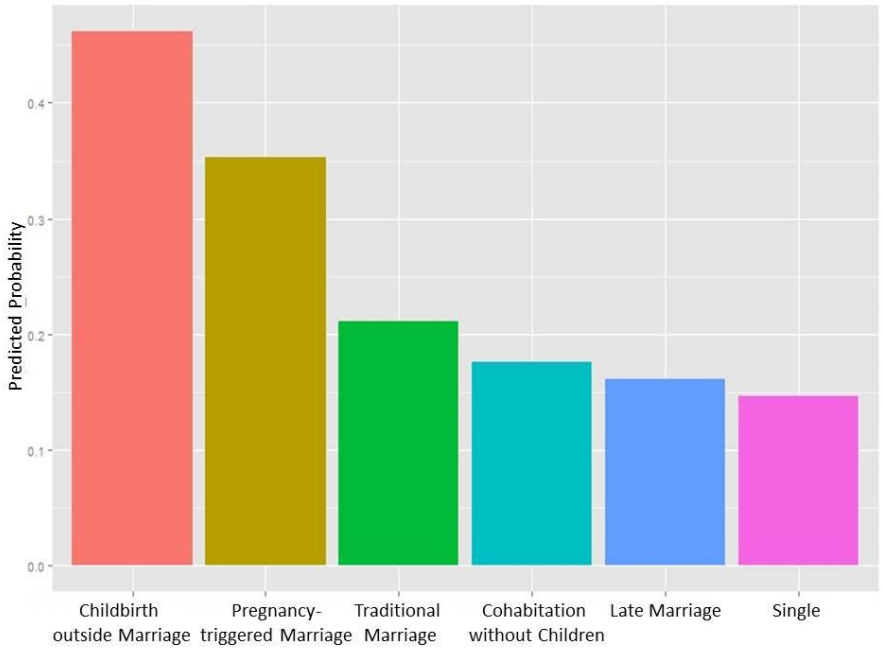
The BIC's of SA-6 are always lower than those of SA-7, implying that the added complexity of SA-7 does not improve the predictive power of the typology sufficiently to warrant the additional complexity.

Things are less clear-cut for the LCA typologies. Based on the BIC's for predicting parental divorce and religion, LCA-6 seems superior to LCA-7, while for predicting education and country, LCA-7 seems superior to LCA-6. To understand this, it is important to remember that in LCA-6, single motherhood and cohabitation with children are jointly classified in one class 'childbirth outside marriage', while in LCA-7, these are separate classes. For predicting parental divorce and religion, the distinction between single motherhood and cohabitation with children does not improve model fit, suggesting that those classified as single mothers and those classified as having a childbirth within cohabitation do not differ much in terms of their odds of experiencing a parental divorce or of being religious. However, for predicting education and country, distinguishing these two groups improves model fit. This suggests that those in the 'single motherhood' class and the 'cohabitation without children' class differ significantly from each other in terms of their distributions across countries and across levels of education. To provide a better interpretation of the meaning of this latter difference, we calculate predicted probabilities (Figure 6) of having no further education after age 15. In LCA-6, those classified as 'childbirth outside marriage' have a 47% chance of having no education after age 15. In LCA-7, this group is split and those classified as single mothers have a higher chance of no additional education (48%) than those classified as having a child within cohabitation (43%). Whether to prefer the LCA-6 or LCA-7 models, depends on how substantively meaningful these differences in educational distributions are. We view them as substantively meaningful, as they show that single motherhood is more strongly linked to social disadvantage than having a child within cohabitation, and thus we prefer LCA-7 to LCA-6.

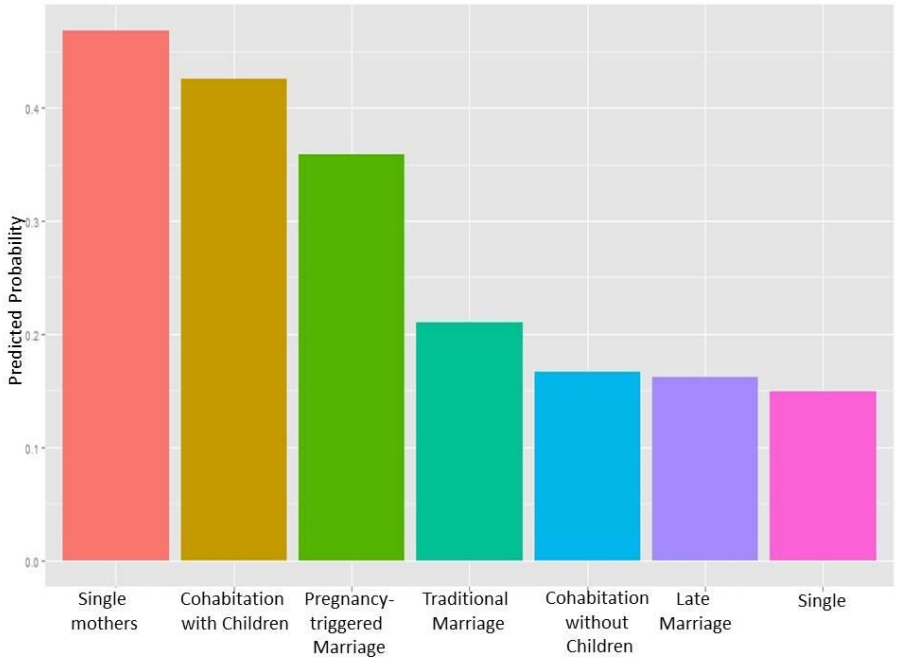
Finally, to choose between the SA-6 and LCA-7 classifications, BIC-values of the models with these two typologies were compared. Table 7 shows that the BICs for all background variables are lower for the LCA-7 typology than for the SA-6 typology. Therefore, we conclude that in in this *particular* illustration the LCA-typology is superior to the SA-typology. Overall, choosing LCA-7 seems the best decision.

Table 7. BICs of SA (OM 6 and OM 7) and LCA (LCA 6 and LCA 7), based on multinomial logistic regression models

	Education	Parental divorce	Religion	Country
OM 6	21612.34	8970.52	20308.92	51758.54
OM 7	21626.74	8986.59	20333.26	51813.44
LCA 6	21448.98	8909.06	20225.49	51583.76
LCA 7	21441.70	8927.17	20230.11	51442.98



(a)



(b)

Figure 6: Predicted probabilities of "no education after age 15" using a multinomial logistic regression model for (a) the LCA-6 solution and (b) the LCA-7 solution

Discussion and conclusion

The key question discussed in this article is whether quite different approaches to develop life course typologies, sequence analysis (SA) and latent class analysis (LCA), lead to the same typologies, and whether it is possible to decide on which of these typologies is to be preferred.

We emphasise three main contributions of this article. First, we suggest a number of statistical tools that aid decision-making about the optimal number of clusters or classes. We propose that one should make use of a combination of statistical information and substantive interpretation. The choice of the number of clusters in SA can be facilitated by using cluster quality statistics – whereas the use of BIC plots supports the choice process in LCA.

The second contribution of this paper is its suggestion to consider cluster stability as an important aspect of cluster quality.

Our third, and major, contribution consists of our proposal to validate the obtained typologies by examining their association with other variables that are known or expected to be related to the pertaining life course trajectories. This approach is based on the idea of construct validity that is central to measurement theory. Analogously, we argue that one can evaluate the quality of a typology by examining how strongly it relates to other variables within its nomological network. This validation approach is not confined to comparing SA and LCA, but can also be used to compare typologies obtained by using other distance metrics within SA or by using other clustering methods applied to the same metric.

Our presentation of SA and LCA was illustrated by a substantive comparison. As we emphasised throughout our presentation, many different decisions have to be taken within each approach, and each one of them has to be based on a combination of substantive and statistical evidence. In Appendix 1, the main practical challenges facing both methods are discussed.

Our example illustrated the steps to be taken when performing SA and LCA. One of the interesting results is that the resulting typologies were quite comparable: the adjusted Rand-index is close to 0.7. Thus, the question arises whether or not this result is a coincidence. To shed light on this question, we elaborated on an old idea of Joseph Kruskal, and present our findings in Appendix 2. Our reasoning suggests that SA and LCA will most often lead to the roughly the same typologies.

In a recent article, Mikolai and Lyons-Amos (2017) compared SA to Latent Class Growth Models (LCGM), a type of LCA that takes the temporal ordering of events into account. Theoretically, LCGM's have the advantage over the simple LCA model that the former incorporates the ordering of events in the life course whereas the latter does not. In practical terms, estimating LCGMs with a larger number of observable states and almost ten times as many respondents is practically infeasible However, Mikolai and Lyons-Amos obtained interesting results comparing LCGM to SA and also showed that in practice, results are roughly the same.

Summarising, we tried to expand upon the pioneering research by Barban & Billari (2012) by proposing guidelines on performing and comparing SA- and LCA-based typologies and by introducing a number of useful statistical tools to aid in choosing between competing typologies.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 324178 (Project: Contexts of Opportunity. PI: Aart C. Liefbroer). The authors would like to thank the anonymous reviewers for their stimulating comments to earlier versions of this article.

References

Aassve, A., Billari, F. C., & Piccarreta, R. (2007). Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population/Revue Européenne de Démographie, 23*(3-4), 369-388. https://doi.org/10.1007/s10680-007-9134-6

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 16*(4), 129–147. https://doi.org/10.1080/01615440.1983.10594107

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History, 16*(3) , 471-494. https://doi.org/10.2307/204500

Agresti, A. (2002) *Categorical Data Analysis* (2nd Edition). Wiley, NJ. https://doi.org/10.1002/0471249688

Bahl, L. R., & Jelinek, F. (1975). Decoding for channels with insertions, deletions and substitutions with applications to speech recognition. *IEEE Tra'nsactions on Information Theory, 21*(4), 404-411. https://doi.org/10.1109/TIT.1975.1055419

Barban, N., & Billari, F. C. (2012). Classifying life course trajectories: a comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 61*(5), 765-784. https://doi.org/10.1111/j.1467-9876.2012.01047.x

Bayne, C. K., Beauchamp, J. J., Begovich, C. L. & Kane, V. E. (1980). Monte Carlo comparison of selected clustering procedures. *Pattern Recognition, 12*(2), 51-62. https://doi.org/10.1016/0031-3203(80)90002-3

Billari, F. C., & Liefbroer, A. C. (2010). Towards a new pattern of transition to adulthood?. *Advances in Life Course Research*, 15(2), 59-75. https://doi.org/10.1016/j.alcr.2010.10.003

Billari, F. C. & Piccarreta, R. (2005). Analyzing demographic life courses through sequence analysis. *Mathematical Population Studies, 12*(2), 81-106. http://dx.doi.org/10.1080/08898480590932287

Brzinsky-Fay, C. (2007). Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review, 23*(4), 409-422. https://doi.org/10.1080/08898480590932287409-422

Brzinsky-Fay, C., & Kohler, U. (2010). New Developments in Sequence Analysis. *Sociological Methods & Research, 38*(3), 359-364. https://doi.org/10.1177/0049124110363371

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. *The Stata Journal, 6*(4), 435-460.

Buchmann, M. C., & Kriesi, I. (2011). Transition to adulthood in Europe. *Annual Review of Sociology, 37*, 481–503. https://doi.org/10.1146/annurev-soc-081309-150212

Cornwell, B. (2015). *Social Sequence Analysis*. Cambridge (UK): Cambridge University Press. https://doi.org/10.1017/CBO9781316212530

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281. https://doi.org/10.1037/h0040957

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

Elzinga, C. H. (2005). Combinatorial representation of token sequences. *Journal of Classification, 22*(1), 87-118. https://doi.org/10.1007/s00357-005-0007-6

Elzinga, C. H. (2014). Sequence A152072. The On-Line Encyclopedia of Integer Sequences (2014), published electronically at http://oeis.org.

Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population/Revue Européenne de Démographie, 23*(3-4), 225–250. https://doi.org/10.1007/s10680-007-9133-7

Elzinga, C. H., & Studer, M. (2015). Spell sequences, state proximities and distance metrics. *Sociological Methods & Research, 44*(1), 3-47. https://doi.org/10.1177/0049124114540707

Elzinga, C. H., & Studer, M. (forthcoming). Normalization of distance and similarity in sequence analysis. *Sociological Methods & Research*.

Elzinga, C. H., & Wang, H. (2013). Versatile string kernels. *Theoretical Computer Science, 495*, 50-65. https://doi.org/10.1016/j.tcs.2013.06.006

Faul, F., Erdfelder, E., Lang, A-G. & Buchman, A. (2007) G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191. https://doi.org/10.3758/BF03193146

Festy, P., & Prioux, F. (2002). An evaluation of the Fertility and Family Surveys project. United Nations Publications.

Frahley, C. & Raftery, A. E. (1998) How many clusters? Which clustering method? Answers via Model Based Cluster Analysis. *The Computer Journal, 41*(8), 578-588. https://doi.org/10.1093/comjnl/41.8.578

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1–37. https://doi.org/10.18637/jss.v040.i04

Greenberg, R. I. (2003) Bounds on the Number of Longest Common Subsequences. arXiv:cs/031030v2[cs.DM] e-print

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511499531

Hand, D.J. & Yu, K. (2001) Idiot's Bayes – Not so stupid after all. *International Statistical Review*, 69(3), 385-398.

Helske, S., Steele, F., Kokko, K., Räikkönen, E., & Eerola, M. (2014). Partnership formation and dissolution over the life course: applying sequence analysis and event history analysis in the study of recurrent events. *Longitudinal and Life Course Studies, 6*(1), 1-25. https://doi.org/10.14301/llcs.v6i1.290

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis, 52*(1), 258–271. https://doi.org/10.1016/j.csda.2006.11.025

Hennig, C. (2008). Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis, 99*(6), 1154–1176. https://doi.org/10.1016/j.jmva.2007.07.002

Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin, 83*(6), 1072. https://doi.org/10.1037/0033-2909.83.6.1072

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218. https://doi.org/10.1007/BF01908075

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kleinepier, T., de Valk, H. A., & Gaalen, R. van (2015). Life paths of migrants: A sequence analysis of Polish migrants' family life trajectories. *European Journal of Population, 31*(2), 155–179. https://doi.org/10.1007/s10680-015-9345-1

Kruskal, J. B. (1983). An overview of sequence comparison: time warps, string edits and macromolecules. *SIAM Review, 25*(2), 201-237. ttps://doi.org/10.1137/1025045

Kuncheva, L., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *Systems, man and cybernetics, 2004 IEEE international conference* (Vol. 2), pp. 1214–1219. https://doi.org/10.1109/ICSMC.2004.1399790

Leary, M. R., Kelly, K. M., Cottrell, C. A., & Schreindorfer, L. S. (2013). Construct validity of the Need to Belong scale: Mapping the nomological network. *Journal of Personality Assessment, 95*(6), 610-624. https://doi.org/10.1080/00223891.2013.819511

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*(8), 707-710.

Liepins, G. E. (1980). Rigorous, systematic approach to automatic data editing and its statistical basis (No. ORNL/TM-7126). Oak Ridge National Lab., TN (USA). https://doi.org/10.2172/5518874

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249–264. https://doi.org/10.3102/10769986022003249

Linzer, D. A., & Lewis, J. B. (2011). polca: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1–29. https://doi.org/10.18637/jss.v042.i10

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook* (Vol. 2). Springer. https://doi.org/10.1007/b107408

Massoni, S., Olteanu, M., & Rousset, P. (2009). Career-path analysis using optimal matching and self-organizing maps. In Advances in Self Organizing Maps. proceedings of the 7th International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 2009 (Vol. 5629, p. 154-162). Springer. https://doi.org/10.1007/978-3-642-02397-2_18

McLachlan, G. & Peel, D. (2000) *Finite Mixture Models.* Wiley Series in Probability and Statistics. New York, Wiley. ISSN 1757-9597191. https://doi.org/10.1002/0471721182

Mikolai, J. & Lyons-Amos, M. (2017) Longitudinal methods for life course research: A comparison of sequence analysis, latent class growth models, and multi-state event history models for studying partnership transitions *Longitudinal and Life Course Studies*, 8(2), 191-208. https://doi.org/10.14301/llcs.v8i2.415

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179. https://doi.org/10.1007/BF02294245

Mills, M. (2011). *Introducing survival and event history analysis*. London: Sage. https://doi.org/10.4135/9781446268360

Moen, P., Kelly, E. & Huang. R. (2008). Fit inside the work-family black box: An ecology of the life course, cycles of control reframing. *Journal of Occupational and Organizational Psychology, 81*(3), 411-433. https://doi.org/10.1348/096317908X315495

Pailhé, A., Robette, N., & Solaz, A. (2013). Work and family over the life course. A typology of French long-lasting couples using optimal matching. *Longitudinal and Life Course Studies, 4*(3), 196-217.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

Rennie, J. D., Shih, L., Teevan, J. & Karger, D.R. (2003) Tackling the poor assumptions of naïve Bayes text classifiers. In Fawcett, T. & Mishra (Eds.) *Proceedings of the Twentieth International Conference on Machine Learning 2003* (ICML-2003), pp 616-623.

Ristad, E. S., & Yianilos, P. N. (1998). Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(5), 522-532. https://doi.org/10.1109/34.682181

Ritschard, G., Gabadinho, A., Studer, M., & Müller, N. S. (2009). Converting between various sequence representations. In *Advances in Data Management* (pp. 155–175). Springer. https://doi.org/10.1007/978-3-642-02190-9_8

Robette, N. (2010). The diversity of pathways to adulthood in France: Evidence from a holistic approach. *Advances in Life Course Research, 15*(2–3), 89-96. https://doi.org/10.1016/j.alcr.2010.04.002

Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodology Sociologique, 116*(1), 5-24. https://doi.org/10.1177/0759106312454635

Rossi, P. H., Wright, J. D., & Anderson, A. B. (2013). *Handbook of survey research*. Academic Press.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review, 17*(2), 119–144. https://doi.org/10.1093/esr/17.2.119

Scholtus, S. (2014). Error localization using general edit operations (Discussion Paper No. 2014-14). The Hague: Statistics Netherlands. (Available at http://www.cbs.nl)

Studer, M. (2012). Analyse de données séquentielles et application à l'étude des inégalités sociales en début de carrière académique (Doctoral dissertation, PhD Thesis, Genève: Université de Genève).

Studer, M. (2013). Weighted cluster library manual. LIVES Working Papers , 24 , 1-32.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481-511. https://doi.org/10.1111/rssa.12125

Vermunt, J. K. & Magidson, J. (2003) Latent class models for classification. *Computational Statistics & Data Analysis, 41*(3-4), 531-537. https://doi.org/10.1016/S0167-9473(02)00179-2

Vermunt, J. K., & Magidson, J. (2005). Latent GOLD 4.0 User's Guide. Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. Belmont, MA: Statistical Innovations Inc.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Watson, M. W., & Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics, 23*(3), 385-400. https://doi.org/10.1016/0304-4076(83)90066-0

Wu, C. J. (1983). On the convergence properties of the EM-algorithm. *Annals of Statistics, 11*, 95-103. https://doi.org/10.1214/aos/1176346060

Zhang, H. (2005) Exploring conditions for the optimality of naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence, 19*(2), 183-198. https://doi.org/10.1142/S0218001405003983

Endnotes

[i] In the present context, picking the "right" clustering technique is an unsolved problem since there is no generally accepted idea about the "structure" of life-course sequence data. Without such an idea, the choice of a clustering method is more or less arbitrary. For example, we know that some clustering techniques are well suited for sets of variables that have non-elliptical multivariate normal distributions with equally sized subpopulations (most hierarchical methods) or work well with particular distance measures (PAM). However, in the context of SA, we do not directly observe these underlying variables but only a diffuse summary measure: a distance between sequences. Bayne, Beauchamp, Begovich and Kane (1980), using bivariate distributions, tested thirteen different techniques for their classification accuracy. Their last, concluding sentence is "*However, as the complexity of the distributions increases, the differences between all of these methods decrease*". Unfortunately, this sentence still well summarises the state of affairs in unsupervised partitioning of distance matrices. Therefore, it is not surprising that often, in the present context, (agglomerative) hierarchical clustering is chosen: most people enter a phase of family formation during their early adulthood and, most often, this involves partnering and reproduction. It is not unreasonable to consider variation in this general pattern has a hierarchical structure and thus a hierarchical clustering seems warranted. Moreover, hierarchical techniques have the advantage of easily visualisable results in the form of a dendrogram. Although PAM is a good alternative, we decided to use Ward's agglomerative hierarchical method since it is by far the most frequently chosen. Of course, this method has disadvantages, which have been amply documented in the vast literature on clustering methods. Unfortunately, good alternatives like PAM also have disadvantages. Therefore, whichever method is picked, substantive validation of a cluster solution is of vital importance. Comparing different clustering methods in the context of SA is, even if possible at all, beyond the scope of this paper.

[ii] Latent class analysis is the simplest form of all latent structure models. There are other models that consider time-dependence of the sequences, such as Markov models. However, LCA is sequence oriented with the fewest assumptions (local independence). Therefore, it is the only latent structure model that can be sensibly compared to SA.

[iii] Country weights were not available for all countries. Given the illustrative purpose of our example, we decided to use unweighted data.

The impact of parental employment trajectories on children's early adult education and employment trajectories in the Finnish Birth Cohort 1987

Pasi Haapakorva The National Institute for Health and Welfare and University of Oulu, Finland
pasi.haapakorva@thl.fi
Tiina Ristikari The National Institute for Health and Welfare, Finland
Mika Gissler The National Institute for Health and Welfare and University of Turku, Finland, Karolinska Institute, Sweden

(Received December 2016, Revised June 2017) http://dx.doi.org/10.14301/llcs.v8i4.441

Abstract

The Finnish Birth Cohort 1987 grew up during the recession that hit Finland in the early 1990s, which had an impact on their parents' activity in the labour market. In this paper we use Finnish register data to build employment and education sequences for all young people born in Finland in 1987 for the period 2005–2012 and employment sequences for all their parents for the entire length of their children's lives from 1987 until 2012. The sequences were analysed and clustered, and four multinomial logistic regression models were used to find how parents' trajectories connect to their children's early adulthood trajectories. Most parents had been on a stable employment trajectory, but we found mothers and fathers who were absent from the labour market during the recession of the 1990s and after it – and some parents never entirely returned to work during this 1987–2012 follow-up. Likewise, most children were either on an employment or education trajectory, but we found groups of children who were on very early child care trajectories, unemployment trajectories, or on a trajectory with no records in the Finnish registers, which in the Finnish context implies that those young people are not employed, not in education and not receiving any of the various benefits. Disadvantageous trajectories were mostly very lasting. We found strong connections between parents' disadvantages in the labour market and children's disadvantageous early adulthood trajectories, even when adjusting for strong background variables. The strongest connections arise from parents' long absences from the labour market.

Keywords

Sequence analysis, unemployment, employment, education, life course, trajectory, birth cohort, education, recession, register data

Introduction

Early adulthood is a period during which individuals begin to move into economic independence. In most industrialised countries, including Finland, this is a period when most further education or labour market entry takes place. This period and the transitions that it involves during youth are impacted by the children's family's educational background (Witting & Keski-Petäjä, 2016). Disadvantageous early adulthood transitions may have their roots in childhood (Caspi, Wright, Mofitt & Silva, 1998) and we know that family background, including parents' educational level and socioeconomic status, continues to have a strong influence in early adulthood regardless of children's educational level (e.g. Barone & Schizzerotto, 2011; Bukodi & Goldthorpe, 2011; Härkönen & Bihagen, 2011; Mastekaasa, 2011; Sirniö, Kauppinen & Marttinen, 2016).

When thinking of mechanisms for how family background affects later life, we take a broad viewpoint, using life course theory (Elder, Johnson & Crosnoe, 2003), which states that individuals make choices in the historical and cultural circumstances in which they find themselves at the time. During childhood, parents create much of those circumstances. Parental poverty (Bäckman and Nilsson, 2011) and unemployment (Gray & Baxter, 2012; Rege, Telle & Votruba, 2007) could affect children's school achievement and later labour market integration. On the other hand, parental unemployment can also raise children's educational aspirations, as children come to recognise the value of higher education (Schoon, 2014), and during hardship, a high parental educational level also fosters children's aspirations (Mortimer, Zhang, Husseman & Wu, 2014).

Previous research related to early adult trajectories—sometimes referred to as school-to-work transitions—has established that in addition to parental educational attainment, several other family-related structural variables, such as poverty and socioeconomic status, have an impact, either directly or indirectly, on these transitions. Previous research has also established connections between several individual characteristics, such as sex and school achievement (Brzinsky-Fay, 2015; Larja et al, 2016; Ristikari et al., 2016; Schoon, 2014). What is less clear is how parental labour market activities impact early adulthood trajectories above and beyond the impact of parental education and other known background variables. No study that we are aware of has measured parental labour market activity longitudinally.

In this study we see early adulthood as a social sequence, which contains ordered set of states (Cornwell, 2015). By analysing the sequences (sequence analysis) we are able to see the transitions holistically and in all their diversity, and look for structure in the mass of information (Brzinsky-Fay, 2011; Schoon & Lyons-Amos, 2016). Many researchers have successfully utilized sequence analysis in the study of school-to-work transitions and life course research in general (Brzinsky-Fay, 2007; Brzinsky-Fay, 2015; Brzinsky-Fay & Solga, 2016; Ilmakunnas, Kauppinen & Kestilä, 2015; McVicar & Anyadike-Danes, 2002; Sackmann & Wingens, 2003).

The Finnish context for early adulthood employment and education trajectories

In Finland, comprehensive school achievement strongly shapes future educational paths. Comprehensive school ends at age 15–16, when most young people make a choice between two types of upper-secondary-level education: upper-secondary-level vocational education and general upper-secondary-level education. The former most often requires significantly lower comprehensive school achievement than the latter, while the latter much more often leads to academic education. In the Finnish Birth Cohort 1987 (FBC 1987), previous research shows that children with a high parental level of education are much more likely to attain a post-compulsory degree than children whose parents have only a comprehensive level education. Parental level of education, along with child's sex, is strongly linked to children's achievement in compulsory school (Ristikari et al., 2016).

Labour markets have changed across cohorts, affecting young people's transition to work. Early employment trajectories have not been simple pathways for some time. Cyclical economic turns, involving downswings and upswings, affected young people entering the labour market already in the 1990s (Gangl, 2002; Zwysen, 2014). At the same time, labour market entry is a major determinant of later labour market integration. Those who enter as unemployed are more likely to become unemployed later (Steijn, Need & Gesthuizen, 2006). In Sweden, those who entered the labour

market during the economic recessions of the mid-1970s and the early 1990s have suffered long-term deficits in career progression (Härkönen & Bihagen, 2011). Mroz and Savage (2006) have shown with US data that youth unemployment causes a catch-up response, increasing the likelihood of training to mitigate the setbacks, but young people do not fully recover from the effects of unemployment. In Finland, during the recessions of the early 1990s and the late 2000s, the employment levels of those aged 15–24 and 25–34 fell more steeply than those of people older than 35. In fact, the only age group in Finland with a positive employment curve during the period 2008–2015 was those aged 55–64 years (Official Statistics Finland: Labour Force Survey, 2015). The latest "Great Recession" (Kangas & Saloniemi, 2013) likely impacted the FBC 1987.

Long-term changes in the Finnish labour market for parents

Long-term changes in the Finnish labour market have been affecting the FBC 1987 parents, whose earliest employment records go back to the 1940s. According to a review by Kangas and Saloniemi (2013) on the history, current situation, and future of the Nordic model in Finland, a number of issues are of great relevance here. First, the structural transformation of the economy in Finland has been late but rapid. Agricultural labour remained important for several decades longer than in other Nordic countries. Primary production was the dominant labour branch up until the late-1950s, but was in strong decline after the 1940s, making up a quarter of the economy in 1970, 14 percent in 1980, and down to 6% in 2000 and 4% in 2012. The decline affected both sexes. The decline in manufacturing jobs, however, was steeper among women than men: the number of women working in manufacturing more than halved in the period 1980–2012, while men dropped by a fifth. Meanwhile the tertiary sector has seen steady growth, with women having a two-thirds majority since the 1970s. The decline in primary production and manufacturing and the growth of the tertiary sector has strongly affected labour force demand by sex and educational level.

Second, the economic depression of the 1990s hit Finland the hardest among the Nordic countries. GDP saw three negative years after stagnation in 1990 and unemployment rose from 3% to 16%. At the end of the millennium, due in part to the structural transformation of the Finnish labour market, Finland faced simultaneously both unemployment and a labour shortage due to skills mismatch. Male dominant branches were hit hardest and women found it easier to find new jobs in the service sector; the same pattern was also true in the 2008 economic crisis. In 2008, the employment rates declined among all men, recovering in the following year for all but those with only a comprehensive education (Kangas & Saloniemi, 2013). These two factors combined, late but rapid decline in traditional sectors followed by skills mismatch, have likely led to a disappearance of job opportunities for the FBC 1987 parents.

Objectives

Our first objective is to identify the most significant early adult trajectories in an entire Finnish birth cohort born in 1987, using rich sequential register data, with sequence analysis and clustering methods. Due to register data availability and completeness (in terms of events relating to moving into adulthood and the length of the FBC 1987 follow-up at present), we limit our study of early adulthood to 18–25 years of age (2005–2012). We expect to find many youth still in education in 2012, but many have already taken up employment after finishing upper-secondary-level vocational school at 19 years of age. We are paying close attention to those young adults who have not had much or any post-compulsory education or employment history. One of the strengths of register data is that we always at least know what people are not doing.

The second objective is to identify parents' most common employment trajectories during the lives of their children born in 1987, following them up to 2012, a period of 26 years, using equally rich register data and sequence analysis methods. We are not considering the parents' age, although we will know if they have retired or are deceased.

What we also seek to explore, and which is our third objective, is how much the children's early trajectories are affected by their parents' employment trajectories. We will look at how unemployment and activities outside the labour market at distinct times affect children, while controlling for parents' education and marital status, children's sex, residence abroad and comprehensive school achievement. We include residence data to explain some missing register

data. The next section will cover the data and methods, the third section will present the results, and the fourth section includes a discussion.

Data and methods

The Finnish Birth Cohort 1987 (N = 59 476) includes data from several registers from both children born in 1987 and their parents. In Table 1 we present the data we have included in the analyses for this paper. All the mothers are present in the data, but 821 children had no registered father. In total, we analyse trajectories for 177,604 individuals. The Finnish Birth Cohort 1987 (Paananen & Gissler, 2011; Ristikari et al, 2016) consists of data from several register holders. In this study, the data on employment, pensions and benefits are from the Finnish Centre for Pension (ETK), and we have added study grant data from the Social Insurance Institution of Finland (Kela), unemployment data from the Finnish Ministry of Economic Affairs and Employment, and finally social assistance data from the Finnish National Institute of Health and Welfare (THL). All linkages were done by using parents' and children's personal identity codes available in all Finnish registers.

For the sequence analysis, we studied the labour market statuses for 96 months for the children, in the period 2005–2012, and 52 half-year periods for their parents, for the period 1987–2012. The children's data are characterised by a substantial number of brief periods of employment or benefit periods, as short-term employment is common during education among young adults. For the children, we look at cumulative periods of at least 10 days per month and for the parents we account for all the cumulative periods with a length of at least one month. It is common to find many overlapping statuses for the study periods, which compelled us to prioritize. Above all others was death, and we prioritised benefit periods that commonly occur during periods of employment, education, and entrepreneurship. Unemployment, different types of pensions and social assistance come last. An employed person receiving a study grant is considered studying, likewise an employed person receiving childcare or sickness benefits is not considered employed (a person receiving a parental or sick leave benefit has their wage suspended but they are allowed to return to their job after the benefit period). However an employed person is never considered unemployed or drawing a pension, but pensions and unemployment benefits are considered over social assistance. A majority of parents' benefit data are not available in the FBC 1987 from 1994 to 2004, but employment and pension data are complete. Data availability improves in 2005, when a new benefit register at Centre for Pensions was formed.

The status "missing" is meaningful in this study. In the sequence analysis, we handle it like all other statuses. Children's missing status denotes not being in employment, education or in receipt of any of the many benefits. The same is true for parents, except for the period before 2005 where only data on employment and pensions are available.

In the analyses, we employ several background variables. We use children's sex, comprehensive school achievement and in part, their residence history for residence abroad to account for some of the missing data. We measure parents' marriage when their children were age 22 (2010) and educational level at children's age 21 (2009). Education is measured in four levels: comprehensive (compulsory education only), upper-secondary (general upper-secondary-level or vocational education), lower academic (community/junior college level, "opistoaste", obsolete but common among parents) and higher academic level (bachelor's and above).

Table 1: Data sources for the sequence analysis and the regression analyses

Register holder	Data	Availability
Centre for Pensions	Employment	–2012 (c + p)
	Entrepreneurship	–2012 (c + p)
	Alternation leave	2005–2012 (c + p)
	Child care (all)	2005–2012 (c + p)
	Sickness benefit	2005–2012 (c + p)
	Accident benefit	2005–2012 (c + p)
	Rehabilitation	2005–2012 (c + p)
	Unemployment benefit	2005–2012 (p)
	Old age pension	–2012 (p)
	Disability pension	–2012 (c + p)
	Other pension	–2012 (p)
Ministry of Economic Affairs and Employment	Unemployment periods	1999–2015 (c)
The Social Insurance Institution of Finland	Study grant	2005–2012 (c)
	Unemployment benefit	1987–1994 (p)
	Child care (parental leave)	1987–1994 (p)
National Institute of Health and Welfare	Social assistance	1988–2012 (c)
	Identity number (sex)	1987 (c + m)
Statistics Finland	Educational level	2008 (p)
	Date of death	–2012 (c +p)
Population Register Centre	Place of residence	1987–2012 (c)
	Marriages and divorces	–2012 (p)
	Identity number (sex)	1987 (f)
The Finnish National Board of Education	Comprehensive school achievement	2003– (c)

c = child, p = parents, m = mother, f = father
Availability of data from different registers varies, and for the parents a majority of the benefit data are not available in the FBC 1987 from 1994 to 2005.

The data on comprehensive school achievement (average of all awarded grades on a scale of four to 10 at the end of comprehensive school, Table 2) is derived from the children's applications for further education, and in 4.4% of the cases, it is missing due to one of two reasons: they have applied earlier than average, for which we do not have data, or they have never applied, at least not through the centralized system. Of those 4.4% with missing achievement records, 46% have obtained a post-comprehensive degree by the end of 2012. We use multiple imputations (10 imputations, 20 iterations, predictive mean matching) using the 'mice' package (Van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2015) to account for the missing data. We have included all the variables present in the regression models along with the degree data in the imputation model to achieve the strongest possible imputation results. We did not impute achievement records for those who had died.

Table 2: Children's comprehensive school achievement* by sex

	n	mean	SE	missing
Female	27,775	8.07	0.006	1,266
Male	29,084	7.49	0.006	1,351
Total	56,859	7.77	0.004	2,617

* Average of all awarded grades at the end of comprehensive school. Scale 4–10.

Some children have no registered father, and some have confidential residence histories, for example, due to their occupation, status, or threats to themselves or to their family. We have accounted for missing data by adding a category for the missing data.

For the distance matrix in sequence analysis, we use the hamming distance method with a substitution cost of 1, and for clustering we use the Ward method along with the non-hierarchical PAM method (Studer, 2013). The children's data are split into two due to large size, so we have created a set of clusters from both and matched them cluster-to-cluster to create a whole data set (see Appendix for more information about the cluster analysis). For the children, we chose a solution with 12 clusters, including one for deceased persons and another for the ones with no or little data. The mothers' data were split into seven clusters, which included five clusters between steady career and death, and for the fathers, 10 clusters were sufficient to capture most of the variance.

To examine the relationship between parents' and children's trajectories we use multinomial logistic regression. The exponentiated coefficients are relative risk ratios (RRR). We have included both parents' clusters in the model as independent variables along with control variables. We will present four regression models. The first model includes only the parents' trajectories as independent variables. In the second model we add parents' highest level of education, in the third model we add children's sex, residence abroad, and parents' marital status, and finally in the fourth model we add the children's average grade at the end of comprehensive school. The results in the fourth model are pooled from the imputed datasets. Due to the multiple imputations, those children who had died before the end of 2012 were left out of all the models. We do not account for a possible sibling effect in the models, because only one per cent of the cohort share a parent, and our data include no data on children born in other years.

Results

Cluster results

Two sets of results are presented here. First, we take a brief look at the data, next a look at the sets of clusters we have created from the children's (C indicating children) and their parents' (M indicating mother, F for father) data, and finally, we present the results of the multinomial logistic regressions.

Finnish young people stay in school for a relatively long time. Compulsory school ends at age 16, but most attend non-compulsory upper-secondary-level school immediately after, which usually lasts for three years. At this point it is common to take a gap year. The Finnish Birth Cohort 1987 finished their upper-secondary-level degrees in spring 2006. A year after, in May 2007, 34% of cohort members had no data in these registers, meaning they were not in employment, education or on any of the many benefits. Another year later, 14% remained in the no-data group and in 2009 some 10% remained. The share of 'no-data' settles down to about 8% at each cross-section.

At the other end of the series, in the beginning of 2012, 29% of cohort members are on a study grant, down from 37% the year before. In 2009–2010 cohort members on a study grant peaks at 42%. Meanwhile cross-section employment at the end of each year steadily rises to 53% in 2012.

We have chosen a rather large set of clusters for our subjects, the children and their fathers and mothers. There are several reasons for this. First, our data set is relatively large, so less common trajectories can be separated into clusters. Second, as we show later in the results, the parent–children trajectory connections are far from random: we obtain stronger results with more clusters. In the parents' case, we use time-variance in the unemployment and pension clusters to identify the effect of parental labour-market inactivity at different times during the children's life-course,

resulting in a few more clusters. Third, real-life trajectories are very complex, and with this data we are able to capture a few more of them than previously was possible.

Table 3: Children's cluster results by sex, C1–C12

Group	Females	%	Males	%	Total	%
C1. Education	10,307	35.5	8,689	28.5	18,996	31.9
C2. Education with secondary level study grant	5,308	18.3	3,316	10.9	8,624	14.5
C3. Employment	2,418	8.3	7,030	23.1	9,448	15.9
C4. Employment with secondary level study grant	2,314	8	4,077	13.4	6,391	10.7
C5. Employment after difficulties	649	2.2	1,370	4.5	2,017	3.4
C6. Early child care	2,548	8.8	16	0.1	2,564	4.3
C7. Late child care via employment or education	2,321	8	347	1.1	2,668	4.5
C8. Fragmented employment / unemployment with secondary level study grant	327	1.1	900	3	1,227	2.1
C9. Fragmented unemployment / income support	363	1.2	796	2.6	1,159	1.9
C10. Fragmented no data / education / employment	1,543	5.4	2,229	7.3	3,772	6.3
C11. Not much or no data	745	2.6	1,259	4.1	2,004	3.4
C12. Deceased	198	0.7	406	1.3	604	1
Total	29,041	100	30,435	100	59,476	100

In the children's clusters (Table 3, Figure 1) we find that it is necessary to keep some clusters split based on whether there is an upper-secondary-level study grant during 2005 and in the first half of 2006 or not. This benefit is granted based on parents' income. By keeping respondents with and without study grants separate we not only distinguish young people who are well off from those who are less well off in their youth; this distinction enables us to discern the paths of both groups more accurately.

The first four clusters include the main education and employment trajectories. Most children born in 1987 finish their upper-secondary-level education in spring 2006 and some will begin employment or education during the following months. However, it is common to take one or two gap years working, applying to schools, or in the military. Most males and some females spend at least half a year in the Finnish army, but this is not designated in our data.

After the education and employment clusters, we find a cluster in which the share of employed reaches 80% only at the end. We named this cluster C5 'employment after difficulties', since employment shows a steady rise up to late 2008, when the latest recession hit Finland, and declines, though it picks up again after a few years. Following that, based on the study period, we have early and late child care, clusters C6 and C7. It is worth noting that nearly a fifth of FBC 1987 females were on either of these trajectories (9% and 8% of females, respectively), but only 1% of males have spent considerable lengths of time receiving child care related benefits.

Clusters C8 to C11 include those with more NEET-type (Not in Education, Employment or

Training) trajectories. In the eighth cluster, unemployment varies with employment, and in the ninth, with social assistance. In the 10th cluster, trajectories show variation between education, work, unemployment and a sizable share of no data which stays at around 30% during the second half. Individuals having little or no data at all, at least not in these registers, are clustered in the 11th cluster. The last one is reserved for the deceased. All these clusters have a male-majority.

We find children's NEET-type trajectories, the unemployment and the no-data trajectories, very stable over a period of five to eight years.

The mothers' (Table 4, Figure 2) and fathers' (Table 4, Figure 3) employment trajectories have been divided in much the same way. In the first place, we have those with continuous careers, while the last clusters are for the deceased. Between them we first have labour market inactivities of different timings and lengths, followed by pensions. In both series of clusters, before the deceased, we have a cluster of individuals having little or no data.

Both mothers and fathers show clear signs of the recession that hit Finland in the early 1990s. The fathers' data also show signs of the latest recession in 2008, while the mothers' data do not, consistent with its greater impact on male-dominated industries. The fathers' data are split into more clusters, 10 versus seven for mothers, since fathers have more complex pension trajectories than mothers, with two clusters for disability pension and an old age pension cluster, while more than twice as many fathers have died than mothers, resulting in deaths split into two clusters (early and late).

A prominent feature of the parents' clusters is the missing benefit data between the mid-1990s and the mid-2000s. While we cannot take a deeper look into unemployment during this period, we do have employment and pension data, which allows us to consider different lengths of labour market inactivities.

Table 4. Parents' clusters, educational level, marital status, and children's residence abroad

	%	n
Mother M1: Employment or entrepreneurship)	68.2	40,550
M2: Outside of labour market in the 90's	16.7	9,928
M3: Long gap in labour market activity	5.3	3,164
M4: Unemployed throughout	2.6	1,526
M5: Disability pension	2.5	1,490
M6: Not much or no data	3.5	2,109
M7: Deceased	1.2	709
Total	100	59,476
Father F1: Employment or entrepreneurship)	72.2	42,952
F2: Unemployed in the early 1990s	4.3	2,537
F3: Unemployed starting in 2009	3.2	1,879
F4: Unemployed after early 90's	2.9	1,750
F5: Early disability pension	2.7	1,594
F6: Late disability pension	2.5	1,490
F7: Retirement pension after work	3.2	1,928
F8: Not much or no data	3.3	1,967
F9: Early death	2.0	1,207
F10: Late death	2.3	1,351
F11: No registered father	1.4	821
Total	100	59,476
Parents' education (2008): Higher academic level	25.9	15,412
Comprehensive level	6.9	4,128
Upper secondary level	43.1	25,639
Lower academic level	24.0	14,297
Total	100	59,476
Parents' marriage (2009): Parents married	57.8	34,352
Parents not married	42.2	25,124
Total	100	59,476
Residence abroad at any time: Never abroad	95.0	56,531
Abroad at some point	4.7	2,788
Residence history not available	0.3	157
Total	100	59,476

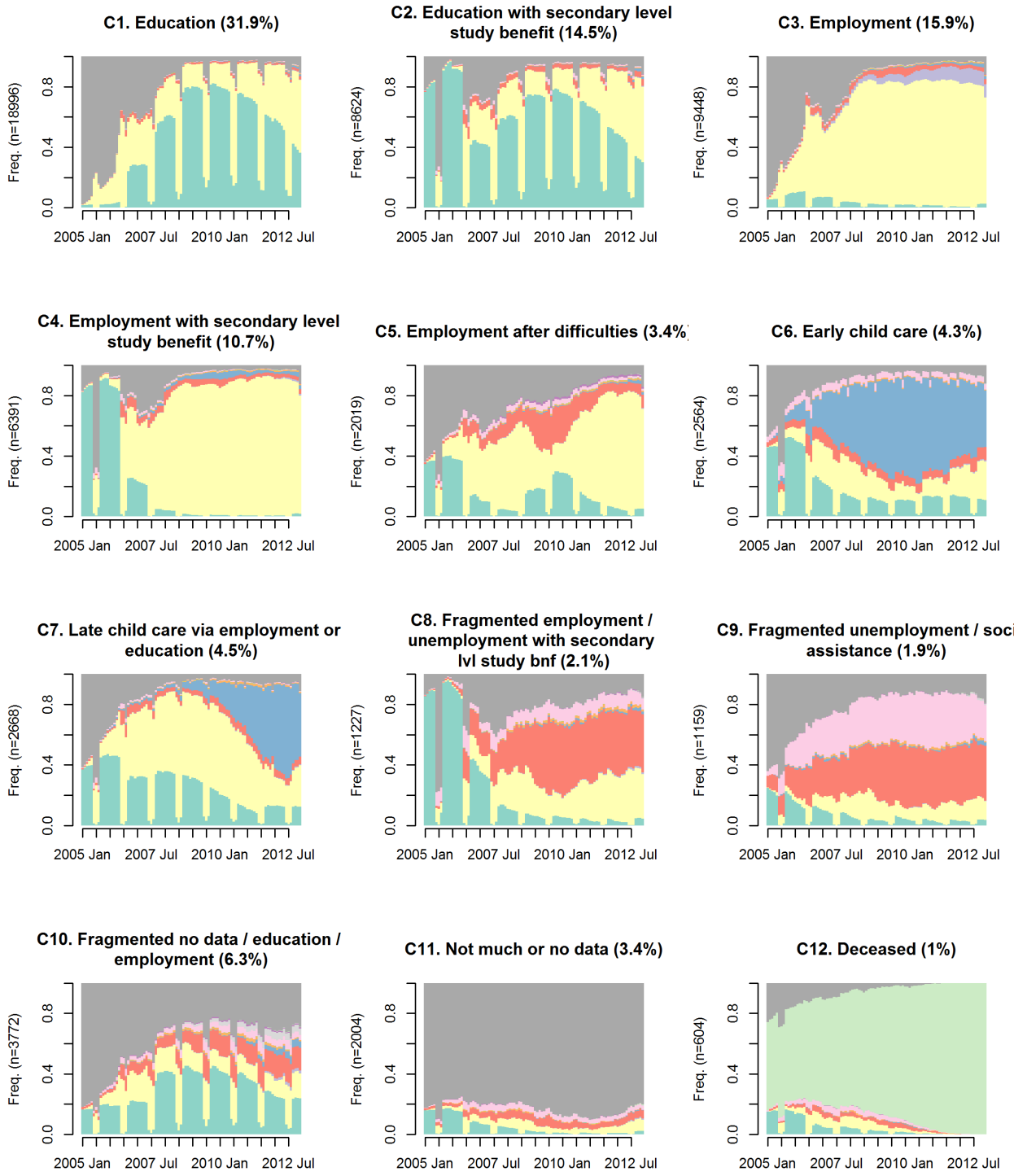


Figure 1. State distribution plot of children's clusters

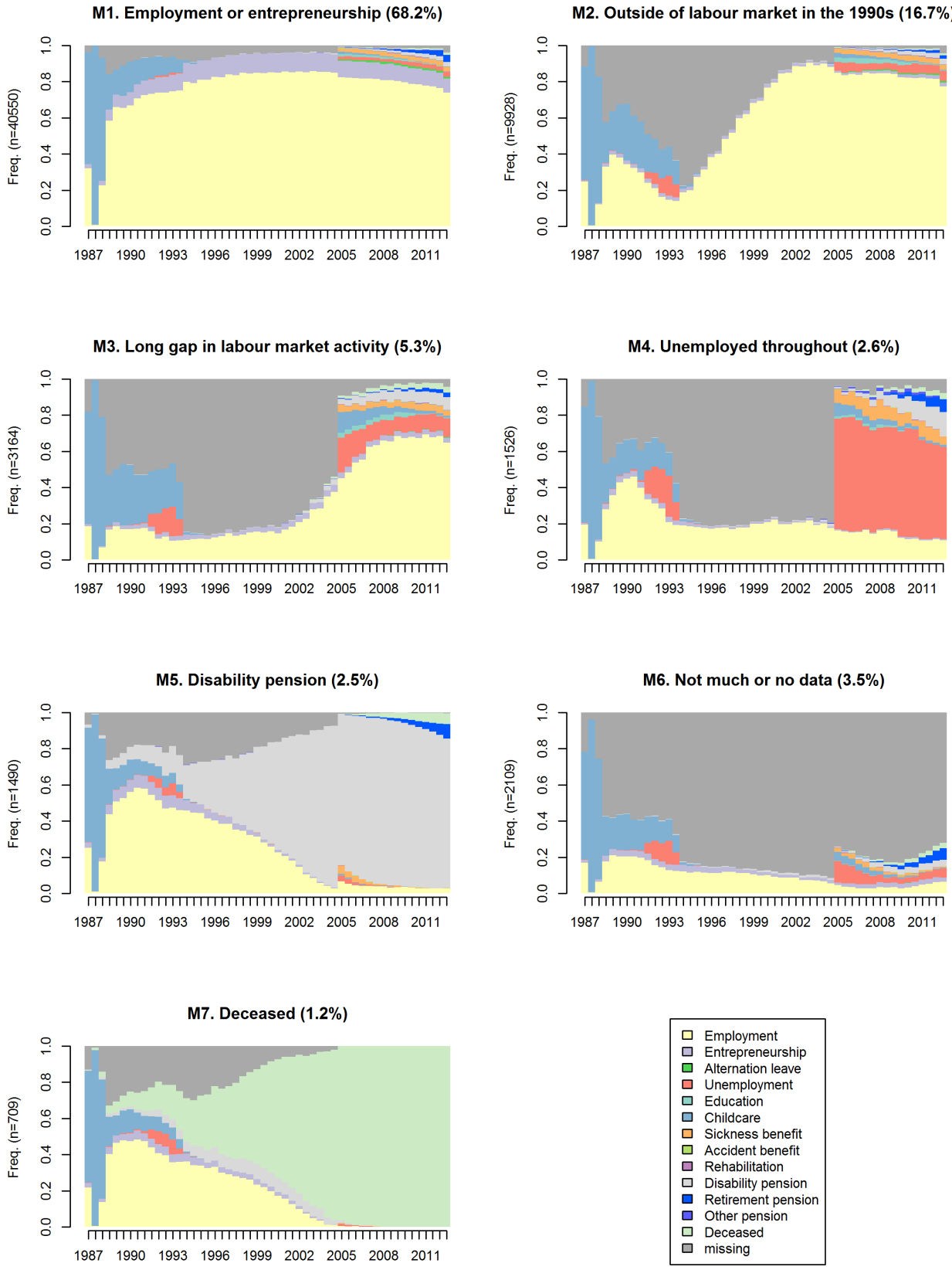


Figure 2. State distribution plot of mothers' clusters. Parents' data on other than on employment, pensions and deaths are missing between 1994 and 2004

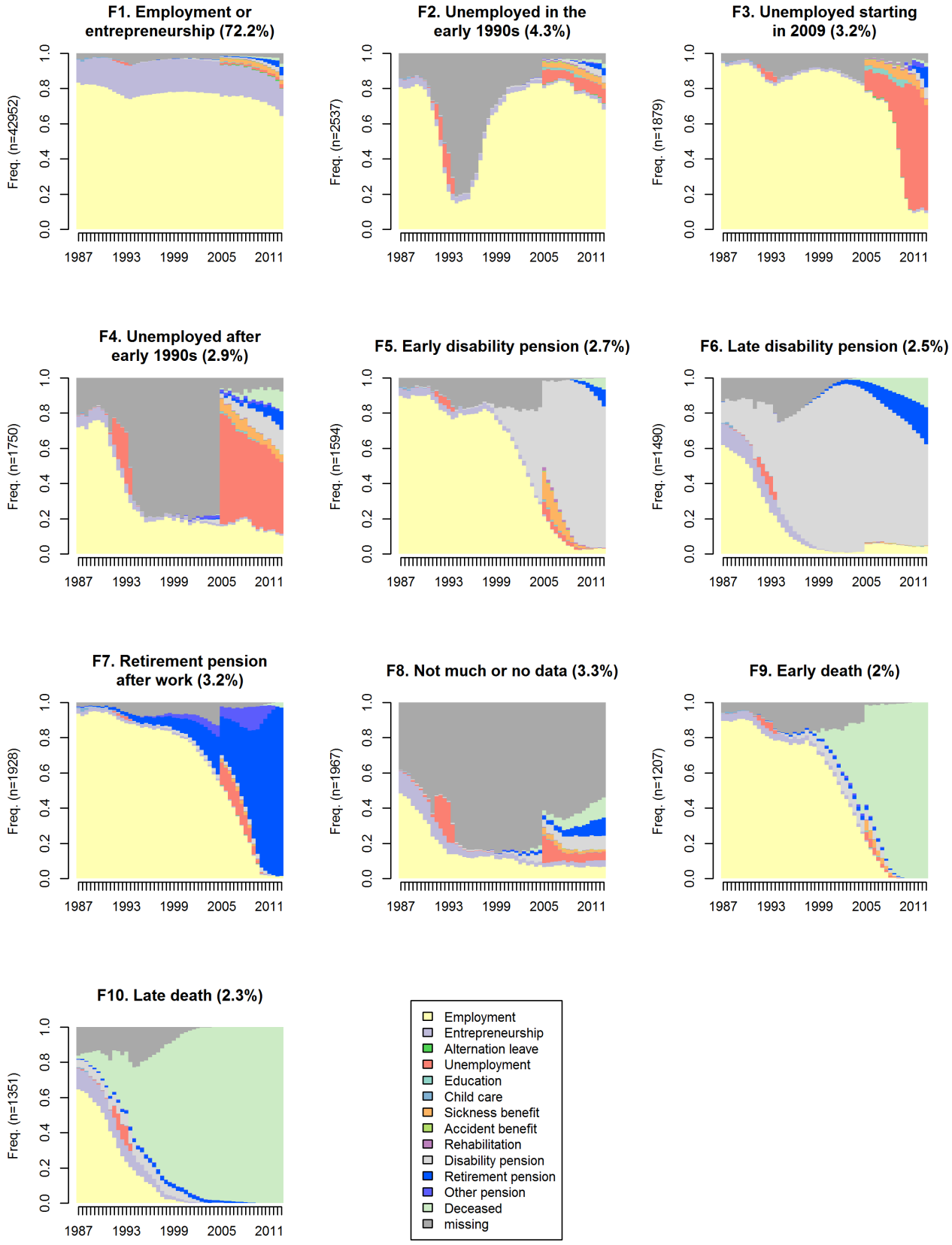


Figure 3. State distribution plot of fathers' clusters. Parents' data on other than employment, pensions and deaths are missing between 1994 and 2004

Regression results

We found strong connections between the parents' and children's trajectories (Table 5). The connections remain even when we bring in strong background and predictor variables in regression models 2–4.

Education and employment trajectories

The reference cluster in the children's trajectory clusters is education. In clusters C2–C5, where we first have education with an upper-secondary-level study grant and subsequently three employment clusters, we can see the interplay between parental trajectories and having received an early study grant and also choosing employment over further education.

In both the upper-secondary-level study grant groups we see a strong connection to parents' divorce or parents never marrying. The study grant indicates a lower level of household income due to the upper secondary level benefit system in Finland; thus marriage between parents' links to a financially more secure youth.

In education cluster C2, parental employment trajectories show mostly highly significant connections with having received an upper-secondary-level study grant. We also see a strong connection to unmarried parents. Females take this path more often than males, but sex is not significant in the last model (model 4, male RRR 0.77, 95% CI [0.58, 1.00]). Of the parents' unemployment trajectories, longer inactivities have stronger connections than short inactivities. Father's early disability pension F5 has a stronger connection than F6, indicating late disability pension. For mothers, the relative risk ratio for any disability pension, M5, remains between these two, resulting possibly from an averaging out of the timing of disability pensions (one disability pension group vs. early and late). Interestingly, both mothers' and fathers' groups (M6 and F8) with no data show slightly weaker connections to the second child cluster than long inactivities, though still significant.

Parental employment trajectories do not show many significant connections to children being in employment over education in cluster C3. The lack of an early study grant indicates financially more stable youth, so other factors are at play here. The strongest connections are with male sex, average school grade, and parents' level of education. When we move on to employment cluster C4 (with a study grant), we see that the parental employment trajectory connections become significant, with a strong connection also to unmarried parents and parents with a low level of education, possibly indicating a low household income during youth resulting from a combination of these factors. Parents' low education level has a stronger connection to employment cluster C4 than to education cluster C2.

Employment cluster C5 (employment after difficulties) shows weaker but still significant connections to parents' trajectories. In this cluster the grouping is a result of a similar situation in the labour market during the 2008 recession and before. The connections to family background here might be more random.

If we now look at the transition from regression model 3 to model 4, where we add the average school grade to the model, we see that the effect of parental education level in the regression model is weakened, much more so than is parental labour market inactivity. Some long periods of parental inactivity connections even become stronger in model 4. This indicates two things: First, parental level of education is a strong predictor of school achievement, and that works here by weakening the statistical connection between parental education level and their children's school achievement related trajectories. School achievement itself strongly predicts the children's trajectory, since upper-secondary-level vocational education most often implies lower comprehensive school achievement than general upper-secondary-level education. General upper-secondary-level education commonly leads to further education, while vocational school rarely does (though further education is at present possible) and instead leads to relatively early employment. Second, parental labour market inactivity trajectories are strong predictors of the children's trajectories regardless of other factors, and there is no strong correlation between the parents' employment trajectories and children's comprehensive school achievement.

What this means in real life is another matter. School achievement and aspirations have been shown to be a mediator of childhood circumstances (Bäckman & Nilsson, 2011; Schoon, 2014). Here we see that parental employment trajectories have a strong effect regardless of school achievement and parental level of education (and other predictors), while parental education level more often works via children's school achievement.

Table 5. Regression results for models 1–4. Relative risk ratios

	C2. Education with 2nd lvl benefit				C3. Employment				C4. Employment with 2nd lvl benefit				C5. Employment after difficulties			
	1	2	3	4[1]	1	2	3	4[1]	1	2	3	4[1]	1	2	3	4[1]
Mother (ref. Employment/entrepreneurship)	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
M2. Outside of labour market in the 1990s	1.42***	1.36***	1.32***	1.3**	1.28***	1.21***	1.2***	1.13	1.63***	1.49***	1.42***	1.32***	1.48***	1.4***	1.38***	1.3**
M3. Long gap in labour market activity	3.23***	2.94***	3.26***	3.54***	1.65***	1.4***	1.4***	1.39***	4.45***	3.56***	3.75***	3.69***	3.37***	2.91***	2.98***	3***
M4. Unemployed throughout	4.98***	3.9***	2.89***	3.35***	2.18***	1.48**	1.22	1.2	7.27***	4.53***	3.1***	3.03***	4.56***	3.2***	2.4***	2.43***
M5. Disability pension	3.25***	2.7***	2.4***	2.38***	1.39**	1.03	0.97	0.96	3.83***	2.67***	2.33***	2.32***	2.5***	1.9***	1.73***	1.68***
M6. Not much or no data	1.59***	1.6***	1.75***	1.92***	1.25*	1.15	1.2*	1.21***	1.72***	1.53***	1.74***	1.75***	1.55**	1.44*	1.52**	1.57***
M7. Deceased	2.6***	2.57***	0.98	1.02	1.32	1.21	0.8	0.9	3.81***	3.43***	1.26	1.39**	2.12**	1.97**	0.97	1.07
Father (ref. Employment/entrepreneurship)	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
F2. Unemployed in the early 1990s	2.59***	2.42***	1.87***	1.9***	1.5***	1.35***	1.28**	1.13**	3.36***	2.93***	2.21***	1.93***	2.89***	2.63***	2.27***	2*
F3. Unemployed starting in 2009	1.84***	1.65***	1.33***	1.35*	1.33***	1.15	1.07	1.03	1.81***	1.49***	1.18	1.13	2.11***	1.85***	1.62***	1.54***
F4. Unemployed after early 1990s	7.21***	6.12***	3.38***	2.96***	2.22***	1.72***	1.43**	1.05	10.43***	7.64***	4.09***	3.03***	7.51***	5.98***	4.03***	2.99***
F5. Early disability pension	5.04***	4.18***	3.58***	3.66***	1.73***	1.29*	1.18	1.14*	5.42***	3.8***	3.08***	3.07***	3.59***	2.76***	2.4***	2.47***
F6. Late disability pension	2.64***	2.26***	1.93***	1.81***	1.31**	1.04	0.93	0.78*	3.37***	2.52***	2.08***	1.75***	3***	2.44***	2.13***	1.78***
F7. Retirement pension after work	1.11	1.12	1.25**	1.24***	0.89*	0.87	0.88	0.91	0.87	0.86	0.99	1.03	1.05	1.03	1.1	1.15
F8. Not much or no data	3.92***	3.65***	2.32***	2.23***	1.45***	1.25*	1.17	0.92	3.93***	3.29***	2.11***	1.69***	3.48***	3.06***	2.34***	1.91***
F9. Early death	5.72***	5.27***	1.78***	1.67***	1.85***	1.6***	1.04	0.92	6.95***	5.75***	1.82***	1.6***	3.71***	3.26***	1.49*	1.32***
F10. Late death	3.76***	3.58***	1.22*	1.2	1.41**	1.29*	0.88	0.86	4.19***	3.74***	1.22	1.19	3.21***	2.97***	1.42*	1.4*
F11. No registered father	4.46***	3.31***	1.14	1.01	2.13***	1.29*	0.85	0.84	4.63***	2.62***	0.86	0.86	2.52***	1.6	0.74	0.72
Parents' education (ref. Higher academic level)		ref.	ref.	ref.		ref.	ref.	ref.		ref.	ref.	ref.		ref.	ref.	ref.
Comprehensive level		3.58***	3.37***	3.32***		9.21***	9.59***	4.27***		16.13***	15.73***	7.46***		6.95***	7.09***	3.37
Upper secondary level		3.54***	3.47***	3.42***		5.46***	5.69***	3.01***		11.39***	11.37***	6.27***		4.39***	4.55***	2.54***
Lower academic level		2.05***	2.04***	1.99***		2.39***	2.46***	1.76*		3.34***	3.38***	2.5***		2.04***	2.1***	1.52***
Parents' marriage (ref. Parents married)			ref.	ref.			ref.	ref.			ref.	ref.			ref.	ref.
Parents not married			5.42***	5.36***			1.76***	1.4***			6.58***	5.23***			3.01***	ref.
Sex (ref. Female)			ref.	ref.			ref.	ref.			ref.	ref.			ref.	
Male			0.78***	0.77			3.63***	1.79***			2.23***	1.16			2.65***	1.39**
Residence abroad at any time (ref. Never abroad)			ref.	ref.			ref.	ref.			ref.	ref.			ref.	ref.
Abroad at some point			0.97	0.98			0.68***	0.92			0.36***	0.46***			0.92	1.21
Residence history not available			1.25	1.21			1.48	2.2***			1.43	1.98***			3.59***	4.62***
Average grade at the end of comp. school				0.91				0.22***				0.25***				0.26***

* p < 0.05, ** p < 0.01, *** p < 0.001. [1] = pooled results due to the multiple imputations.

Table 5. Regression results for models 1–4 continued. Relative risk ratios

	C6. Early child care				C7. Late child care				C8. Frag. unemp. / emp				C9. Frag. unemp. / social ass.			
	1	2	3	4[1]	1	2	3	4[1]	1	2	3	4[1]	1	2	3	4[1]
Mother (ref. Employment/entrepreneurship)	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
M2. Outside of labour market in the 1990s	2.40***	2.21***	2.16***	1.9***	1.71***	1.61***	1.58***	1.5***	2.23***	2.03***	1.93***	1.72***	2.12***	1.94***	1.85***	1.66***
M3. Long gap in labour market activity	7.60***	6.03***	7.07***	6.44***	3.07***	2.6***	2.93***	2.97***	8.44***	6.67***	6.91***	6.34***	6.95***	5.48***	5.67***	5.07***
M4. Unemployed throughout	12.29***	7.45***	6.28***	5.59***	4.72***	3.22***	2.82***	2.82***	13.96***	8.52***	5.77***	5.47***	18.32***	11.05***	7.7***	6.94***
M5. Disability pension	3.53***	2.40***	2.16***	2.11***	2.64***	1.97***	1.84***	1.84***	4.91***	3.35***	2.93***	2.78***	6.03***	4.08***	3.6***	3.46***
M6. Not much or no data	3.37***	2.88***	3.39***	3.15***	1.46***	1.34*	1.55**	1.62***	3.77***	3.25*	3.68***	3.56***	4.82***	4.09***	4.45***	4.14***
M7. Deceased	2.80***	2.42***	1.03	1.13	1.81***	1.66*	0.85	0.89	4.19***	3.66*	1.35	1.46*	3.58***	3.07***	1.2	1.38***
Father (ref. Employment/entrepreneurship)	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
F2. Unemployed in the early 1990s	2.82***	2.43***	2.01***	1.69***	2.56***	2.3***	2.01***	1.82*	6.29***	5.43***	4.01***	3.18***	4.95***	4.25***	3.28***	2.56***
F3. Unemployed starting in 2009	1.80***	1.49***	1.24*	1.18	1.59***	1.38**	1.22	1.2	4***	3.28**	2.6***	2.23***	2.74***	2.25***	1.83***	1.59***
F4. Unemployed after early 1990s	9.12***	6.55***	3.88***	2.78***	5.36***	4.19***	2.86***	2.23***	17.9***	12.95***	6.91***	4.63***	15.43***	11.09***	6.28***	4.17***
F5. Early disability pension	4.33***	2.96***	2.80***	2.65***	2.84***	2.14***	2.05***	2.1***	10.59***	7.3***	5.74***	5.31***	6.05***	4.13***	3.36***	3.08***
F6. Late disability pension	3.76***	2.78***	2.58***	2.03***	2.04***	1.63***	1.55**	1.33	3.14***	2.32***	1.88**	1.46	4.07***	3.01***	2.52***	1.9***
F7. Retirement pension after work	0.79	0.76	0.84	0.90	0.70***	0.69**	0.73*	0.75	1.35	1.32	1.54*	1.47***	0.93	0.89	1.01	1.11
F8. Not much or no data	4.83***	3.90***	2.81***	2.08***	2.89***	2.5***	1.98***	1.71***	7.48***	6.15***	3.87***	2.8***	9.91***	8.02***	5.24***	3.58***
F9. Early death	4.58***	3.70***	1.45*	1.29***	3.51***	3.04***	1.5*	1.37***	10.44***	8.51***	2.6***	2.09***	7.71***	6.2***	2.1***	1.68***
F10. Late death	2.77***	2.45***	0.93	0.95	2.85***	2.61***	1.26	1.26	6.52***	5.77***	1.85***	1.64*	4.06***	3.59***	1.27	1.18
F11. No registered father	4.73***	2.49***	1.04	1.01	3.14***	1.94**	1	0.99	7.84***	4.29**	1.37	1.18	8.08***	4.28***	1.49	1.23
Parents' education (ref. Higher academic level)		ref.	ref.	ref.		ref.	ref.	ref.		ref.	ref.	ref.		ref.	ref.	ref.
Comprehensive level		16.89***	14.54***	5.53***		7.56***	6.82***	3.71***		16.42***	16.05***	5.94***		16.84***	16.63***	5.62***
Upper secondary level		8.68***	7.76***	3.46***		4.78***	4.39***	2.7***		10.06***	10.08***	4.33***		8.96***	9.08***	3.77***
Lower academic level		2.54***	2.37***	1.56***		1.9***	1.81***	1.38***		2.8***	2.84***	1.81***		2.55***	2.6***	1.63**
Parents' marriage (ref. Parents married)			ref.	ref.			ref.	ref.			ref.	ref.			ref.	ref.
Parents not married			3.78***	2.78***			2.59***	2.16***			7.38***	5.18***			5.65***	4.01***
Sex (ref. Female)			ref.	ref.			ref.	ref.			ref.	ref.			ref.	ref.
Male			0.01***	0.03***			0.18***	0.13***			3.46***	1.57*			2.78***	1.17
Residence abroad at any time (ref. Never abroad)			ref.	ref.			ref.	ref.			ref.	ref.			ref.	ref.
Abroad at some point			0.47***	0.69***			0.43***	0.55*			0.48**	0.68***			0.87	1.07
Residence history not available			2.47*	2.89***			2.29*	2.62***			0.39	0.77			1.67	2.26***
Average grade at the end of comp. school				0.22***				0.34***				0.21***				0.16***

* p < 0.05, ** p < 0.01, *** p < 0.001. [1] = pooled results due to the multiple imputations.

Table 5. Regression results for models 1–4 continued. Relative risk ratios

	C10. Frag. no data / edu. / empl.				C11. Not much or no data			
	1	**2**	**3**	**4**[1]	**1**	**2**	**3**	**4**[1]
Mother (ref. Employment/entrepreneurship)	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
M2. Outside of labour market in the 1990s	1.46***	1.42***	1.41***	1.36***	1.37***	1.31***	1.28**	1.21**
M3. Long gap in labour market activity	2.75***	2.54***	2.55***	2.61***	3.8***	3.42***	3.39***	3.4***
M4. Unemployed throughout	4.66***	3.82***	3.15***	3.3***	7.83***	5.97***	4.5***	4.3***
M5. Disability pension	2.36***	2.04***	1.9***	1.92*	3.89***	3.17***	2.83***	2.7***
M6. Not much data available	2.26***	2.18***	1.98***	2.09***	10.89***	10.4***	7.45***	7.01***
M7. Deceased	2.31***	2.25***	1.39	1.54***	2.73***	2.62***	1.34	1.56***
Father (ref. Employment/entrepreneurship)								
F2. Unemployed in the early 1990s	2.2***	2.11***	1.93***	1.79**	2.64***	2.48***	2.1***	1.77***
F3. Unemployed starting in 2009	1.24	1.16	1.09	1.06	1.71***	1.56**	1.42*	1.32
F4. Unemployed after early 1990s	4.73***	4.19***	3.28***	2.57**	6.77***	5.74***	4***	2.88***
F5. Early disability pension	2.44***	2.13***	1.97***	2.04***	4.73***	3.91***	3.45***	3.38***
F6. Late disability pension	1.74***	1.56***	1.43**	1.27***	2.18***	1.87***	1.67**	1.36***
F7. Retirement pension after work	1.02	1.01	1.05	1.07	1.29*	1.28	1.42**	1.46**
F8. Not much or no data	2.64***	2.47***	1.9***	1.63***	9.73***	8.87***	4.8***	3.49***
F9. Early death	3.75***	3.54***	2.16***	1.97	4.76***	4.39***	2.2***	1.87***
F10. Late death	2.05***	1.97***	1.24	1.23	2.66***	2.52***	1.26	1.26
F11. No registered father	2.54***	1.97***	1.14	1.05	10.3***	7.33***	2.74***	2.44***
Parents' education (ref. Higher academic level)								
Comprehensive level		2.65***	2.81***	1.67**		3.91***	4.55***	2.02***
Upper secondary level		2.04***	2.21***	1.47**		2.82***	3.42***	1.8**
Lower academic level		1.17**	1.23***	0.97		1.47***	1.62***	1.16***
Parents' marriage (ref. Parents married)								
Parents not married			1.91***	1.67***			2.74***	2.18***
Sex (ref. Female)								
Male			1.83***	1.18***			2.33***	1.14*
Residence abroad at any time (ref. Never abroad)								
Abroad at some point			2.41***	2.87***			5.35***	6.77***
Residence history not available			1.55	1.81***			1.38	1.81**
Average grade at the end of comp. school				0.38				0.21***

* p < 0.05, ** p < 0.01, *** p < 0.001. [1] = pooled results due to the multiple imputations.

Child care trajectories

Next, we have two child care trajectories. In the early child care trajectory (cluster C6), almost all recipients (model 4, male RRR 0.027, 95% CI [0.020, 0.036]) of child care benefits relatively early in their early twenties are female. In cluster C7, child care starts a few years later, so is still relatively early, after a period of education or employment (model 4, male RRR 0.12, 95% CI [0.096, 0.170]). The difference between these two is clear with respect to the connections to the parental trajectories. Long parental labour market inactivities and lower levels of parental education have stronger connections to child care at an early age than to later child care. Higher school grade and residence abroad at any time are strong protective measures in both clusters. Mothers' long labour market inactivities have a much stronger connection to early child care than those of fathers'. Of the mothers' clusters, the long gap (cluster M3) has the strongest connection (RRR 6.44, 95% CI [5.99, 6.91]), stronger than the next cluster M4 with continuous unemployment (RRR 5.59, 95% CI [4.24, 7.38]). The difference between these two mothers' clusters is that in the former, some mothers have been on a very long child care career, which might, in some circumstances, encourage their daughters to bear children early.

Females have a strong majority in the child care clusters. Of the early child care cluster, 99% are female, and in the late child care cluster 87% are female.

When we move from model 3 to 4, we are again seeing that adding the school grade weakens the effect of parental level of education, but at times even increases the effect of parental employment trajectories.

Unemployment trajectories

Parental trajectories that are characterised by long labour market inactivities, have their strongest connections to the children's unemployment trajectories in clusters C8 and C9. The difference between these two outcomes is that in the former (C8) unemployment varies with employment, while in the latter (C9), unemployment varies with social assistance. In these two trajectories, we also have the strongest protective effect of higher average school grade (model 4 RRRs 0.21 and 0.16, with 95% CIs [0.18, 0.23] and [0.14, 0.19], respectively). In cluster C9 the crude relative risk ratios of the longest labour market inactivities of mothers (M4) and fathers (F4) in model 1 are at 18 and 15 respectively, and are gradually brought down to 7 and 4 in model 4 by other variables. Again long parental inactivities have stronger connections than shorter periods of absence.

Associations vary between mothers' two clusters with long inactivities (clusters M3 and M4) and children's two unemployment clusters (C8 and C9). In children's unemployment trajectory C9, with social assistance, the stronger connection is with mothers' continuous inactivity (M4), while the mothers' long gap (M3) has the stronger connection to children's unemployment / employment cluster C8. As can be seen in Figure 2, most mothers in cluster M3 are indeed employed during the children's period of study in 2005–2012, which might offer some assistance to their children, while mothers in cluster M4 are unemployed. Also mothers' no data trajectory (cluster M6) has strong connections to children's unemployment clusters (C8 and C9).

Other strong predictors are parents' level of education and unmarried parents. Male sex is a highly significant predictor in model 3, but loses most of its power in model 4 due to the addition of school achievement.

Adding school achievement in model 4 has the same effect that is present in other trajectory connections above: adding it strongly weakens the effect of a low level of parental education, but has a milder effect on parental employment trajectories. This again indicates that school achievement is a very strong predictor for these children's trajectories, and that it highly correlates with parental education level. Meanwhile, children's school achievement did not have an independent association with parental employment trajectories.

No data

Finally, we have the two clusters with various degrees of missing data (C10 and C11). Unlike in surveys, missing data can be meaningfully interpreted in many ways. We know that the children in cluster C11 are not in employment and not on unemployment benefit or any other benefit or pension. There might be some who are so well off that they do not need employment or any benefits, and some might be seriously excluded from society. Most likely, this group (cluster C11) is a mix of many things. We do see a statistically

significant connection to most parental labour market trajectories with data and also to the trajectory of parents belonging to a no-data group, most likely due to childhood spent with family abroad. In C11, with most data missing, the share of expatriates is at 15% in 2005 and rises to 20% in 2012; while in C10 the proportion rises from less than 1% to 5% in 2005–2012. Some 14% of young adults in C11 have resided abroad the entire study time, leaving their mark in some other country's registers, but not in Finnish ones.

Here for the first time in the regression models, in the no-data groups, residence abroad is a strong predictor (C10 and C11 in model 4 RRRs 2.87 and 6.77, 95% CIs [2.07, 3.96] and [5.02, 9.14], respectively). The connection to not having a registered father is at its strongest in C11, with a possible link to unknown family connections abroad.

The children's cluster C10, where missing data interchanges with education or employment, the connection to parental trajectories with long labour market inactivities is relatively weak, but still highly significant in most cases. What is apparent is that there might be some other factors, unknown life events or circumstances, not present in the model that push children to this trajectory. This is also indicated by school achievement, which is a slightly less protective factor than in most of the other children's trajectories. This group is in part a mix of children on either education or employment trajectories, who cluster together due to gaps in their sequences, which could result from a wide variety of parental backgrounds, thus weakening the statistical connections. When taking a closer look, we see that some of the least common trajectories are clustered here, for instance, disability pensions.

Discussion

In this article, our first objective was to identify the most significant early adulthood trajectories in the Finnish Birth Cohort 1987, and the second objective was to identify their parents' most common employment trajectories during the lives of the birth cohort. We studied the early adulthood trajectories of children and their parents' employment trajectories using sequence analysis and cluster analysis. We identified 12 children's trajectories, or clusters, and 10 fathers' and 7 mothers' trajectories. More than two thirds of mothers (68%) and fathers (72%) have continuously been in employment, and we saw that the same proportion of children is on a stable education or employment trajectory (73%). We are left with between a quarter and one-third of the population who are not on these trajectories.

In addition to the most common employment or education trajectories, we found that children were on less continuous or more disrupted education or employment trajectories, on early child care trajectories, unemployment trajectories or on a trajectory for which we do not have much data. Most NEET-type and missing data trajectories were very persistent during the study years. We did not find common unemployment trajectories in which the situation for the young adults improved over the eight years of study.

Some mothers have been outside of the labour market for a long while. The most common gap in labour market activity took place during the 1990s recession and in its aftermath. Following a gap, some 5% of mothers returned to the labour market only during the late 2000s, with some 3% never entirely returning. The rest of the mothers were on disability pension, the no-data group, or deceased.

Likewise, fathers had long absences from the labour market. Some 7% of fathers suffered in the 1990s recession, 3% never entirely returning to the labour market. The latest recession of 2008 took a toll on 3% of fathers, who mostly did not return to employment during this follow-up. More fathers than mothers were on disability pension, on old age pension or had died, so we saw fit to split groups based on the timing of pension or death, early and late. Similar to children and mothers, some of the fathers were in a no-data group.

Our third objective was to see how much children's early trajectories are affected by their parents' employment trajectories, and we used multinomial logistic regression to find connections between them. Generally, the results show that parents' long absences from the labour market have strong links to children's early child care, NEET-type and missing data paths.

Parents' labour market trajectories did not have many significant connections to children being on an employment trajectory rather than further education. Here other factors were at play: sex, parental education and comprehensive school achievement. On the other hand, parental trajectories characterised by long absences from

the labour market were strongly linked to children's early employment trajectories that begin with a study grant. Thus, in circumstances of low family income the benefit system assists children in moving forward. Females are more strongly represented in clusters that receive early upper-secondary-level study grants, which could indicate that males are less prepared to use the benefit system.

When looking at the children's child care trajectories, family background distinguished those with very early and early child care. Those who were on the earlier child care trajectory had stronger links to childhood disadvantages than those on the later child care trajectory. Mothers' long labour market absences had strong connections to these trajectories, which were almost entirely female trajectories, while fathers' long absences from work had slightly weaker but mostly significant connections. Parents' low education and children's low comprehensive school achievement had strong connections to both trajectories. The strong link to mothers' late return to the labour market could indicate a connection between mother's long child care career and children's early child care.

As our investigation focused on transitions during early adulthood, it is not surprising that we did not identify major continuous unemployment trajectories for children. Instead we found two clusters in which unemployment interchanges with employment or with social assistance. In these two trajectories, we have the strongest connections to parents' long labour market absences. These clusters have a male majority, but we saw stronger connections to mothers' employment trajectories than to those of fathers. In Finland, mothers are more often lone parents than fathers (Official Statistics Finland: Families, 2015), but joint custodies are not registered.

Since we sequenced an entire birth cohort, we also had thousands of cases with not much or no data. In the Finnish context, with its strong reliance on registers, no data equals no employment, no entrepreneurship and no benefits, which might include people with other kinds of support, possibly support from family or a spouse, or at the other end, in a situation of serious exclusion from society. Given the presence of these groups, we included residence data in the regression model to see if there is a connection to residence abroad, which would explain the lack of register data. But while residence is a strong predictor, there are other strong factors. In the first child no-data group the connections to family background are relatively weak (and some statistically insignificant), resulting possibly from the group mix, which includes children from a wide variety of backgrounds. In the latter group, where data are almost entirely or entirely missing, we saw strong connections to parents' long labour market absences and also to the mothers' no-data group. Future research should try to identify the background selection processes leading to these cluster memberships. They may have something to do with somatic or mental health, substance abuse, or phenomena related to institutional arrangements.

Throughout the study, we saw that children's average comprehensive school grade is a strong predictor of their early adulthood trajectory. If grades are too low some education doors will be closed, which will in most cases result in an employment trajectory over further education. Very low grades are connected to disadvantages during early adulthood. We found that grades are strongly connected to parents' educational level, but the connections between parents' labour market activity and children's school grades are weak. Earlier evidence (Schoon, 2014) shows that parental unemployment can raise children's educational aspirations, as they might see value in education to avoid their parents' plight. On the other hand, parental unemployment trajectories could be discouraging. If parental worklessness has countervailing effects, we would expect a weakening of its relationship to child outcomes. Most children's trajectories vary by sex, which relates to school achievement difference between males and females. Could it be that males need more interventions during hardship than females during their years in school?

In Norway, fathers experiencing job loss had a stronger effect on children than mothers, possibly due to traditional role models at home (Rege, Telle & Votruba, 2007). This was not true in the current Finnish study, where even though the subjects are nearly the same age as those in Rege et al.'s study (2007), mothers' trajectories had a stronger connection to disadvantageous male-dominated trajectories than those of fathers, and equally strong connection elsewhere. Female labour market participation in Finland has been at a high

level throughout the lives of this cohort, so traditional roles may have weakened.

Overall parents' employment trajectories had very strong connections to children's early adulthood trajectories, at times stronger than parents' education, although the impact of parents' education is most likely funnelled through children's comprehensive school achievement. Especially parents' long absences from the labour market showed up as strong predictors of the more NEET-typed and missing data trajectories of children.

By using sequence analysis and 52 data points with sampling, not cross-sectioning, from the entire length of the birth cohort's life, we obtained a good picture of parents' long periods of labour market exclusion. The excellent quality data allowed us to draw a picture of highly diverse early adult pathways and to identify the significant role that parental employment trajectories have upon them.

Acknowledgements

An earlier version of this paper was presented at The 8th Nordic Working Life Conference in Tampere, Finland, in October 2016. This study was supported by grants from the Academy of Finland strategic fund, The Finnish Work Environment Fund, Alli Paasikivi Foundation, Palkansaajasäätiö, and European Social Fund in project "Six City Strategy". The Finnish Birth Cohort 1987 study obtained a positive statement from the research ethical committee of the National Institute for Health and Welfare (Ethical committee §28/2009), and permissions to use the register data was obtained from all register holding organisations.

References

Bäckman O. & Nilsson A. (2011). Pathways to Social Exclusion – A Life-Course Study. European *Sociological Review, 27*(1), 107–123. https://doi.org/10.1093/esr/jcp064

Barone, C. & Schizzerotto, A. (2011). INTRODUCTION: Career mobility, education, and intergenerational reproduction in five European Societies. *European Societies*, *13*(3), 331-345. https://doi.org/10.1080/14616696.2011.568248

Brzinsky-Fay, C. (2007). Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review*, *23*(4), 409-422. https://doi.org/10.1093/esr/jcm011

Brzinsky-Fay, C. (2011). *School-to-work Transitions in International Comparison* (Doctoral dissertation). Tampere, Finland: Tampere University Press.

Brzinsky-Fay, C. (2015). Gendered School-to-Work Transitions? A Sequence Approach to How Women and Men Enter the Labor market in Europe. In Blossfeld, H-P., Skopek, J., Triventi, M. & Buchholz, S. (Eds.), *Gender, Education and Employment – An Intergenerational Comparison of School-to-Work Transitions* (pp. 39–61). Cheltenham, United Kingdom: Edward Elgar Publishing. https://doi.org/10.4337/9781784715038.00010

Brzinsky-Fay, C. & Solga, H. (2016). Compressed, postponed, or disadvantaged? School-to-work-transition patterns and early occupational attainment in West Germany. *Research In Social Stratification And Mobility, 46*, Part A, 21–36. https://doi.org/10.1016/j.rssm.2016.01.004

Bukodi, E. & Goldthorpe, J. H. (2011). Social class returns to higher education: chances of access to the professional and managerial salariat for men in three British birth cohorts. *Longitudinal and Life Course Studies, 2*(2), 185–201.

Caspi, A., Wright, B. R. E, Mofitt, T. E. & Silva, P. A. (1998). Early Failure in the Labor Market: Childhood and Adolescent Predictors of Unemployment in the Transition to Adulthood. *American Sociological Review, 63*(3), 424–451. https://doi.org/10.2307/2657557

Cornwell, B. (2015). *Social Sequence Analysis – Methods and Applications*. Campridge, United Kingdom: Campridge University Press. https://doi.org/10.1017/CBO9781316212530

Elder, G. H. Jr., Johnson, M. K. & Crosnoe, R. (2003). The Emergence and Development of Life Course Theory. In Mortimer, J. T., Shanahan, M. J. (Eds.), *Handbook of the Life Course* (pp. 3–19). New York: Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-0-306-48247-2_1

Gabadinho, A., Ritschard, G., Müller, N. S. & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1–37. https://doi.org/10.18637/jss.v040.i04

Gangl M. (2002). Changing labour markets and early career outcomes: labour market entry in Europe over the past decade. *Work, employment and society. 16*(1), 67–90. https://doi.org/10.1177/09500170222119254

Gray, M. & Baxter, J. (2012). Family joblessness and child wellbeing in Australia. In Kalil, A., Haskins, R. & Chesters, J. (Eds.), *Investing in children: work, education, and social policy in two rich countries*. Washington, DC: Brookings Institution.

Hollister, M. (2009). Is Optimal Matching Suboptimal? *Sociological Methods & Research, 38*(2), 235–264. https://doi.org/10.1177/0049124109346164

Härkönen, J. & Bihagen, E. (2011). Occupational Attainment and Career Progression in Sweden. *European Societies, 13*(3), 451–479. https://doi.org/10.1080/14616696.2011.568261

Ilmakunnas, I, Kauppinen, T. M. & Kestilä, L. (2015). Sosioekonomisten syrjäytymisriskien kasautuminen vuonna 1977 syntyneillä nuorilla aikuisilla. *Yhteiskuntapolitiikka, 80*(3), 247–262.

Kangas, O. & Saloniemi, A. (2013). *Historical making, present and future challenges for the Nordic welfare state model in Finland*. NordMod 2030, Sub-report 6.

Larja, L., Törmäkangas, T., Merikukka, M., Ristikari, T., Gissler, M. & Paananen, R. (2016). NEET-indikaattori kuvaa nuorten syrjäytymistä. *Tieto & Trendi, 2/2016*.

Mastekaasa, A. (2011). Social Origins and Labour Market Success – Stability and Change over Norwegian Birth Cohorts 1950–1969. *European Sociological Review, 27*(1), 1–15. https://doi.org/10.1093/esr/jcp050

McVicar, D. & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A, 165*(2), 317–334. https://doi.org/10.1111/1467-985X.00641

Mortimer J. T., Zhang L., Husseman J. & Wu C (2014). Parental economic hardship and children's achievement orientations. *Longitudinal and Life Course Studies, 5*(2), 105–128. https://doi.org/10.14301/llcs.v5i2.271

Mroz T. A. & Savage T. H. (2006). The Long-Term Effects of Youth Unemployment. *The Journal of Human Resources, 41*(2), 259–293. https://doi.org/10.3368/jhr.XLI.2.259

Müllner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software, 53*(9), 1–18. https://doi.org/10.18637/jss.v053.i09

Official Statistics Finland: Families (2015). Helsinki: Statistics Finland. Accessed 22th June 2017 from http://www.stat.fi/til/perh/index_en.html

Official Statistics Finland: Labour Force Survey (2015). Helsinki: Statistics Finland. Accessed 22th June 2017 from http://www.stat.fi/til/tyti/index_en.html

Paananen, R. & Gissler, M. (2012). Cohort profile: the 1987 Finnish Birth Cohort. *International Journal of Epidemiology, 41*(4), 941–945. https://doi.org/10.1093/ije/dyr035

R Core Team (2015). R: A language and environment for statistical computing (Version 3.2.3) [Software]. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/

Rege, M., Telle, K. & Votruba, M. (2007). *Parental Job Loss and Children's School Performance*. Discussion Papers No. 517, October 2007. Statistics Norway: Research Department.

Ristikari, T., Törmäkangas, L., Lappi, A., Haapakorva, P., Kiilakoski, T., Merikukka, M., Hautakoski, A., Pekkarinen, E. & Gissler, M. (2016). Suomi nuorten kasvuympäristönä – 25 vuoden seuranta vuonna 1987 Suomessa syntyneistä nuorista aikuisista. THL Raportti 9/2016 / Nuorisotutkimusverkosto/Nuorisotutkimusseura, verkkojulkaisuja 101. Tampere, Finland: Juvenes Print.

Sackmann, R. & Wingens, M. (2003). From Transitions to Trajectories – Sequence Types. In Heinz, W. R. & Marshall, V. W. (Eds.), Social Dynamics of the Life Course – Transitions, Institutions and Interrelations. New York, US: Aldine De Gruyter.

Schoon I. (2014). Parental worklessness and the experience of NEET among their offspring. Evidence from the Longitudinal Study of Young People in England (LSYPE). *Longitudinal and Life Course Studies, 5*(2), 129–150. https://doi.org/10.14301/llcs.v5i2.279

Schoon, I. & Lyons-Amos, M. (2016). Diverse pathways in becoming an adult: The role of structure, agency and context. *Research in Social Stratification and Mobility, 46*, Part A, 11–20. https://doi.org/10.1016/j.rssm.2016.02.008

Sirniö, O., Kauppinen, T. M. & Marttinen, P. (2016). Income trajectories after graduation: an intergenerational approach. *Advances in Life Course Research, 30*, 72–83. https://doi.org/10.1016/j.alcr.2016.04.001

Steijn, B., Need, A. & Gesthuizen M. (2006). Well begun, half done? Long-term effects of labour market entry in the Netherlands, 1950–2000. *Work, employment and society. 20*(3), 453–472. https://doi.org/10.1177/0950017006066996

Studer, M. (2013). WeightedCluster Library Manual – A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES working papers, 24.*

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03

Witting, M. & Keski-Petäjä, M. (2016). Vanhempien koulutus vaikuttaa lasten valintoihin. *Tieto & Trendi, 2/2016.*

Zwysen, W. (2014). A disadvantaged childhood matters more if local unemployment is high. *SOEPpapers on Multidisciplinary Panel Data Research*, *725.* https://doi.org/10.2139/ssrn.2547430

Appendix

For the sequence analysis, we used TraMineR (Gabadinho, Ritschard, Müller & Studer, 2011) on 64-bit R 3.2.3 (R Core Team, 2015) on a Windows computer. Distance and clustering methods were chosen due to the size of our data. For the distance method, we opted for hamming distances, which is, in short, optimal matching without insertion-deletion costs. This saved us several hours of computer time per run, but did not affect the cluster quality much if at all (Hollister, 2009). What insertion-deletion essentially does is allow the distance calculation to shift a sequence around on the time axle to match it with other sequences. By how much depends on the ratio of the insertion/deletion and substitution costs, where a lower ratio allows for a longer time shift. Too low an insertion/deletion cost interferes with substitution cost, however, since with a ratio below 0.5, 'deletion+insertion' is a cheaper operation than a substitution. However, we are not concerned with the lack of insertions and deletions, since we consider the timing of the transitions as important.

For the clustering, we used the package 'fastcluster' (Müllner, 2013) in R. According to the package's author and our tests, it significantly speeds up hierarchical clustering with the Ward method, but offers the same results as other packages. Hierarchical clustering is nevertheless not always optimal (Studer, 2013), so we experimented with non-hierarchical partitioning around medoids (PAM) and also with a combination of these two, where the hierarchical clustering is given as an initial partitioning for the PAM method. For children's data, we found hierarchical clustering to be sufficient, while with PAM we discovered more interesting unemployment clusters for their parents.

With this data we discovered an internal limit to the number of individual sequences that the distance methods in TraMineR can handle (46,431 at the most), possibly due to internal limits in the size of R's 32-bit vector. Depending on the sequence complexity, one can try to reduce the number of identical sequences and go below the limit by means of the wcAggregateCases function in the R package WeightedCluster (Studer, 2013). We were able to process the data of both parents by aggregating the sequences (matching identical sequences and calculating distances only once per match), but in the children's more complex data, only 2,381 identical sequences were found from the original 59,476. As no other software packages for sequence analysis were available to us, we had to split the data and run two sequence analyses.

The children's distance and cluster analysis was split due to the split sequence data. We created a surplus of clusters from both parts, where mainly the largest resulting clusters formed several initial small clusters based on the time entering further education or employment. After forming the initial clusters, we joined the two data together. Since our objective was to find the most significant early adulthood trajectories with a focus on disadvantaged trajectories, we joined all the education and employment trajectory groups respectively by visual inspection (based on status distribution and timing), keeping early study benefit (which links to low income in the family) trajectories separate. This formed two large education and two large employment groups. All the other clusters were matched and joined to other clusters until a visual match could not be found. This resulted in our final cluster solution. Both parents' analyses were performed using the same principle but without the split: clusters were joined together until no match was found. Creating a surplus of clusters allowed us to find more interesting groups than was possible with fewer clusters.

The number of clusters has been debated at length and indicators for their quality also exist (Studer, 2013). We find that the quality indicators do not necessarily measure the usefulness of the clusters. We might, for example, discover clusters with both very low and very high within-sequence variances. A steady career, with low variance, is just as important to identify as a highly fragmented career, with high variance. The data also matter, since as the number of available labour market statuses and the length of the sequence grow, the variance necessarily grows as well. Large amounts of data and a high number of statuses might require more clusters. Usefulness depends on the research questions – here we want to identify the most significant trajectories.

Psychiatric diagnoses as grounds for disability pension among former child welfare clients

Miia Bask Norwegian Social Research, Norway
miia.bask@nova.hioa.no
Tiina Ristikari National Institute for Health and Welfare, Finland
Ari Hautakoski National Institute for Health and Welfare, Finland
Mika Gissler National Institute for Health and Welfare, Finland; University of Turku, Finland, and Karolinska Institute, Sweden

(Received March 2017 Revised May 2017) http://dx.doi.org/10.14301/llcs.v8i4.459

Abstract

This study used the 1987 Finnish Birth Cohort, which included all children born in Finland in 1987 (N=59,476), to investigate psychiatric diagnoses as grounds for disability pensions (DPs) among child welfare clients and explored the background factors associated with such diagnoses. Descriptive statistics show that DP is substantially more common among child welfare clients than among other children.

Logistic regressions revealed that the factors most strongly related to psychiatric diagnoses among girls were mother's somatic DP, child protection history, and parental social assistance. Psychiatric diagnoses among boys were most strongly related to mother's psychiatric DP, child protection history, and parental divorce.

The factors related to DP among girls included child protection history, father's psychiatric DP, father's somatic DP, and parental social assistance. DP among boys was related to child protection history, mother's psychiatric DP, parental social assistance, father's somatic DP, and father's psychiatric care in specialised hospitals.

A child welfare history that includes out-of-home care indicates that there were severe problems in the home environment during upbringing. Detailed investigations should therefore be undertaken, such as examining the role of mediating and moderating factors, including the ability of social and educational services to ameliorate the effects of challenging childhood conditions.

Keywords

Child welfare, disability pension, Finland, out-of-home care, psychiatric diagnoses

Introduction

States with generous welfare systems have become increasingly concerned about the future financing of pension systems. Because of aging populations in many industrialised countries, the working population must remain in the workforce longer to keep the welfare and pension systems afloat. It is therefore important to both the research community and decision-makers to identify the causes of the disbursement of disability pensions (DPs) to younger recipients. Thus, there is a need for more detailed knowledge regarding the factors underlying DP decisions.

More importantly, young recipients of DP are at risk for economic hardship later in their lives. This is because they contribute to (and later benefit from) the earnings-based occupational pension insurance.

If their work career is short, payments in the future will be small. There is substantial evidence regarding child welfare clients' experience of intergenerational transmission of inequality and their disadvantaged positions in adulthood. Previous research has shown that individuals with a childhood history in social services are more disadvantaged than their counterparts in many areas of life, including labour market integration (Vinnerljung, Brännström, & Hjern, 2015). However, to our knowledge, no previous research has investigated the specific diagnoses that underlie DP decisions for individuals who have childhood histories with social services. Therefore, our objective is to investigate psychiatric diagnoses as grounds for DP among child welfare clients, exploring whether child welfare clients are more prone to DP on specific psychiatric grounds than their counterparts, and to scrutinize the background factors related to psychiatric diagnoses and DP.

Emergency out-of-home care placements in Finland increased during the early 2000s but began to decrease between 2013 and 2014. Altogether, 10,675 children were placed in care during 2014, which represents 1% of the population of this age group (0-17 years). In total, 17,958 (1.1%) children and youth were placed outside of their homes in 2014. Of these placements, 53% were male. More than half of these children were placed with foster families; of these, 13% were placed with their relatives (THL, 2014a, 2014b).

Short-term absence from the labour market due to health reasons is typically covered by sickness allowance in Finland. However, individuals with permanently reduced work capacities are entitled to DP. If the disability lasts less than one year, a sickness allowance is paid by the Social Insurance Institution. If the illness, injury or handicap reduces work capacity for a year or more, an individual is compensated by cash rehabilitation benefits or DP (for a more detailed description, see ETK, 2017). During a fixed-term DP, an individual may be offered rehabilitation or an opportunity to change occupation if the pension provider considers that there are possibilities to return to work. The process leading to permanent DP typically involves thorough medical examinations to evaluate the capacity for work. Even if the majority of the DPs granted to the cohort in this paper are fixed-term, individuals with a fixed-term DP tend to depart from the work force permanently.

In 2014, there were 232,475 individuals on DP in Finland. Of these, 1,615 were under 20 years of age, and 43,129 were between 20 and 44 years old. The most common grounds for DP are mental health and behavioural diagnoses. This group of diagnoses includes mental and behavioural disorders due to psychoactive substance abuse; schizophrenia, schizotypal and delusional disorders; mood affective disorders; neurotic, stress-related and somatoform disorders; and disorders of psychological development and intellectual disability (ETK, 2015, p. 137). These diagnoses are particularly common among younger DP recipients; 32,779 individuals between 16 and 44 years old are on DP due to mental health and behavioural diagnoses (ETK, 2015).

The share of new retirees granted DP in Finland has decreased. However, the share of mental health problems as grounds for DP is substantial. A total of 20,987 individuals were granted DP in 2014; within this group, 6,757 individuals were granted DP due to mental health and behavioural problems. Moreover, 3,575 of these were between 16 and 44 years old (ETK, 2015, p. 114).

Because mental health and behavioural issues are the main reasons for DP among younger DP recipients in Finland, these diagnoses are of considerable interest. Moreover, there is substantial political concern regarding the working abilities of the working-age population because aging and high unemployment are already challenging the sustainability of the Nordic welfare state model in Finland. This threatens to place larger numbers of individuals in conditions of economic hardship when they reach old age retirement. We hope that this research can shed light on the challenges child welfare clients experience in their transition to the labor market and increase interest in developing measures to support these children in this transition.

Previous research

Intergenerational transmission of inequality

In their review, Ben-Shlomo and Kuh (2002) discuss several factors that affect an individual's health over the life course. The mechanisms behind the intergenerational transmission of inequality include parents passing on economic, human and cultural capital. Two classes of models of adult

health appear in the life course literature. The first class is called critical period models. These models emphasise the timing of an experience, meaning that exposure to a certain experience during a particular period in an individual's development can have long-term consequences on the physiological functions of the individual and may lead to ill health. Thus, poverty may be particularly harmful to children during important life course transitions such as the beginning of school.

The other class of models concentrates on accumulated risk factors and experiences. These models emphasise the accumulation of effects over the individual life course. In some cases, the main element is the number of risk factors, whereas in other cases, it is the duration of the risk-experience that matters (Ben-Shlomo & Kuh, 2002; Lynch & Smith, 2005).

The processes that lead a disadvantaged youth to DP can be described as cumulative disadvantage. The cumulative advantage model proposed by Crystal and Shea (1990) is a popular model in life-course research that has achieved widespread acceptance in the literature. In attempting to explain inequality in society, the cumulative advantage model focuses on how inequality can be exaggerated over the life course because individuals accumulate different amounts of advantages and disadvantages over time: "Those who are initially advantaged […] are more likely to receive a good education, leading to good jobs, leading to better health and better pension coverage, leading to higher savings and better postretirement benefit income" (p. 437). Correspondingly, those who are disadvantaged from the start are less likely to receive these types of positive reinforcement, resulting in increased intra-cohort inequality over the life course (Bask & Bask, 2015). This model also portrays the phenomena discussed in this paper. A history with child welfare services is itself an indicator of negative life experiences early in the life course. DP is an additional indicator of a disadvantaged position that involves worse health and worse pension coverage. Furthermore, we consider the accumulated risk factors that the parents of the children in our study possess to study intergenerational transmission of disadvantage.

A substantial body of literature demonstrates that poor socioeconomic conditions during the early life course affect adult health. In her review of the literature, Reiss (2013) shows that children from disadvantaged socioeconomic backgrounds are clearly more prone to mental health problems. The review further shows that the persistence of low socioeconomic status (SES), usually measured as a combination of parental educational, economic and occupational status, is related to higher rates of mental health problems. Studies show that accumulated risk factors have more severe consequences than exposure to a single risk factor for the development process over the early life course, and the dangerousness of the effects increases with the sum of the risk factors. Therefore, even when an individual shows extraordinary resilience in many cases, exposure to multiple risks has a permanent effect on the individual (Evans, Li, & Whipple, 2013; Franzén, Vinnerljung, & Hjern, 2008; Lynch & Smith, 2005).

Evans and Cassells (2014) show that children who experience poverty in their early childhood follow a developmental trajectory that involves worse behavioural adjustment. Cumulative risk experience is found to be an explanatory factor. More specifically, these authors link early childhood poverty to behavioural adjustment problems in early adulthood by showing that longer periods of poverty at age nine correlate positively with worse mental health at age 17. Moreover, there is recent evidence that early childhood poverty has a negative effect on mental wellbeing in adulthood and that cumulative risk experience acts as an explanatory mechanism linking childhood poverty and young adulthood mental health problems (see also Costello, Erkanli, Copeland, & Angold, 2010; Najman et al., 2010; Paananen, Ristikari, Merikukka, & Gissler, 2013).

Multiple studies have shown that individuals' socioeconomic background affects adult health and wellbeing outcomes (Ristikari, Hakovirta, & Gissler, 2016). SES also influences physical and psychosocial living conditions. Parents with fewer resources can often afford only disadvantageous living conditions, including lower-quality schools and more dangerous and segregated neighborhoods for their children (Cohen, Janicki-Deverts, Chen, & Matthews, 2010). In addition to SES, another background factor that has been shown to be important is family type. After controlling for several relevant explanatory factors, a Swedish study showed that children of single mothers are clearly more likely to enter out-of-home care (Franzén et al., 2008).

Parents who are struggling to make ends meet have fewer resources – both economic and non-economic – to offer their children. Parental SES influences parents' expectations for their children, and these expectations, in turn, influence children's outcomes (Bask, Ferrer-Wreder, Salmela-Aro, & Bergman, 2014). Thus, values and expectations are transmitted from parents to their children. Some studies indicate that there is a certain social inheritance in welfare program participation. The probability of an individual becoming dependent on welfare programs is higher for those whose parents were also welfare recipients. A Norwegian study shows that the norms and values related to DP are passed on from one generation to the next and that the probability of receiving DP is partly dependent on parental behaviour regarding DP (Bratberg, Nilsen, & Vaage, 2015; see also Dahl, Kostøl, & Mogstad, 2014).

The remainder of this paper is organised as follows. In the next section, we present previous research on child welfare clients and their mental health and on psychiatric diagnoses as grounds for DP. We then present our aims, materials and methods, and results before concluding the paper with a discussion.

Child welfare clients and health

There is substantial evidence that child welfare clients are disadvantaged in many areas of life. Child welfare clients have been shown to have higher rates of illness, particularly mental illness, but they also tend to be characterised by lower educational attainment (Berlin, Vinnerljung, & Hjern, 2011; Jackson & Cameron, 2012; Kestilä, Väisänen, Paananen, Heino, & Gissler, 2012). They also have a higher likelihood of becoming involved in criminality; this is particularly true among children with a history of out-of-home care (Doyle, 2007; Mersky & Janczewski, 2013). Child welfare clients are also overrepresented in statistics involving suicide (Farand, Chagnon, Renaud, & Rivard, 2004; Vinnerljung, Hjern, & Lindblad, 2006).

In their review of the relationship between SES and child health, Chen, Matthews, and Boyce (2002) present several potential mechanisms that link parental SES to children's health. In addition to prenatal factors, they present studies that reveal emotional/cognitive, social, environmental, behavioral and biological mechanisms that link parental SES with their children's health. For example, children from lower SES families have increased risk of injury, more severe and higher prevalence of asthma, and increased risk of high blood pressure. These children are also more likely to become smokers and to exercise less than their wealthier counterparts.

As discussed above, individuals with low SES have worse health than their wealthier counterparts. Low SES is also related to behaviours that are known to be health risks. For instance, smoking, excessive alcohol consumption and lower levels of physical activity vary based on SES. Although interest in the grounds and processes for entry into DP among the younger population has recently increased, there is scant research investigating the risk factors for young DP recipients because most disability studies address the older population (Bowen & González, 2010).

Therefore, low parental SES background can be seen as a risk factor for health problems in the younger population. As previously discussed, studies of the determinants of DP are mainly focused on adulthood predictors of DP (Harkonmäki et al., 2007). However, children with a history of social services suffer from mental health problems more often than their counterparts do (Heneghan et al., 2013), which leads to the suspicion that child welfare clients are more prone to DP than children without that history and that psychiatric diagnoses may be important grounds for such DP decisions. To our knowledge, there are no studies that address these research questions.

Psychiatric diagnoses as grounds for DP

The number of DP recipients has remained stable in Finland since the final decades of the twentieth century. However, the share of the various grounds for entry into DP has changed. Since the mid-1990s, mental disorders have increased as the primary grounds for DP (Organisation for Economic Co-operation and Development, 2014).

According to the disability process model, disability is an outcome of a long-term process with great variability in illness types and severity as well as health behaviours and personal and environmental factors (Verbrugge & Jette, 1994). The relationship between adverse childhood living conditions and DP has been documented in previous research (Upmark, Lundberg, Sadigh, & Bigert, 2001; Upmark & Thundal, 2002).

In a recent study, Laaksonen, Blomgren, and Tuulio-Henriksson (2016) show that sickness allowances due to mental health problems

predicted DP based on mental health grounds. Negative factors seem to accumulate; for example, bipolar disorder often involves other types of problems, such as excessive alcohol consumption. This accumulation of problems negatively affects the labour market prospects of these individuals. Sickness allowances due to mental disorders also predicted DP in cases of schizophrenia and depression. A Norwegian study showed that anxiety and depression were significant factors that explained subsequent DP. In addition to the separate effects, the combination of depression and anxiety was an even stronger explanatory factor for DP. These factors were more noticeable in the younger population (Mykletun et al., 2006).

Aims

There is a vast body of literature on this topic. We know a great deal about the life courses of child welfare clients, and we know a great deal about psychiatric diagnoses as grounds for DP in the general population. However, no studies have examined specific psychiatric diagnoses as the grounds for DP among former child welfare clients with a history of out-of-home care. Therefore, this study aims to investigate psychiatric diagnoses as grounds for DP among child welfare clients and to explore the background factors that are related to psychiatric diagnoses and DP. Furthermore, we consider the cumulative risk assumption and test whether multiple risk factors increase the likelihood of DP within this cohort. The motivation for this research is to gain greater insight into the adult life course, labour market attachment and psychiatric challenges related to that transition among the very vulnerable group of child welfare clients.

Materials and methods

Study population

The study uses the 1987 Finnish Birth Cohort (Paananen et al., 2013). The data include all children born in Finland in 1987 (N=59,476) and their parents. Children who died before the age of 18 were removed from the analyses in this paper. The children's life courses until the age of 25 were followed using official registers. The study was approved by the Ethical Committee of the National Institute for Health and Welfare (§28/2009) and received appropriate permission to use the confidential register data in scientific research from all register-keeping organisations.

For the purposes of this study, we use data on the members of the cohort who were subjected to child welfare actions at some point during their lives. The data describe their life conditions prior to age 16, including information about the health and education of the parents of the cohort members, information about child welfare actions, and information regarding the diagnostic grounds for the DP received by the cohort member. All register data were combined using the unique personal identification numbers (IDs).

Study variables

Child protection. Data on child welfare actions were obtained from the National Institute for Health and Welfare, Child Protection Register. In our analyses, we use information about whether the cohort member has a history of out-of-home care. In total, 1,891 individuals in this cohort had been subjected to child welfare actions (see more details in Table 1). Child welfare actions include support community care, out-of-home placement, including emergency and involuntary placement, and after-care. The average length of a single placement was 616 days.

Disability pension (DP). Cohort members' DP data were gathered from the Social Insurance Institution of Finland (2003-2012) and from the Pension Register that is maintained by the Finnish Centre for Pensions (2006-2012). The data included fixed-term DPs (891 individuals) and DPs that will continue until further notice (481 individuals). The data also included information about the diagnostic grounds for the DP using the ICD 10. DP data were gathered from age 16 onwards. This cohort included 1,372 individuals with DP, of whom 866 were granted DP based on psychiatric diagnoses (see more details in Table 1). Those with intellectual disability diagnoses (F70-F79) were removed from the regression analyses (Table 2a-2c) but are included in the analysis investigating whether child welfare clients are more prone to specific psychiatric DP grounds than their counterparts (Table 3).

Parents' DP data were also obtained from the Pension Register. Parental DPs were divided into two groups: somatic and psychiatric (F00-F99). Before the children turned 16 years old, 2,115 fathers were on DP based on somatic grounds and 1,173 based on psychiatric grounds. The

corresponding figures for mothers were 969 and 907, respectively.

Parental psychiatric in-/outpatient care. The Finnish Hospital Discharge Register (HDR), which is maintained by the National Institute for Health and Welfare (THL), includes all inpatient care episodes from all Finnish hospitals since 1969 and all specialised-level outpatient visits in public hospitals since 1998. Data on parental psychiatric care were collected from the HDR for psychiatric inpatient care and/or outpatient care between the cohort member's birthdate and the date when the cohort member reached age 16. In total, 3,031 fathers and 3,796 mothers had psychiatric hospital inpatient care episodes before the children turned 16 years old.

Parental social assistance. Recipients of social assistance are registered by the THL. Social assistance refers to last-resort financial assistance provided by social services to a household from municipal funds when other sources of income are insufficient to ensure that the basic needs of a person or a family are met. Parental social assistance was registered for either the biological mother or biological father or both parents during the follow-up 1987-2003 period. In total, 21,234 parents had received social assistance at least once before the children turned 16 years old.

Family characteristics

Mother's age under 20 years at the time of the child's birth. Data on the mother's age at the time of childbirth were obtained from the Medical Birth Register, maintained by the THL. In total, 1,884 mothers were under 20 years of age when the child was born.

Parental education. Data on the highest educational level of cohort members' parents when the cohort member was below 16 years old were obtained from Statistics Finland and classified as 'high school or higher' (12 years or more of education; 10,675 fathers and 9,383 mothers had this educational level), 'lowest level tertiary' (11 years; 8,121 fathers and 13,604 mothers), 'lower secondary' (10-11 years; 25,560 fathers and 26,600 mothers), or 'primary' (up to nine years; 14,540 fathers and 9,291 mothers).

.

Divorce. Data on cohort members' biological parents' divorces (classified in the analyses as divorced vs. not divorced during the follow-up) were obtained from the Finnish Central Population Register. This cohort comprises 13,327 cases in which biological parents were divorced before the child turned 16 years old.

Death of a parent. Information on parents' death during the follow-up was received from the Finnish Central Population Register. In total, 1,755 fathers and 595 mothers died before the child turned 16 years old.

Cumulative risk factors. This is a cumulative risk factor variable (i.e., Parents' social assistance; Mother's psychiatric in-/outpatient care; Father's psychiatric in-/outpatient care; Mother's psychiatric DP; Mother's somatic DP; Father's psychiatric DP; Father's somatic DP; Mother's death; Father's death; Divorce; and Mother younger than 20 years old). We found that 21,832 individuals had no risk factors, 26,768 had 1-2 risk factors, 8,773 had 3-4 risk factors, and 1,505 had 5-9 risk factors.

Analysis

The first part of the analysis involves logistic regressions that seek to determine whether DP is more likely among child welfare clients than among others. Binary logistic regression analyses were used to define the odds ratios (ORs) and 95% confidence intervals (95% CIs). The analyses were performed using SPSS Statistics version 24. The first model involves child welfare experience as the only covariate. The second model includes all the covariates, and the third model differs from the second model by excluding the child welfare experience and reveals the relationship between DP and social background characteristics in this cohort. Model 4 investigates the role of the cumulative risk factors, and Model 5 involves cumulative risk factors and child welfare experience. Finally, we present crosstabs with χ^2 and Fisher's tests to reveal whether specific psychiatric diagnoses are more common DP grounds among individuals with child welfare experience than among those without that experience.

Results

The descriptive statistics of our empirical material are shown in Table 1.

Table 1. Description of the dataset

	All (59116)		Boys (30221)		Girls (28895)	
	N	%	N		N	
Disability pension	1372	2.3	668	2.2	704	2.4
Disability pension with psychiatric diagnosis	866	1.5	383	1.3	483	1.7
Child welfare	1891	3.2	931	3.1	960	3.3
Child welfare + Disability pension	181	0.3	89	0.3	92	0.3
Child welfare + Disability pension with psychiatric diagnosis	135	0.2	68	0.2	67	0.2

Note: Persons who died before 18 years of age were excluded from the figures.
Original population 59,476, of which 30,435 are boys and 29,041 are girls.

The results from the logistic regression analyses are shown in Tables 2a-2c. We present odds ratios (ORs) and their 95% confidence intervals (95% CI). In the first model, including child welfare experience as the only covariate, we find a statistically significant association between child protection history and DP status (Table 2a, joint model for boys and girls). Compared to children without child welfare experience, individuals who have been subject to child welfare measures are overrepresented as DP recipients. The odds ratios in the model are 7.36 for boys (Table 2b) and 5.02 for girls (Table 2c).

Table 2a. Associations between DP and child welfare experience and covariates, including OR and 95% confidence intervals. Both genders' significant OR (0.05 level) in bold. Persons who died before 18 years of age and persons with DP with F70-79 diagnosis were excluded from the figures.

	N	Model 1			Model 2			Model 3			Model 4			Model 5		
		OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper
Child welfare	1861	**6.02**	4.98	7.28	**4.38**	3.52	5.44							**4.56**	3.68	5.64
Parents' social assistance	21234				**1.43**	1.22	1.68	**1.63**	1.39	1.90						
Mother's psych. in-/outpatient care	3796				1.15	0.90	1.45	**1.37**	1.09	1.73						
Father's psych. in-/outpatient care	3031				**1.42**	1.11	1.83	**1.54**	1.20	1.97						
Mother's psychiatric DP	907				**1.81**	1.28	2.57	**2.32**	1.65	3.26						
Mother's somatic DP	969				1.52	0.99	2.34	1.51	0.98	2.31						
Father's psychiatric DP	1173				1.34	0.93	1.92	1.43	1.00	2.05						
Father's somatic DP	2115				1.28	0.93	1.74	1.28	0.94	1.75						
Mother's death	595				0.58	0.31	1.08	0.82	0.44	1.51						
Father's death	1755				1.08	0.78	1.49	1.23	0.89	1.70						
Divorce	13327				1.14	0.98	1.34	1.17	1.00	1.37						
Mother's highest education																
High school or higher (ref.)	9383															
Lowest level tertiary	13604				0.89	0.71	1.13	0.89	0.70	1.12						
Lower secondary	26600				0.90	0.72	1.11	0.91	0.73	1.13						
Primary	9291				0.94	0.73	1.22	1.06	0.82	1.37						
Father's highest education																
High school or higher (ref.)	10657															
Lowest level tertiary	8121				0.92	0.71	1.18	0.90	0.70	1.16						
Lower secondary	25560				**0.81**	0.66	1.00	0.81	0.66	1.00						
Primary	14540				**0.72**	0.57	0.91	**0.77**	0.61	0.97						
Mother below 20 years old	1884				1.08	0.78	1.50	1.24	0.90	1.71						
Parents' cumulative risk, 0 (ref.)	21832										1.00			1.00		
Parents' cumulative risk, 1-2	26768										**1.28**	1.08	1.51	**1.20**	1.02	1.42
Parents' cumulative risk, 3-4	8773										**2.22**	1.84	2.68	**1.70**	1.39	2.07
Parents' cumulative risk, 5-9	1505										**3.90**	2.93	5.19	**2.06**	1.50	2.82

Table 2b. Associations between DP and child welfare experience and covariates, including OR and 95% confidence intervals for boys. Significant OR (0.05 level) in bold. Persons who died before 18 years of age and persons with DP with F70-79 diagnosis were excluded from the figures.

	N	Model 1			Model 2			Model 3			Model 4			Model 5		
		OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper
Child welfare	914	**7.36**	5.62	9.65	**4.88**	3.55	6.71							**5.47**	4.01	7.45
Parents' social assistance	10813				**1.78**	1.39	2.27	**2.00**	1.58	2.54						
Mother's psych. in-/outpatient care	1820				1.23	0.87	1.75	**1.56**	1.11	2.19						
Father's psych. in-/outpatient care	1485				1.39	0.96	2.02	**1.49**	1.02	2.16						
Mother's psychiatric DP	469				**2.08**	1.29	3.34	**2.68**	1.69	4.26						
Mother's somatic DP	492				0.73	0.30	1.80	0.74	0.30	1.81						
Father's psychiatric DP	632				1.38	0.83	2.29	1.58	0.96	2.60						
Father's somatic DP	1087				0.99	0.59	1.66	1.01	0.61	1.69						
Mother's death	304				0.39	0.12	1.26	0.53	0.17	1.69						
Father's death	881				0.95	0.57	1.59	1.09	0.66	1.82						
Divorce	6719				1.18	0.93	1.49	1.22	0.96	1.54						
Mother's highest education																
High school or higher (ref.)	4801															
Lowest level tertiary	6899				0.86	0.61	1.23	0.86	0.60	1.22						
Lower secondary	13655				0.93	0.67	1.29	0.95	0.69	1.32						
Primary	4732				0.97	0.66	1.43	1.13	0.77	1.65						
Father's highest education																
High school or higher (ref.)	5449															
Lowest level tertiary	4169				0.80	0.55	1.14	0.78	0.54	1.12						
Lower secondary	13074				**0.59**	0.44	0.81	**0.60**	0.44	0.82						
Primary	7395				**0.57**	0.40	0.80	**0.62**	0.44	0.88						
Mother below 20 years old	961				0.92	0.56	1.51	1.12	0.69	1.83						
Parents' cumulative risk. 0 (ref.)	11260										1.00			1.00		
Parents' cumulative risk. 1-2	13657										**1.42**	1.10	1.83	**1.33**	1.03	1.71
Parents' cumulative risk. 3-4	4398										**2.49**	1.86	3.31	**1.80**	1.33	2.45
Parents' cumulative risk. 5-9	772										**4.70**	3.10	7.13	**2.15**	1.35	3.43

Table 2c. Associations between DP and child welfare experience and covariates, including OR and 95% confidence intervals for girls. Significant OR (0.05 level) in bold. Persons who died before 18 years of age and persons with DP with F70-79 diagnosis were excluded from the figures.

	N	Model 1			Model 2			Model 3			Model 4			Model 5		
		OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper
Child welfare	947	**5.02**	3.85	6.55	**3.98**	2.94	5.38							**3.87**	2.89	5.19
Parents' social assistance	10421				1.21	0.98	1.51	**1.38**	1.12	1.70						
Mother's psych. in-/outpatient care	1976				1.05	0.75	1.46	1.20	0.87	1.66						
Father's psych. in-/outpatient care	1546				**1.43**	1.02	2.00	**1.56**	1.12	2.17						
Mother's psychiatric DP	438				1.52	0.89	2.58	**1.90**	1.13	3.19						
Mother's somatic DP	477				**2.15**	1.31	3.52	**2.12**	1.30	3.47						
Father's psychiatric DP	541				1.29	0.76	2.19	1.30	0.77	2.19						
Father's somatic DP	1028				**1.52**	1.03	2.25	**1.51**	1.02	2.23						
Mother's death	291				0.76	0.36	1.58	1.08	0.52	2.25						
Father's death	874				1.19	0.78	1.81	1.33	0.88	2.03						
Divorce	6608				1.12	0.90	1.39	1.13	0.91	1.40						
Mother's highest education																
High school or higher (ref.)	4582															
Lowest level tertiary	6705				0.91	0.67	1.24	0.90	0.66	1.23						
Lower secondary	12945				0.86	0.64	1.16	0.87	0.65	1.16						
Primary	4559				0.92	0.65	1.30	1.01	0.72	1.42						
Father's highest education																
High school or higher (ref.)	5208															
Lowest level tertiary	3952				1.04	0.74	1.47	1.03	0.73	1.45						
Lower secondary	12486				1.04	0.78	1.39	1.05	0.78	1.39						
Primary	7145				0.87	0.63	1.21	0.92	0.67	1.28						
Mother below 20 years old	923				1.24	0.80	1.91	1.34	0.87	2.06						
Parents' cumulative risk. 0 (ref.)	10572										1.00			1.00		
Parents' cumulative risk. 1-2	13111										1.17	0.94	1.46	1.12	0.90	1.39
Parents' cumulative risk. 3-4	4375										**2.02**	1.57	2.60	**1.62**	1.24	2.10
Parents' cumulative risk. 5-9	733										**3.34**	2.25	4.97	**1.98**	1.29	3.05

In Model 2, we include parental social assistance recipiency, parental psychiatric health variables, parental DP variables and other background characteristics. Child protection experience, parental social assistance, father's psychiatric care, mother's psychiatric DP, and father's low level of education are statistically significant in the model for the entire cohort. For boys, the statistically significant factors are child protection experience, parental social assistance, mother's psychiatric DP, and father's lowest and second-lowest educational level. In the model for girls, the statistically significant factors are child protection experience, father's psychiatric care, and mother's and father's somatic DP.

In the third model, we investigate the relationship between background characteristics excluding the child welfare experience. The results are mostly similar to the previous model, but the ORs are slightly higher. In the model for the entire cohort, mother's psychiatric care is statistically significant when the child welfare experience is not accounted for. For boys, both mother's and father's psychiatric care are statistically significant factors in this model. The difference in the model for girls compared to the previous model is that parental social assistance and mother's psychiatric DP are statistically significant factors.

To examine the effect of the accumulation of risk factors on DP, we estimated logistic regression analyses with only accumulated risk factors (Model 4) and risk factors and child welfare experience (Model 5) as covariates. Model 4 reveals that there is a statistically significant relationship between the accumulation of risk factors and DP. For individuals with parents with 1-2 risk factors, compared to those without any risk factors, the ORs for DP are 1.28 (entire cohort), 1.42 (boys) and 1.17 (girls, insignificant). An increase in risk factors involves an increase in ORs. For individuals with parents with 5-9 risk factors, the ORs are 3.90 (whole cohort), 4.70 (boys) and 3.34 (girls).

The final model involves accumulated risk variables and child welfare experience. Including the child welfare variable lowers the ORs for the risk factor variables. However, other than the cumulative risk with 1-2 risk factors among girls, all the cumulative risk variables are statistically significant. Children with child welfare experience are more likely to have DP. The ORs for those with child welfare experience compared to those without child welfare experience are 4.56 for the entire cohort, 5.47 for boys and 3.87 for girls.

To summarise the main findings from Tables 2a-2c, we find that child welfare experience has a statistically significant association with DP since ORs for the entire cohort are considerable. Parental social assistance has a statistically significant association with DP among boys, but it loses its significance in the model with the most controls among girls. Furthermore, mother's psychiatric DP makes a difference for boys, whereas father's psychiatric care and both father's and mother's somatic DP are statistically significant factors for girls. Parents' accumulated risk factors are important, and the more risk factors there are, the larger the OR is.

Finally, to investigate whether specific psychiatric diagnoses are more common DP grounds among individuals with child welfare experience than among those without that experience, we show several statistically significant differences between individuals with and without child welfare experience in Table 3. We use χ^2 tests to determine significant associations and ORs to indicate how strong the association is. For instance, girls with child welfare experience are more likely to receive DP than are girls without child welfare experience for all of the grounds except mania/bipolar disorder. Compared to those without child welfare experience, girls with that experience have an OR of 7.32 for having DP due to neurotic, stress-related and somatoform disorders. For boys, the pattern is very similar except that no statistically significant difference is found regarding neurotic, stress-related and somatoform disorders. There is no statistically significant difference regarding other psychiatric diagnoses among boys. Schizophrenia and schizotypal and delusional disorders have an OR of 10.12 and are the DP ground among boys, among whom child welfare clients most strongly differ from those without that experience. Multiple diagnoses also have a large OR for both genders, indicating that compared to those without child welfare experience, child welfare clients are clearly more prone to having DP due to multiple diagnoses.

Table 3. Crosstabs with OR and χ^2 tests (Fisher's Exact Test is reported when assumptions for χ^2 tests are not fulfilled)

		Disability pension				Schizophrenia. schizotypal and delusional disorders				Mania/bipolar disorder				Depression and mood disorders				Neurotic. stress-related and somatoform disorders				Other diagnoses				Multiple diagnoses				Intellectual disability				Other than F-diagnoses				Total
All		No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	No	Yes N	Yes %	OR p χ^2	
Child welfare	No	56034	1191	2.08	4.98 <0.001	57037	188	0.33	7.73 <0.001	57163	62	0.11	2.44 0.063	57045	180	0.31	4.25 <0.001	57179	46	0.08	5.94 <0.001	57134	91	0.16	4.68 <0.001	57061	164	0.29	6.56 <0.001	57017	208	0.36	4.42 <0.001	56973	252	0.44	1.93 0.01	57225
	Yes	1710	181	9.57		1844	47	2.49		1886	5	0.26		1866	25	1.32		1882	9	0.48		1877	14	0.74		1856	35	1.85		1861	30	1.59		1875	16	0.85		1891
Total		57744	1372	2.32		58881	235	0.40		59049	67	0.11		58911	205	0.35		59061	55	0.09		59011	105	0.18		58917	199	0.34		58878	238	0.40		58848	268	0.45		59116
Boys																																						
Child welfare	No	28711	579	1.98	5.24 <0.001	29184	106	0.36	10.12 <0.001	29270	20	0.07	1.57 0.482	29229	61	0.21	4.68 <0.001	29272	18	0.06	3.50 0.125	29238	52	0.18	5.49 <0.001	29232	58	0.20	7.69 <0.001	29173	117	0.40	4.64 <0.001	29143	147	0.50	0.86 1.000	29290
	Yes	842	89	9.56		898	33	3.54		930	1	0.11		922	9	0.97		929	2	0.21		922	9	0.97		917	14	1.50		914	17	1.83		927	4	0.43		931
Total		29553	668	2.21		30082	139	0.46		30200	21	0.07		30151	70	0.23		30201	20	0.07		30160	61	0.20		30149	72	0.24		30087	134	0.44		30070	151	0.50		30221
Girls																																						
Child welfare	No	27323	612	2.19	4.73 <0.001	27853	82	0.29	5.03 <0.001	27893	42	0.15	2.78 0.066	27816	119	0.43	3.96 <0.001	27907	28	0.10	7.32 <0.001	27896	39	0.14	3.74 0.015	27829	106	0.38	5.87 <0.001	27844	91	0.33	4.20 <0.001	27830	105	0.38	3.36 0.001	27935
	Yes	868	92	9.58		946	14	1.46		956	4	0.42		944	16	1.67		953	7	0.73		955	5	0.52		939	21	2.19		947	13	1.35		948	12	1.25		960
Total		28191	704	2.44		28799	96	0.33		28849	46	0.16		28760	135	0.47		28860	35	0.12		28851	44	0.15		28768	127	0.44		28791	104	0.36		28778	117	0.40		28895

Note: Persons who died before 18 years of age were excluded from the figures.
OR from binary logistic regression model without controls. Comparison group is individuals without child welfare experience.
Disability pension (DP on any of the grounds).
Schizophrenia, schizotypal and delusional disorders (DP based on only these psychiatric diagnoses; may involve somatic diagnoses).
Mania/bipolar disorder (DP based on only these psychiatric diagnoses; may involve somatic diagnoses).
Depression and mood disorders (DP based on only these psychiatric diagnoses; may involve somatic diagnoses).
Neurotic, stress-related and somatoform disorders (DP based on only these psychiatric diagnoses; may involve somatic diagnoses).
Other diagnoses (DP based on other psychiatric grounds, excluding intellectual disability (F70-79); may involve somatic diagnoses).
Multiple diagnoses (DP based on at least two of the psychiatric diagnoses above; may involve somatic diagnoses).
Intellectual disability (involves all who have received DP with diagnoses related to intellectual disability (F70-79); may also involve other psychiatric or somatic diagnoses).
Other than F-diagnoses (DP based only on somatic diagnoses).

To summarise the results in Table 3, children with child welfare experience are more likely than children without that experience to be granted DP on certain psychiatric grounds, such as schizophrenia, schizotypal and delusional disorders, depression and mood disorders, and neurotic, stress-related and somatoform disorders (girls only).

Discussion

Mental disorders are major grounds for DP in Finland. Even though most DPs are fixed-term, individuals who start on DP typically depart from the workforce permanently. Mental health and behavioural issues are the main reasons for DP among the younger DP recipients in Finland. Young DP recipients are problematic because the efficient functioning of the welfare state depends substantially on high labour market participation. The Finnish welfare state is already challenged by a population structure with large cohorts on their way to retirement and an unemployment level that is higher than it should be. Thus, there is substantial political concern regarding the working ability of the working-age population. The reasons for entry into DP among younger people are of great interest for both the research community and for those involved in decision-making, policy design and policy implementation.

This study aimed to investigate psychiatric diagnoses as grounds for DP among child welfare clients and to explore the background factors that are associated with psychiatric diagnoses and DP. Therefore, we reviewed the grounds for DP in a Finnish cohort. We investigated the grounds for DP in this cohort in general, but we also focused on a particularly vulnerable group, children who have been placed outside their homes by the child protection authorities. We found that these individuals are more prone to DP on psychiatric grounds than are individuals without that history. As previous research indicates, the impact of the accumulation of risk factors during childhood is essential even for this cohort. The accumulation of risk factors during childhood was found to be a statistically significant factor explaining DP in young adulthood. Furthermore, we investigated specific psychiatric diagnoses as grounds for a DP decision. To our knowledge, this is a novel undertaking because no previous studies have investigated psychiatric diagnoses as grounds for DP among individuals with a child protection history. Individuals with child welfare experience were overrepresented as DP recipients with regard to most of the specific psychiatric diagnoses as DP grounds.

We found a gender difference in the psychiatric grounds for DP. Schizophrenia, schizotypal and delusional disorders were more common grounds for DP among boys, whereas depression and mood disorders were more common grounds for DP among girls. This finding is not surprising because we know from previous research that women are clearly more prone to depressive disorders than men are. Similarly, we know that schizophrenia is somewhat more common among men than among women. Compared to those without child welfare experience, girls with that experience were clearly more prone to having DP due to neurotic, stress-related and somatoform disorders. Schizophrenia and schizotypal and delusional disorders is the DP ground among boys, among whom child welfare clients most strongly differ from those without that experience. Multiple diagnoses also have a strong association for both genders, indicating that child welfare clients are clearly more prone to obtain DP due to multiple diagnoses than individuals without that experience.

Schizophrenia has a substantial hereditary component, and there is a need for more detailed investigations relating to this group. We cannot draw more accurate conclusions based on the analysis conducted in this paper. However, we recognise that a closer investigation of the grounds on which child welfare officers based their decisions to place these children outside their homes may shed light on the pathways between child welfare clients' childhood histories and their DP. Currently, this information is not available in national registers due to difficulties in defining and classifying the reasons for out-of-home placement.

The results in this paper are consistent with previous empirical and theoretical research regarding cumulative advantage and disadvantage. Our results show that children whose parents have psychiatric or somatic problems for which they are on DP are also more likely to have a psychiatric diagnosis or to become a DP recipient in young adulthood. This intergenerational transmission of disadvantages is well established in the literature, and our findings support theories related both to the intergenerational transmission of inequality and

cumulative disadvantage. We also find that cumulative parental risk factors increase the risk for psychiatric DP, confirming the importance of cumulative risk factors.

Moreover, our results are in line with studies showing that socioeconomic circumstances during childhood are related to adult health status. Previous research shows that the probabilities for DP and participation in other welfare programs are higher for those individuals whose parents participated in these programs. Similarly, our results show that DP has a statistically significant association with both parental social assistance receipt and parental DP. More detailed analyses that involve an attempt to separate socially and biologically inherited behaviours and qualities would be an interesting but complex task.

There might also be interesting interactions between being outside the labor force and mental health, and the mechanisms behind these life course trajectories are of considerable interest. There are likely multiple reasons why individuals with a difficult path to adulthood fail in their transition to the labour market. For example, the roles that they have learned at home may involve learned helplessness or a lack of social skills and behaviours that are needed to function in the labour market. However, previous research has also shown that the economic stress that is related to unemployment is harmful to mental health (cf. Barr, Kinderman, & Whitehead, 2015). Thus, an individual who has satisfactory labour market prospects but for some reason fails in the transition to the labour market may experience economic stress due to unemployment, and this stress, in turn, has negative effects on mental health. The interconnections between these phenomena are obviously complex. Detailed investigations should therefore be undertaken, such as examining the role of mediating and moderating factors, including the ability of social and educational services to ameliorate the effects of challenging childhood conditions.

A child welfare history that includes out-of-home care indicates that there were severe problems in the home environment during upbringing. Our findings confirm previous research showing that when the number of risk factors increases, the dangerousness of the effects increases as well. Thus, exposure to multiple risks seems to have lasting effects on individuals. Future research should also consider whether the age when the out-of-home care occurs makes a difference for child welfare clients and their long-term outcomes. Are there critical periods related to out-of-home care? This is a difficult question to answer because there may be many different mechanisms, including selection mechanisms, that affect the outcome.

Limitations

Our variable information does not include any information about why the child was placed. From previous research, we know that reasons for placement for teenagers often involve a complex mixture of behavioural and school-related difficulties that may or may not involve psychiatric conditions. Further research is needed to completely rule out reverse causality.

There are also reasons to believe that some psychiatric diseases are underdiagnosed. It may be difficult to seek medical help for these because of social stigma, and those cases of psychiatric illness would not be identified in this study. Child welfare clients may also have less trust in the authorities, including medical practitioners, and their tendency to seek help may differ from the behaviour of individuals without child welfare experience.

Acknowledgements

This paper has benefitted from comments at a seminar at Norwegian Social Research in Oslo and at a meeting organised by the Nordic network on register-based child welfare research in Stockholm. The usual disclaimer applies.

This study was supported by two grants from the Academy of Finland, grant number 288960 (Time trend changes of child and adolescent mental health, service use and well-being in multiple Finnish cohorts) and grant number 308552 (PSYCHORT). The work also benefited from two personal grants from the Finnish Work Environment Fund and Alli Paasikivi Foundation (Tiina Ristikari).

The Finnish Birth Cohort 1987 study obtained a positive statement from the research ethical committee of the National Institute for Health and Welfare (Ethical committee §28/2009), and permissions to use the register data was obtained from all register holding organisations.

References

Barr, B., Kinderman, P., & Whitehead, M. (2015). Trends in mental health inequalities in England during a period of recession, austerity and welfare reform 2004 to 2013. *Social Science and Medicine*, *147*, 324-331. https://doi.org/10.1016/j.socscimed.2015.11.009

Bask, M., & Bask, M. (2015). Cumulative (dis)advantage and the Matthew effect in life-course analysis. *PLoS ONE*, *10*(11), e0142447. https://doi.org/10.1371/journal.pone.0142447

Bask, M., Ferrer-Wreder, L., Salmela-Aro, K., & Bergman, L. R. (2014). Pathways to educational attainment in middle adulthood: the role of gender and parental educational expectations in adolescence. In J. S. Eccles, & I. Schoon (Eds.), *Gender differences in aspirations and attainment: a life course perspective* (pp. 389-411). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139128933.023

Ben-Shlomo, Y., & Kuh, D. (2002). A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, *31*(2), 285-293. https://doi.org/10.1093/intjepid/31.2.285

Berlin, M., Vinnerljung, B., & Hjern, A. (2011). School performance in primary school and psychosocial problems in young adulthood among care leavers from long term foster care. *Children and Youth Services Review*, *33*(12), 2489-2497. https://doi.org/10.1016/j.childyouth.2011.08.024

Bowen, M. E., & González, H. M. (2010). Childhood socioeconomic position and disability in later life: results of the health and retirement study. *American Journal of Public Health*, *100*, S197-S203. https://doi.org/10.2105/AJPH.2009.160986

Bratberg, E., Nilsen, Ø. A., & Vaage, K. (2015). Assessing the intergenerational correlation in disability pension recipiency. *Oxford Economic Papers*, *67*(2), 205-226. https://doi.org/10.1093/oep/gpu028

Chen, E., Matthews, K. A., & Boyce, W. T. (2002). Socioeconomic differences in children's health: how and why do these relationships change with age? *Psychological Bulletin*, *128*(2), 295-329. https://doi.org/10.1037/0033-2909.128.2.295

Cohen, S., Janicki-Deverts, D., Chen, E., & Matthews, K. A. (2010). Childhood socioeconomic status and adult health. *Annals of the New York Academy of Sciences*, *1186*, 37-55. https://doi.org/10.1111/j.1749-6632.2009.05334.x

Costello, E. J., Erkanli, A., Copeland, W., & Angold, A. (2010). Association of family income supplements in adolescence with development of psychiatric and substance use disorders in adulthood among an American Indian population. *JAMA*, *303*(19), 1954-1960. https://doi.org/10.1001/jama.2010.621

Crystal, S., & Shea, D. (1990). Cumulative advantage, cumulative disadvantage, and inequality among elderly people. *Gerontologist*, *30*(4), 437-443. https://doi.org/10.1093/geront/30.4.437

Dahl, G. B., Kostøl, A. R., & Mogstad, M. (2014). Family welfare cultures. *Quarterly Journal of Economics*, *129*(4), 1711-1752. https://doi.org/10.1093/qje/qju019

Doyle, J. J. Jr. (2007). Child protection and child outcomes: measuring the effects of foster care. *American Economic Review*, *97*(5), 1583-1610. https://doi.org/10.1257/aer.97.5.1583

ETK. (2015). Statistical yearbook of pensioners in Finland. Accessed July 8, 2017, from http://www.etk.fi/wp-content/uploads/Statistical_yearbook_of_pensioners_in_Finland_2014.pdf

ETK. (2017). Disability pension. Accessed July 8, 2017, from http://www.etk.fi/en/the-pension-system-2/the-pension-system/pension-benefits/earnings-related-pensions/disability-pension/

Evans, G. W., & Cassells, R. C. (2014). Childhood poverty, cumulative risk exposure, and mental health in emerging adults. *Clinical Psychological Science*, *2*(3), 287-296. https://doi.org/10.1177/2167702613501496

Evans, G. W., Li, D., & Whipple, S. S. (2013). Cumulative risk and child development. *Psychological Bulletin*, *139*(6), 1342-1396. https://doi.org/10.1037/a0031808

Farand, L., Chagnon, F., Renaud, J., & Rivard, M. (2004). Completed suicides among Quebec adolescents involved with juvenile justice and child welfare services. *Suicide and Life-Threatening Behavior*, *34*(1), 24-35.

Franzén, E., Vinnerljung, B., & Hjern, A. (2008). The epidemiology of out-of-home care for children and youth: a national cohort study. *British Journal of Social Work*, *38*(6), 1043-1059. https://doi.org/10.1093/bjsw/bcl380

Harkonmäki, K., Korkeila, K., Vahtera, J., Kivimäki, M., Suominen, S., Sillanmäki, L., & Koskenvuo, M. (2007). Childhood adversities as a predictor of disability retirement. *Journal of Epidemiology and Community Health*, *61*(6), 479-484. https://doi.org/10.1136/jech.2006.052670

Heneghan, A., Stein, R. E. K., Hurlburt, M. S., Zhang, J., Rolls-Reutz, J., Fisher, E., Landsverk, J., & Horwitz, S. M. (2013). Mental health problems in teens investigated by U.S. child welfare agencies. *Journal of Adolescent Health*, *52*(5), 634-640. https://doi.org/10.1016/j.jadohealth.2012.10.269

Jackson, S., & Cameron, C. (2012). Leaving care: looking ahead and aiming higher. *Children and Youth Services Review*, *34*(6), 1107-1114. https://doi.org/10.1016/j.childyouth.2012.01.041

Kestilä, L., Väisänen, A., Paananen, R., Heino, T., & Gissler, M. (2012). Kodin ulkopuolelle sijoitetut nuorina aikuisina: rekisteripohjainen seurantatutkimus Suomessa vuonna 1987 syntyneistä [Children placed in out-of-home care as young adults. A register-based follow-up study on children born in 1987 in Finland]. *Yhteiskuntapolitiikka*, *77*, 599-620.

Laaksonen, M., Blomgren, J., & Tuulio-Henriksson, A. (2016). Sickness allowance histories among disability retirees due to mental disorders: a retrospective case-control study. *Scandinavian Journal of Public Health*, *44*(3), 291-299. https://doi.org/10.1177/1403494815618314

Lynch, J., & Smith, G. D. (2005). A life course approach to chronic disease epidemiology. *Annual Review of Public Health*, *26*, 1-35. https://doi.org/10.1146/annurev.publhealth.26.021304.144505

Mersky, J. P., & Janczewski, C. (2013). Adult well-being of foster care alumni: comparisons to other child welfare recipients and a non-child welfare sample in a high-risk, urban setting. *Children and Youth Services Review*, *35*(3), 367-376. https://doi.org/10.1016/j.childyouth.2012.11.016

Mykletun, A., Overland, S., Dahl, A. A., Krokstad, S., Bjerkeset, O., Glozier, N., Aarø, L. E., & Prince, M. (2006). A population-based cohort study of the effect of common mental disorders on disability pension awards. *American Journal of Psychiatry*, *163*(8), 1412-1418. https://doi.org/10.1176/ajp.2006.163.8.1412

Najman, J. M., Hayatbakhsh, M. R., Clavarino, A., Bor, W., O'Callaghan, M. J., & Williams, G. M. (2010). Family poverty over the early life course and recurrent adolescent and young adult anxiety and depression: a longitudinal study. *American Journal of Public Health*, *100*(9), 1719-1723. https://doi.org/10.2105/AJPH.2009.180943

Organisation for Economic Co-operation and Development. (2014). OECD economic surveys: Finland 2014. Accessed July 8, 2017, from http://www.oecd-ilibrary.org/economics/oecd-economic-surveys-finland-2014_eco_surveys-fin-2014-en

Paananen, R., Ristikari, T., Merikukka, M., & Gissler, M. (2013). Social determinants of mental health: a Finnish nationwide follow-up study on mental disorders. *Journal of Epidemiology and Community Health*, *67*(12), 1025-1031. https://doi.org/10.1136/jech-2013-202768

Reiss, F. (2013). Socioeconomic inequalities and mental health problems in children and adolescents: a systematic review. *Social Science and Medicine*, *90*, 24-31. https://doi.org/10.1016/j.socscimed.2013.04.026

Ristikari, T., Hakovirta, M., & Gissler, M. (2016). The impact of timing and duration of parental social assistance receipt on early adult outcomes. Manuscript submitted for publication.

THL. (2014a). Lastensuojelu 2014 [Child welfare 2014]. Accessed July 8, 2017, from http://www.julkari.fi/bitstream/handle/10024/129537/Tr25_15.pdf?sequence=4

THL. (2014b). Liitetaulukko 1. Kodin ulkopuolelle sijoitetut lapset ja nuoret viimeisimmän sijoitustiedon mukaan, 2012-2014 [Children and young people placed outside the home based on the last placement decision, 2012-2014]. Accessed July 8, 2017, from https://www.thl.fi/tilastoliite/tilastoraportit/2015/liitetaulukot/Tr25_15_liite1.xls

Upmark, M., Lundberg, I., Sadigh, J., & Bigert, C. (2001). Conditions during childhood and adolescence as explanations of social class differences in disability pension among young men. *Scandinavian Journal of Public Health*, *29*(2), 96-103. https://doi.org/10.1177/14034948010290020601

Upmark, M., & Thundal, K. L. (2002). An explorative, population-based study of female disability pensioners: the role of childhood conditions and alcohol abuse/dependence. *Scandinavian Journal of Public Health*, *30*(3), 191-199. https://doi.org/10.1177/140349480203000305

Verbrugge, L. M., & Jette, A. M. (1994). The disablement process. *Social Science and Medicine*, *38*(1), 1-14. https://doi.org/10.1016/0277-9536(94)90294-1

Vinnerljung, B., Brännström, L., & Hjern, A. (2015). Disability pension among adult former child welfare clients: a Swedish national cohort study. *Children and Youth Services Review*, *56*, 169-176. https://doi.org/10.1016/j.childyouth.2015.07.001

Vinnerljung, B., Hjern, A., & Lindblad, F. (2006). Suicide attempts and severe psychiatric morbidity among former child welfare clients – a national cohort study. *Journal of Child Psychology and Psychiatry*, *47*(7), 723-733. https://doi.org/10.1111/j.1469-7610.2005.01530.x

Adverse childhood experiences, non-response and loss to follow-up: Findings from a prospective birth cohort and recommendations for addressing missing data

James C. Doidge University College London, UK, and University of South Australia
j.doidge@ucl.ac.uk
Ben Edwards Australian National University and Australian Institute of Family Studies
Daryl J. Higgins Australian Catholic University and Australian Institute of Family Studies
Leonie Segal University of South Australia

(Received April 2016 Revised January 2017) http://dx.doi.org/10.14301/llcs.v8i4.414

Abstract

Adverse childhood experiences have wide-ranging impacts on population health but are inherently difficult to study. Retrospective self-report is commonly used to identify exposure but adult population samples may be biased by non-response and loss to follow-up. We explored the implications of missing data for research on child abuse and neglect, domestic violence, parental mental illness and parental substance use. Using 15 waves of data collected over 28 years in a population-based birth cohort, the Australian Temperament Project, we examined the relationship between retrospective self-reports of adverse childhood experiences and parent- and cohort-responsiveness at other time points. We then compared prevalence estimates under complete case analysis, inverse probability-weighting using baseline auxiliary variables, multiple imputation using baseline auxiliary variables, multiple imputation using auxiliary variables from all waves, and multiple imputation using additional measures of participant responsiveness. Retrospective self-reports of adverse childhood experiences were strongly associated with non-response by both parents and cohort members at all observable time points. Biases in complete case estimates appeared large and inverse probability-weighting did not reduce them. Multiple imputation increased the estimated prevalence of any adverse childhood experiences from 30.0% to 36.9% with only baseline auxiliary variables, 39.7% with a larger set of auxiliary variables and 44.0% when measures of responsiveness were added. Close attention must be paid to missing data and non-response in research on adverse childhood experiences as data are unlikely to be missing at random. Common approaches may greatly underestimate their prevalence and compromise analysis of their causes and consequences. Sophisticated techniques using a wide range of auxiliary variables are critical in this field of research, including, where possible, measures of participant responsiveness.

Keywords

Adverse childhood experiences, child abuse and neglect, missing data, selection bias, response bias, survey non-response, loss to follow-up, cohort attrition, multiple imputation, inverse probability-weighting

Introduction

Adverse childhood experiences such as child maltreatment and exposure to parental mental illness or substance abuse have widespread health and socioeconomic consequences (Fang, Brown, Florence, & Mercy, 2012; Gilbert et al., 2009; Norman et al., 2012). There are at least two significant hurdles in the measurement of adverse childhood experiences. One relates to the elicitation of information about potentially traumatic or illegal events, often long after the fact (Dube, Williamson, Thompson, Felitti, & Anda, 2004; Wyatt & Peters, 1986), and aligning this information with a consistent set of definitions (Besharov, 1981). The other, which is the focus of this paper, is the problem of eliciting any information at all from a population group with many barriers to participation in research (Edwards et al., 2001; Haugaard & Emery, 1989). Most of the risk factors and outcomes associated with adverse childhood experiences are likely to be associated with higher rates of non-response. The more severe outcomes, such as homelessness (Herman, Susser, Struening, & Link, 1997), incarceration (Widom & Maxfield, 2001) and death (Brown et al., 2009), are likely to result in loss to follow-up in longitudinal studies and may exclude affected individuals from cross-sectional sampling frames altogether. These types of 'missing data' are likely to lead to underrepresentation of people with adverse childhood experiences in population-based research and higher levels of incomplete data. This will at least affect estimates of prevalence and may also have implications for research on the causes and consequences of childhood adversity.

A common method for identifying exposure to adverse childhood experiences is through retrospective self-report. Samples may be recruited either in adulthood (e.g. cross-sectional surveys) or in childhood (e.g. prospective cohorts). Birth cohorts provide an opportunity to collect some information about participants prior to exposure occurring and prior to any possible influence of outcomes on participant responsiveness. They may therefore offer some of the greatest potential for measuring the prevalence of adverse childhood experiences—but only if subsequent cohort attrition and missing data can be dealt with effectively.

The implications of missing data depend on the associations between any variables of interest and the probability that some or all of the relevant data are missing (Rubin, 1976). 'Variables of interest' include any variables that are necessary for estimating results; in our case, indicators of adverse childhood experiences and the correlates that we wished to investigate—risk factors for and outcomes of adverse childhood experiences (although for simplicity, this paper focuses on estimates of the prevalence of adverse childhood experiences). The joint distribution of missingness in variables of interest conditional on the data is known as the *missingness mechanism* (Schafer & Graham, 2002). It is important to note that *missingness mechanism* does not refer to the real-world process that results in data being missing (e.g. the participant died or could not be contacted).

There are three classifications of missing data that have arisen in the literature: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR, also called *not missing at random,* NMAR). There is some variation in the definitions that have been proposed, partly because their formal definition depends on the framework for inference that is being used (for a detailed explanation, see Seaman, Galati, Jackson, & Carlin, 2013). These can be thought of as either assumptions about the missingness mechanism that are made for a specified analysis or as classifications of the missingness mechanism in the context of the analysis. If missingness is unrelated to observed or unobserved value of any variables of interest then the data are said to be MCAR. If data are MCAR, many types of analysis can produce unbiased estimates, including simpler approaches such as complete case analysis (the default in most statistical software) (Rubin, 1976). For an estimate of the prevalence of adverse childhood experiences, MCAR requires that there be no association between adverse childhood experiences and the probability of there being missing data about adverse childhood experiences.

Sometimes there is an association between some variables of interest and the probability of data being missing, but this can be explained by the observed values of other variables (i.e. after conditioning on the other variables, there is no association). This

would be the case, for example, if men were more likely to have adverse childhood experiences and more likely to be lost to follow-up, but there were no other relationships between adverse childhood experiences and the probability of data being missing and the distribution of gender with respect to missing data was known (for example, because there was no missing data about gender). In this case, data are MAR conditional on gender, a less restrictive and generally more plausible assumption than MCAR. Under MAR, valid estimation may still be obtained using three groups of techniques: multiple imputation, maximum likelihood and inverse probability-weighting (Schafer & Graham, 2002). This paper focuses multiple imputation and inverse probability-weighting. Recent and comprehensive reviews of these procedures can be found in Carpenter and Kenward (2013) and Seaman and White (2013), respectively.

When there are associations between variables of interest and the probability of data being missing, which persist after conditioning on observed data, then data are said to be *missing not at random* (MNAR; or *not missing at random, NMAR*) and estimates will generally be biased (Rubin, 1976). This would be the case, for example, if people who experienced adverse childhood events were less likely to participate, and this systematic difference in participation could not be explained by observed variables. There are certain instances where valid inference can be made under MNAR (e.g. Bartlett, Carpenter, Tilling, & Vansteelandt, 2014), although this arguably reflects that limitations of the MCAR/MAR/MNAR classification system for reflecting the nuanced variation in assumptions about missing data that can be are implied in different analyses (strictly, it is the *ignorability* of the missingness mechanism for a given analysis which matters, rather than its classification). One important thing to note, though, is that whether data are MAR or MNAR (whether the MAR assumption holds) depends critically on the observed data that are fed into an analysis; the more informative they are about the missingness mechanism, the more plausible the MAR assumption becomes. Put simply, the more that is known about people with missing data, the more reliable the analysis.

Life course studies in which a broad range information is collected at many points in time offer good potential for addressing missing data through auxiliary variables—additional observed variables that are not required for analysis models but are associated with missing variables of interest or the probability of data being missing. Methods for addressing missing data under MAR involve creating at least two models: a substantive or analytic model for the parameters of interest, and either an imputation model for missing values or a selection (response) model for the probability of being data not being missing (Cole, 2008). Including auxiliary variables in the imputation or selection models can reduce bias because of information that is gleaned from their association with the incompletely observed variable and its probability of missingness. However, including variables that are only associated with the probability of missingness may reduce precision without reducing bias (Collins, Schafer, & Kam, 2001; Seaman & White, 2013) while, conversely, auxiliary variables that are only associated with the variable of interest and not its probability of missingness will increase precision but not reduce bias (White, Royston, & Wood, 2011). The stronger the associations between auxiliary variables and the incompletely observed variables or the probability of data being missing, the greater their potential for reducing bias (Hardt, Herke, & Leonhart, 2012).

As well as the explicitly recorded variables, longitudinal studies also include many opportunities to observe participants' responsiveness, such as the proportion of surveys completed and items missed within each survey. The potential value of directly utilising measures of participant responsiveness when addressing missing data was recently demonstrated in a simulation study (Doidge, 2016) but has not yet been applied to real-world data. *Indirectly* or descriptively utilising measures of participant responsiveness is relatively common; this is the basis of follow-up studies of non-respondents and related approaches (Fielding, Fayers, & Ramsay, 2009).

The aims of this study were (1) examine the relationship between adverse childhood experiences, non-response and loss to follow-up in prospective birth cohort, and (2) to compare estimates for the prevalence of adverse childhood experiences

obtained using different approaches to addressing missing data, differing primarily in their use of auxiliary variables. The motivation for this analysis was to establish an optimal basis upon which a set of related analyses concerning adverse childhood experiences could be conducted in a population-based birth cohort with substantial missing data from non-response and loss to follow-up. While the methods evaluated are primarily relevant to longitudinal studies, the findings may be generalisable to cross-sectional settings.

Methods

Participants

All data were derived from the Australian Temperament Project (ATP), a prospective birth cohort that has been previously described (Prior, Sanson, Smart, & Oberklaid, 2000; Vassallo & Sanson, 2013). The sampling frame was designed to select a cohort that represents people born in 1983 in the Australian state of Victoria in terms of socioeconomic status and urban/rural locality (Sanson & Oberklaid, 1985). Questionnaires were completed by Maternal and Child Health Nurses and caregivers of infants aged four – eight months during a two-week period in 1983. Since the initial survey, 15 waves of follow-up questionnaires have been administered to parents, teachers (3 waves) and cohort members (nine waves) over 32 years. The cohort initially consisted of 2,443 infants and their parents who are the focus of this study. Follow-up and response are summarised in Table 1 and illustrated in Figure 1.

Table 1. Response rates in the Australian Temperament Project

Year	Wave	Age (years)	Families in contact	Number of responses received[a]			
				Parent	Cohort	Teacher	Nurse
1983	1	<1	—	2443			2443
1984	2	1-2	2226	1280			
1985	3	2-3	2234	1357			
1986	4	3-4	2286	1717			
1988	5	5-6	1785	1727		1428	
1990	6	7-8	1874	1603		1256	
1992	7	9-10	1799	1544			
1994	8	11-12	1743	1471	1452	1238	
1995	9	12-13	1661	1275	1228		
1996	10	13-14	1670	1391	1358		
1998	11	15-16	1666	1379	1306		
2000	12	17-18	1650	1308	1259		
2002	13	19-20	1580	1103	1158		
2006	14	23-24	1505	968	1000		
2010	15	27-28	1701	940	1052		

[a]Includes responses relating to 71 cohort members who were recruited after Wave 1 and excluded from analyses reported in this paper.

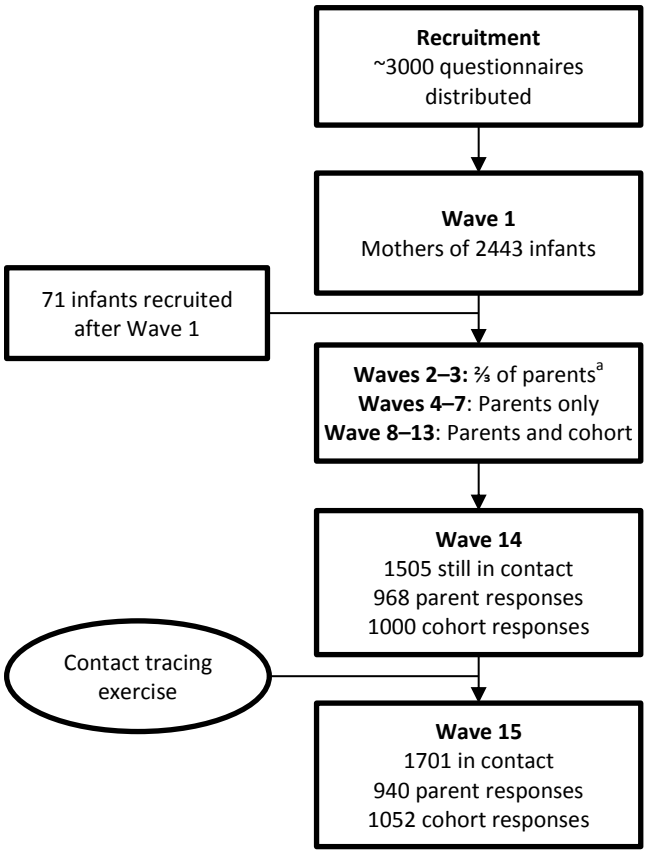


Figure 1. Participation in the Australian Temperament Project

[a]In Waves 2 and 3 a random ⅔ of families were selected for surveying.

Data

Exposures to adverse childhood experiences (physical, sexual and emotional abuse, neglect, witnessing domestic violence and parental mental illness or substance use) were collected by retrospective self-report in Wave 14 (22–23 years) and are described in Table 2.

We identified an initial set of potential auxiliary variables from theory and literature on potential risk factors for child maltreatment (sociodemographic factors, economic factors, parental alcohol and tobacco use, child health and temperament) and outcomes associated with adverse childhood experiences (economic outcomes, social outcomes, mental health and substance use, and physical health). These were then descriptively and visually analysed to identify variables and categories that were associated with adverse childhood experiences and with missingness in questions about adverse childhood experiences. Most risk factors were reported by parents early in the study and outcomes were reported by cohort members in Waves 14 (22–23 years) and 15 (27–28 years).

An additional set of auxiliary variables was derived directly from indicators of participant responsiveness. Theoretical justifications for including measures of responsiveness are discussed in Doidge (2016). Figure 2 is a directed acyclic graph illustrating the hypothesised causal relationships between adverse childhood experiences, responsiveness and other auxiliary variables (risk factors and outcomes of adverse childhood experiences). Solid arrows represent relationships that we consider to be justified by theory or literature and the dashed arrow represents a possible effect of poor parenting practices on reducing parents' willingness to participate in research.

Measures of responsiveness included: the proportion of surveys returned by parents prior to their final response (set to missing if lost before Wave 4 so as early loss to follow-up would not intrinsically imply low levels of responsiveness), the proportion of items completed by parents in the first questionnaire (observed for everyone), the proportion of items completed on average in the remaining waves to which they responded, whether the cohort members responded to both waves 14 and 15 or just one (neither was set to missing), and the average proportion of items completed by cohort members in the waves to which they responded. Cohort response prior to wave 14 was excluded because of the close dependence of cohort responsiveness on parent responsiveness during adolescence. These measures of participant responsiveness were derived from 363 items that were selected to indicate 119 variables of interest (risk factors and outcomes of child maltreatment) across the domains listed above. A complete list of auxiliary variables is provided in the Supplementary Appendix.

Table 2. Identification of adverse childhood experiences

Item	Response coding
Emotional abuse *You experienced verbal treatment from your parent/s that made you feel embarrassed, humiliated or scared (e.g. shouting, name calling, threats)*	*1 = Very true** *2 = Somewhat true** *3 = Uncertain* *4 = Somewhat untrue* *5 = Not at all true*
Neglect *The care taken of you by your parent/s was the right amount (e.g. they watched out for you, fed you properly, gave you attention)*	*1 = Very true* *2 = Somewhat true* *3 = Uncertain* *4 = Somewhat untrue** *5 = Not at all true*
Physical abuse *1. Your parent/s used harsh physical treatment (e.g. smacking hitting) to discipline you* *2. Did you ever suffer effects that lasted to the next day or longer (e.g. bruising, marking, pain, soreness)?*	*1 = No* *2 = Yes* Coded if response = 'yes' to both questions*
Sexual abuse *1. A family member did, or tried to do sexual things to you* *2A. You had a sexual experience with a person who was not a family member before you were 16* *2B. Was this consensual?*	*1 = No* *2 = Yes* Coded if respondent answered 'yes' to 1, or 'yes' to 2A and 'no' to 2B*
Witnessing domestic violence *There was physical violence between the adults caring for you*	*1 = Very true* *2 = Somewhat true** *3 = Uncertain* *4 = Somewhat untrue* *5 = Not at all true*
Parental mental illness and substance use problems *1. Your mother or father had a mental illness or substance use problem.* *2.* If yes: *Who experienced the problem?* *3. Please describe the problem/s*	1. *No/Yes** 2. *Mother/Father/Both parents* [free text, coded by researchers as mental illness, substance use problem or both]

* threshold adopted for classification (in the case of emotional abuse, we subdivided participants into those reporting less severe (somewhat true) or more severe (very true) abuse)

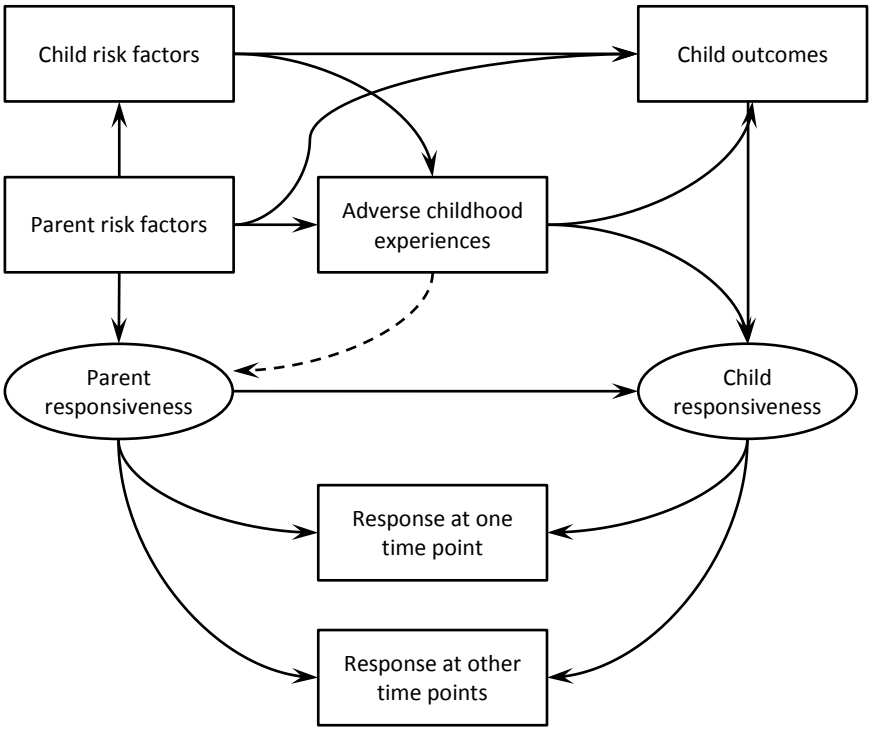


Figure 2. Directed Acyclic Graph (DAG) representing theoretical relationships between adverse childhood experiences and auxiliary variables

Arrows indicate postulated effects and the dashed arrow indicates a potential effect of abusive or neglectful parenting on making the parents less willing to participate in research. While direction or causality is not required for multiple imputation or inverse probability weighting to work, these relationships provide a rationale for expecting associations between adverse childhood experiences and auxiliary variables, including measures of responsiveness. Further: if, as illustrated, variables can only be associated with missing data through their associations with responsiveness, then adequate measures of responsiveness would be sufficient to satisfy the *missing at random* assumption. However, responsiveness is unobserved and 'measures' of responsiveness are based primarily on indicators of response (missing data) at other time points.

Statistical analysis

The association between child maltreatment and missing data was first explored by directly examining correlations (odds ratios) between child maltreatment and measures of non-response by parents in prior waves and by cohort members in Wave 15. Non-response by parents was examined from Wave 4, as only a subset of the cohort were sampled in Waves 2 and 3 and it was not possible to distinguish participant non-response from exclusion at these points.

The prevalence of adverse childhood experiences were estimated using five methods, with progressively more auxiliary variables in each: (1) complete case analysis, (2) inverse probability-weighting (IPW) using a limited set of baseline variables (cohort sex, mother/father educations < diploma, mother/father occupations class < professional or managerial, mother/father aged < 22 years, birthweight < 3rd percentile, premature), (3) multiple imputation using the same baseline variables ('MI baseline'), (4) multiple imputation using all explicitly measured auxiliary variables (risk factors and outcomes associated with adverse childhood experiences; 'MI full'), and (5) multiple imputation using all explicitly measured auxiliary variables plus measures of responsiveness (termed 'responsiveness-informed multiple imputation', 'RMI', for short, but differing from previous forms of multiple imputation only in the inclusion of responsiveness among auxiliary variables). Measures of responsiveness were not able to added to an inverse probability-weighted approach because they had missing data for participants lost to follow-up within the first three waves of data collection, when there was insufficient time over which to observe responsiveness.

The only differences between the multiple imputation methods were in the sets of auxiliary variables included in imputation models. Generally, the inclusion of additional auxiliary variables is little threat to the validity of an analysis (Collins et al., 2001; Enders, 2010a; Seaman & White, 2013). However, the noise created by large numbers of weak auxiliary variables may bias regression analyses towards the null hypothesis in small samples (Hardt et al., 2012) and Thoemmes and Rose (2014) describe a special case in which conditioning on auxiliary variables may introduce dependence between variables of interest and the probability of missing data through a form of collider bias. Targeted selection of auxiliary variables, based on their observed correlations with variables of interest and the probability of missing data, is likely to avoid both of these concerns, and our use of discrete outcome imputation models limits the potential for introduction of error from imputation model misspecification. In his simulation study, Doidge (2016) observed reductions in bias whenever measures of responsiveness where added as auxiliary variables. We therefore interpret any substantial differences in the point estimates obtained using different methods as reflecting lower levels of bias in the methods with more inclusive (but still targeted) selection of auxiliary variables.

Auxiliary variables were collapsed into binary or ordinal indicators that best discriminated risk of adverse childhood experiences. Imputation was performed using chained equations (fully conditional specification) (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001; Royston, 2004). This is an approach to multiple imputation in which a separate imputation model is specified for each variable with missing data. Estimating imputation models iteratively ('chaining') allows for implementation with nonmonotone patterns of missing data (or non-nested imputation models in monotone missing data patterns). This approach accommodates large numbers of non-normal and discrete variables, and allows for different auxiliary variables to be selected for each imputation model. All imputation models were either logit or ordered logit. As independent variables in imputation models, ordinal variables were treated as continuous, to allow inclusion of a greater variety of independent variables in imputation models. Convergence of chained equations was visually assessed using trace plots of 100 imputations, which indicated that 10 burn-in iterations would be sufficient to achieve convergence. Factoring advice from the literature (Rubin, 1987) and the significant computational demands of imputing large numbers of variables in multiple ways for comparison, we selected 20 imputations as being likely to be sufficient for maximising power while also being computationally feasible. Using Rubin's (1987)

formula for relative efficiency and 65% proportion of missing values (the highest level of missingness for any variable), the relative efficiency of 20 imputations was estimated to be 98.5%.

For multiple imputation using baseline measures only (MI baseline), each imputation model included every other variable (i.e. other measures of adverse childhood experiences and baseline variables). For multiple imputation using all explicitly-measured auxiliary variables (MI full), imputation models for adverse childhood experiences included sex, all economic risk factors, mother aged < 22 years at baseline, parental immigration from non-English-speaking countries, parental separation, school mobility, household mobility, parental smoking, maternal alcohol use during childhood, at least two investigated health problems by age three, premature birth, birthweight > third percentile, retrospective self-report of cognitive or behavioural and physical health problems while growing up, weight status in Wave 14 and 15, mental health conditions in Wave 14, been charged by police in Wave 14, frequency of antisocial behaviour in Wave 14, occupational class in Wave 15 and income in Wave 15. For MI full, imputation models for auxiliary variables at least included indicators of adverse childhood experiences, sex and as many theoretically relevant variables (e.g. those from the same conceptual domain) as could be incorporated without computation errors. Responsiveness-informed multiple imputation (RMI) models additionally included indicators of participant responsiveness in every imputation model.

All analyses were conducted using Stata 14 (StataCorp 2015, College Station, TX).

Results

At Wave 14, when cohort members were aged 22–23 years and were asked about adverse childhood experiences, 1,505 families were still enrolled and contactable and responses were received from 1,000 cohort members. Of these, 20 were twins that had been enrolled post-baseline and were excluded from analysis to maintain population representation (in light of the higher risk of child maltreatment associated with multiple births (Wu et al., 2004)), 40 had partial data from questions about adverse childhood experiences and 940 had complete data on these items.

Participant characteristics by completeness on questions about adverse childhood experiences are presented in Table 3. Comparing characteristics of the cohort at baseline with this subset of 'complete cases', it can be seen that men were more likely to be missing, as were those whose parents were young (strongly so), immigrants, less educated, or with lower occupational classes. Inverse-probability weighting did not appear to fully adjust for the strong relationships that were observed between young parental age and loss to follow-up/non-response. Conversely, as parental ages were recorded at baseline and had very little missing data, the prevalence estimates obtained using multiple imputation were almost identical to the item-complete estimates for the whole cohort.

Examining the relationship between adverse childhood experiences reported in Wave 14 and non-response in other waves, strong associations were observed with both non-response by parents and non-response by cohort members, and the association with parent responsiveness appeared relatively stable over time (Figure 3). Similar patterns of associations were observed for most combinations of specific adverse childhood experience and response measures (Supplementary Table S2). Parents of those reporting adverse childhood experiences also exhibited a higher level of incomplete responses although this relationship varied across adverse childhood experiences among cohort members (Supplementary Table S2). Indicators of responsiveness were all strongly associated with missingness in indicators of adverse childhood experiences (Supplementary Table S3).

Nearly all of the hypothesised covariates (risk factors and outcomes of adverse childhood experiences) were associated with both adverse childhood experiences and with the probability of data about adverse childhood experiences being missing (results not shown). Ordinal, multinomial and continuous covariates were collapsed into binary categories that best discriminated risk of adverse childhood experiences.

Table 3. Baseline participant characteristics by completeness on adverse childhood

Variable	Estimated prevalence, by method, % (SE)						
	Whole cohort	Missing	CCA	IPW	MI (baseline)	MI (full)	RMI
n	*2443*		*940*	*940*	*2443*	*2443*	*2443*
Cohort characteristics							
Female	48.1 (1.0)	0.0	61.3 (1.6)	48.1 (1.9)	51.9 (1.0)	51.9 (1.0)	51.9 (1.0)
Birthweight < 3rd percentile	3.2 (0.4)	11.3	2.9 (0.6)	3.1 (0.7)	3.2 (0.4)	3.4 (0.4)	3.6 (0.4)
Parent characteristics							
Either parent immigrated from non-English-speaking country	22.0 (0.8)	2.3	16.8 (1.2)	22.0 (1.7)	22.0 (0.8)	22.0 (0.8)	22.2 (0.9)
Father aged < 22 years at baseline	2.5 (0.3)	1.6	1.1 (0.3)	1.7 (0.7)	2.8 (0.3)	2.7 (0.3)	2.8 (0.4)
Father's first reported education < diploma	70.8 (0.9)	2.7	62.6 (1.6)	69.7 (1.6)	71.1 (0.9)	71.1 (0.9)	71.1 (0.9)
Father's first reported occupation < professional/managerial	60.6 (1.0)	1.6	53.1 (1.6)	59.5 (1.8)	60.8 (1.0)	60.7 (1.0)	60.9 (1.0)
Mother aged < 22 years at baseline	7.3 (0.5)	0.1	3.4 (0.6)	6.0 (1.2)	7.3 (0.5)	7.3 (0.5)	7.4 (0.5)
Mother's first reported education < diploma	76.1 (0.9)	0.9	69.5 (1.5)	74.9 (1.5)	76.2 (0.9)	76.2 (0.9)	76.3 (0.9)
Mother's first reported occupation < professional/managerial	73.9 (0.9)	1.9	67.2 (1.5)	73.5 (1.5)	74.2 (0.9)	74.2 (0.9)	74.2 (0.9)

CCA: complete case analysis (complete on all questions about adverse childhood experiences); IPW: inverse probability-weighting using only baseline auxiliary variables (cohort sex, mother/father educations < diploma, mother/father occupations class < professional or managerial, mother/father aged < 22 years, birthweight < 3rd percentile, premature); MI (baseline): multiple imputation by chained equations using only baseline auxiliary variables; MI (full): multiple imputation by chained equations using all explicitly measured auxiliary variables; RMI: responsiveness-informed multiple imputation by chained equations using all explicitly measured auxiliary variables plus measures of participant responsiveness (refer to Statistical Analysis for further explanation); SE: standard error.

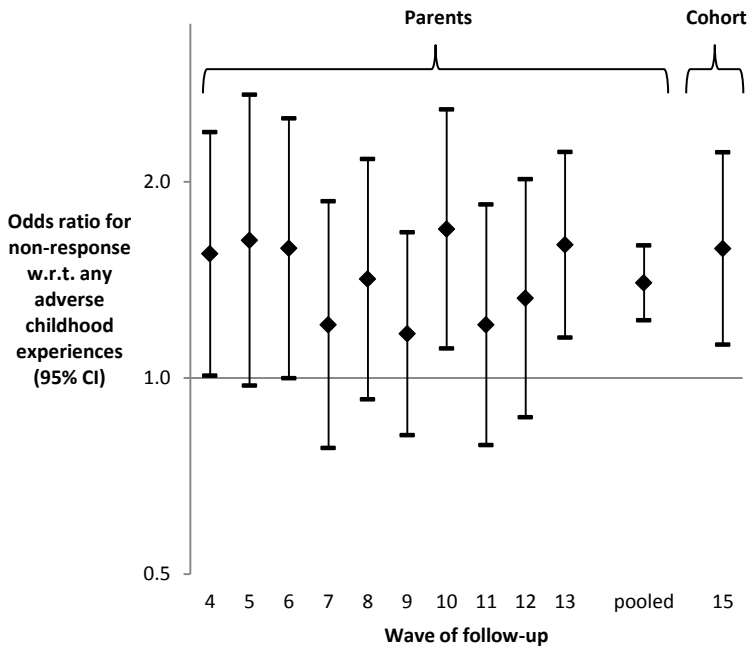


Figure 3. Adverse childhood experiences and non-response by parents and cohort members in waves prior and subsequent to self-report

Adverse childhood experiences were recorded by retrospective self-report in Wave 14. Figure shows point estimates and 95% confidence intervals of odds ratios for non-response at other time points with respect to any adverse childhood experiences. No information was available for non-respondents at baseline (Wave 1). In Waves 2 and 3 only a subset of the cohort was sampled and non-response could not be differentiated from exclusion. The pooled estimate shows non-response by parents across Waves 4–13, estimated by binomial regression with a logit link. Estimates based on complete case analysis.

Table 4. Prevalence of adverse childhood experiences, by method of estimation

Indicator of adverse childhood experiences	Prevalence by method, % (SE)				
	CCA	IPW	MI (baseline)	MI (full)	RMI
Missingness assumption (plausibility)[a]	*MCAR (least)*	*MAR (less)*	*MAR (less)*	*MAR (more)*	*MAR (most)*
Child abuse and neglect					
Emotional abuse (somewhat true)	13.5 (1.1)	13.8 (1.3)	15.6 (1.8)	17.1 (1.4)	18.9 (2.2)
Emotional abuse (very true)	3.2 (0.6)	3.2 (0.6)	4.3 (0.9)	4.9 (1.0)	6.2 (1.6)
Neglect	2.9 (0.5)	3.0 (0.6)	4.4 (1.1)	5.9 (0.9)	7.9 (1.4)
Physical abuse	5.9 (0.8)	6.5 (1.0)	7.8 (1.2)	8.2 (1.1)	9.6 (1.8)
Sexual abuse	5.6 (0.8)	6.2 (0.9)	7.9 (1.1)	10.0 (1.5)	11.0 (1.8)
Witnessed domestic violence	4.4 (0.7)	4.5 (0.8)	6.4 (0.9)	7.9 (1.1)	8.9 (1.4)
Any child abuse or neglect	23.9 (1.4)	24.9 (1.6)	30.6 (2.2)	33.3 (1.8)	37.2 (1.9)
Any child abuse or neglect (emotional = very true)	16.1 (1.2)	16.7 (1.4)	21.4 (1.8)	24.8 (1.7)	28.6 (2.1)
Multiple maltreatment	8.1 (0.9)	8.5 (1.1)	10.6 (1.5)	13.6 (1.1)	16.5 (1.9)
Multiple maltreatment (emotional = very true)	4.0 (0.6)	4.7 (0.8)	6.3 (1.0)	8.0 (1.0)	10.2 (1.8)
Parental mental health					
Parental mental illness	6.6 (0.8)	6.5 (0.9)	7.8 (1.2)	8.8 (1.1)	11.2 (1.5)
Parental substance use problems	4.9 (0.7)	4.9 (0.8)	6.1 (1.1)	6.5 (1.0)	8.1 (1.3)
Number of adverse childhood experiences					
Any	30.0 (1.5)	30.4 (1.7)	36.9 (2.2)	39.7 (1.8)	44.0 (1.9)
1	18.9 (1.3)	19.0 (1.5)	22.7 (1.3)	22.3 (1.3)	22.8 (1.5)
2	7.0 (0.8)	7.1 (1.0)	8.6 (1.0)	9.7 (0.9)	11.2 (1.0)
3	2.7 (0.5)	2.7 (0.6)	3.4 (0.6)	4.5 (0.6)	5.7 (0.8)
4	1.0 (0.3)	1.1 (0.4)	1.4 (0.4)	2.0 (0.5)	2.8 (0.6)
5	0.4 (0.2)	0.6 (0.3)	0.9 (0.5)	1.2 (0.5)	1.6 (0.6)

[a]We propose that the missingness assumption implied by each analysis becomes more plausible from left to right, i.e. bias decreased because of the addition of auxiliary variables that improved imputation of adverse childhood experiences; CCA: complete case analysis; IPW: inverse probability-weighting using only baseline characteristics as auxiliary variables (cohort sex, mother/father educations < diploma, mother/father occupations class < professional or managerial, mother/father aged < 22 years, birthweight < 3rd percentile, premature); MAR: missing at random (given included auxiliary variables); MCAR: missing completely at random; MI (baseline): multiple imputation by chained equations using only baseline characteristics as auxiliary variables; MI (full): multiple imputation by chained equations using all explicitly measured auxiliary variables but excluding measures of responsiveness; RMI: responsiveness-informed multiple imputation by chained equations (refer to Statistical Analysis for further explanation); SE: analytic standard error, combined using Rubin's rules for multiple imputation.

Prevalence estimates for adverse childhood experiences are summarised in Table 4, by method of estimation. Weighting by baseline variables made no substantial or significant differences to prevalence estimates. Compared with complete case analysis, multiple imputation using only baseline measures increased the estimated prevalence of any adverse childhood experiences from 30.0% to 36.9%. Inclusion of explicitly-measured auxiliary variables increased it again to 39.7%), with small gains or losses to efficiency depending on the variable. Adding indicators of parent and cohort responsiveness to the multiple imputation procedure increased it further still, to 44.0%, with small-to-moderate losses of efficiency in all cases.

Relative increases in individual adverse childhood experiences ranged from 39.7% (emotional abuse) to 173.6% (neglect). Using responsiveness-informed multiple imputation, we estimated that experience of these seven adverse childhood experiences in the ATP ranged from 6.2% to 18.9%, with 44.0% experiencing any adverse childhood experience and 37.2% experiencing any child abuse or neglect.

Discussion

This focused examination of missing data in the context of research about adverse childhood experiences demonstrates the high level of susceptibility of surveys in this field to bias arising from non-response and loss to follow-up. Adverse childhood experiences were associated with non-response by both parents and cohort members. We interpret the differences in our estimates across methods as implying that the MAR assumption which underlies even the best commonly employed analyses (e.g. multiple imputation) may be unlikely to hold in many surveys – at least unless a large amount of relevant auxiliary variables are available and utilised, ideally including indicators of participant responsiveness. Simple approaches such as weighting on baseline characteristics alone or multiple imputation using a small set of baseline auxiliary variables appeared insufficient to remove bias with respect to adverse childhood experiences in this cohort and this finding is likely to be generalisable to other studies.

It must be acknowledged that our interpretation of these findings as representing less bias in the analyses with additional relevant auxiliary variables is based on indirect evidence from simulation studies and literature review (Collins et al., 2001; Doidge, 2016; Enders, 2010a, 2010b; Hardt et al., 2012; Seaman & White, 2013). We cannot know the true prevalence of adverse childhood experiences in this or any cohort with missing data; we can only compare our results and interpret the differences in light of other evidence. The study also did not explore the potential influence of misspecification of imputation models, other than through the inclusion of auxiliary variables. It is possible that the way variables were measured and transformed, or the way they interact with each other influenced our results. However, these factors were kept constant across each of the methods, so it is reasonable to interpret the differences between the methods as reflecting only the conditioning of their respective *missing at random* assumptions.

All of the auxiliary variables used in this analysis, including measures of responsiveness, were selected because of their association with adverse childhood experiences and the probability of data about adverse childhood experiences being missing. Auxiliary variables are only valuable if they improve the imputation model; selecting auxiliary variables that are unrelated to the variables of interest, while unlikely to increase bias can be expected to reduce efficiency. Measures of responsiveness may not always be appropriate auxiliary variables. They appear to be valuable, however, at least when the variables of interest are associated with participants being generally less likely to respond across time. This might be expected to be the case for variables associated with things like social marginalisation, disorganisation, cognitive impairment, mental health or geographic mobility.

There was a cost associated with the inclusion of responsiveness measures in this study: the precision of estimates decreased in every case. It may have been that one of the responsiveness measures were responsible for this more than others, and it is likely that more efficient measures of responsiveness could have been derived. Further research is required to identify the most efficient and effective ways to

measure and model responsiveness. Loss of efficiency may be justified by a sufficient reduction in bias, and it appears that this may have been the case in this application. The most precise estimates were made using complete case analysis but these estimates also appear to be the most biased. One interesting observation was that the MI (full) estimates appeared to be both more precise and less biased than the MI (baseline) estimates, despite the additional auxiliary variables. This demonstrates that the inclusion of appropriate auxiliary variables does not always reduce efficiency, and selecting fewer auxiliary variables will not always result in more precise estimates.

One recent study reported no appreciable differences in multiple imputation estimates of adolescent substance use that differed in their inclusion of auxiliary variables (Romaniuk, Patton, & Carlin, 2014). As identified by the authors, this is likely to be because of the sufficiency of the main variables for maximising plausibility of the missing at random assumption; i.e. no additional information was provided by the auxiliary variables. In their case, the main variables were repeated measures of the same things, so this seems reasonable. However, it is rare that repeated measures of adverse childhood experiences are available.

The strength of bias in prevalence estimates using simple approaches in this cohort appeared substantial. Surveys with less missing data are likely to be less distorted but the rate of missing data in the ATP is not unusual for a longitudinal studies of this duration and detail (Dobson et al., 2015; Hawkes & Plewis, 2006; Najman et al., 2015; Straker et al., 2015). Cross-sectional surveys of adults can appear to have less missing data but the potentially greater influence of non-response at the point of recruitment and exclusion from sampling frames must be considered—especially exclusion with respect to the potential outcomes of adverse childhood experiences.

The objectives and scope of a survey are also likely to influence recruitment; when adverse childhood experiences are the focus, people exposed to them may be more or less inclined to participate. This may explain some of the conflicting observations by Haugaard and Emery (1989) and Edwards et al. (2001), who reported some positive correlations between child sexual abuse and response to surveys about adverse childhood experiences. The ATP has a broad scope so is unlikely to be affected by this type of response bias. It is, however, likely to have been influenced by parent-responsiveness at the point of recruitment, which was not addressed in any of our analyses. Given the consistency of the relationship between parent-responsiveness and maltreatment over time, this is likely to have resulted in further downward bias on the prevalence estimates presented.

Research on the correlates of non-response usually focuses on sociodemographic characteristics (e.g. Hawkes & Plewis, 2006; Mostafa & Wiggins, 2015). Weighting on baseline variables is commonly used to correct analyses of adverse childhood experiences and typically indicates little bias (e.g. Fergusson, Boden, & Horwood, 2008). Unlike these approaches, we used a wide range of specifically selected auxiliary variables and supplemented them with measures of participant responsiveness over time to draw comparisons and enhance imputation models. Another opportunity to examine the relationship between adverse childhood experiences and non-response is through linkage of survey data with administrative records. Mills, Alati, Strathearn, and Najman (2014) reported a strong association between child protection notifications and loss to follow-up/non-response in a cohort of children at age 14 (OR = 2.39 calculated from published data) but then did not account for this relationship in their analysis. Linkage of survey and administrative data may be used to estimate probabilities of response with respect to child protection involvement, while at the same time combining two methods for identification of exposure, creating a strong basis for estimating prevalence. While combining these data sources to assess validity of retrospective self-reports has been implemented (Della Femina, Yeager, & Lewis, 1990; Hardt & Rutter, 2004; Smith, Ireland, Thornberry, & Elwyn, 2008), combining them to address missing data is area for future research.

While protocols and incentives can be implemented to maximise retention in population-based cohorts, attrition will always be substantial in the context of active participation under informed

consent (Booker, Harding, & Benzeval, 2011). When data are MNAR, even modest rates of attrition can produce large bias in analyses (Kristman, Manno, & Côté, 2004) but we must remember that the missingness mechanism is a property of the data in the context the analysis—not a property of the data themselves. The selection and utilisation of auxiliary variables is critical to maximising the plausibility of the MAR assumption and minimising bias.

If the auxiliary variables include direct observations of participant responsiveness, then there may be good reason to expect MAR to be plausible with respect to differences in responsiveness that are stable over time. The measurement of participant responsiveness is, however, susceptible to potentially erroneous assumptions about participants with high proportions of missing data. For example, participants lost to follow-up should not be treated as comparable to participants who are not lost but not respondent, as their reasons for non-response are likely to differ (this was the rationale for our measurement of parent response only up until the point of their final response and treatment of cohort non-response in both Waves 14 and 15 as missing). It must also be acknowledged that there are several different forms of possible response behaviour (e.g. non-response, refused response, partial response, inability to respond) and many different reasons that can underlie these. Depending on the variables concerned, the type of data collection and paradata available (information about response time, reminders, etc.), consideration must be given to selecting measures of responsiveness that are most appropriate for a given application. Until further research can provide empirical guidance on this, analysts will have to rely on theory and exploration.

Including direct measures of participant responsiveness in our multiple imputation is likely to have further mitigated bias from missing data but the extent to which this was achieved cannot be assessed without additional information about non-respondents. It is clear though that caution should be exerted when interpreting any quantitative analysis of adverse childhood experiences but particularly in the context of substantial missing data and sampling designs that exclude marginalised populations. Sophisticated analyses based on MAR are always warranted and supplementary information should be sought wherever possible, whether it be from observed responsiveness in longitudinal settings, additional follow-up of non-respondents or linking to administrative data. The potential bias is not trivial and could have important implications for estimating the burden of adverse childhood experiences and responding with appropriate policy and services.

Finally, the utilisation of responsiveness measures as auxiliary variables in multiple imputation or inverse probability weighting appears to hold significant potential for influencing the results of certain analyses in longitudinal studies. We propose that the differences observed in this case reflect a further reduction in bias, which is supported by the findings of simulation study that was designed to mimic this application (Doidge, 2016). This is, however, a relatively unexplored extension of established techniques and there may well be limitations and pitfalls that have not been acknowledged or addressed. We strongly recommend that further simulation research be conducted and that applications of these methods be interpreted with caution while continuing to be explored in other longitudinal studies.

Acknowledgements

The ATP is located at the Royal Children's Hospital in Melbourne and is a collaboration between Deakin University, the University of Melbourne, the Australian Institute of Family Studies, the University of New South Wales, the University of Otago (NZ), and the Royal Children's Hospital; further information available at www.aifs.gov.au/atp. Funding for this analysis was supported by a PhD scholarship from the University of South Australia, and the South Australian Health Economics Collaborative (funded by the South Australian Department of Health). The views expressed in this paper are those of the authors and may not reflect those of their organisational affiliations, nor of other collaborating individuals or organisations. We acknowledge all collaborators who have contributed to the Australian Temperament Project, especially Leah Bromfield, who was involved with the original design of survey questions concerning adverse childhood experiences, and Professors Ann Sanson, Margot Prior, and Frank Oberklaid, and Ms Diana Smart. We would also like to sincerely thank the participating families for their time and invaluable contribution to the study.

References

Bartlett, J. W., Carpenter, J. R., Tilling, K., & Vansteelandt, S. (2014). Improving upon the efficiency of complete case analysis when covariates are mnar. *Biostatistics, 15*(4), 719-730. https://doi.org/10.1093/biostatistics/kxu023

Besharov, D. J. (1981). Toward better research on child abuse and neglect: Making definitional issues an explicit methodological concern. *Child Abuse and Neglect, 5*(4), 383-390. https://doi.org/10.1016/0145-2134(81)90048-X

Booker, C. L., Harding, S., & Benzeval, M. (2011). A systematic review of the effect of retention methods in population-based cohort studies. *BMC Public Health, 11*(1), 1-12. https://doi.org/10.1186/1471-2458-11-249

Brown, D. W., Anda, R. F., Tiemeier, H., Felitti, V. J., Edwards, V. J., Croft, J. B., & Giles, W. H. (2009). Adverse childhood experiences and the risk of premature mortality. *American Journal of Preventive Medicine, 37*(5), 389-396. https://doi.org/10.1016/j.amepre.2009.06.021

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119942283

Cole, J. C. (2008). How to deal with missing data: Conceptual overview and details for implementing two modern methods. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 214-239). Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781412995627.d19

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. https://doi.org/10.1037/1082-989X.6.4.330

Della Femina, D., Yeager, C. A., & Lewis, D. O. (1990). Child abuse: Adolescent records vs. Adult recall. *Child Abuse and Neglect, 14*(2), 227-231. https://doi.org/10.1016/0145-2134(90)90033-P

Dobson, A. J., Hockey, R., Brown, W. J., Byles, J. E., Loxton, D. J., McLaughlin, D., Tooth, L. R., & Mishra, G. D. (2015). Cohort profile update: Australian longitudinal study on women's health. *International Journal of Epidemiology*. https://doi.org/10.1093/ije/dyv110

Doidge, J. C. (2016). Responsiveness-informed multiple imputation and inverse probability-weighting in cohort studies with missing data that are non-monotone or not missing at random. *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280216628902

Dube, S. R., Williamson, D. F., Thompson, T., Felitti, V. J., & Anda, R. F. (2004). Assessing the reliability of retrospective reports of adverse childhood experiences among adult hmo members attending a primary care clinic. *Child Abuse and Neglect, 28*(7), 729-737. https://doi.org/10.1016/j.chiabu.2003.08.009

Edwards, V. J., Anda, R. F., Nordenberg, D. F., Felitti, V. J., Williamson, D. F., & Wright, J. A. (2001). Bias assessment for child abuse survey: Factors affecting probability of response to a survey about childhood abuse. *Child Abuse and Neglect, 25*(2), 307-312. https://doi.org/10.1016/S0145-2134(00)00238-6

Enders, C. K. (2010a). Improving the accuracy of maximum likelihood analyses *Applied missing data analysis*. New York, NY: Guilford Publications.

Enders, C. K. (2010b). The imputation phase of multiple imputation *Applied missing data analysis*. New York, NY: Guilford Publications.

Fang, X., Brown, D. S., Florence, C. S., & Mercy, J. A. (2012). The economic burden of child maltreatment in the united states and implications for prevention. *Child Abuse and Neglect, 36*(2), 156-165. https://doi.org/10.1016/j.chiabu.2011.10.006

Fergusson, D. M., Boden, J. M., & Horwood, L. J. (2008). Exposure to childhood sexual and physical abuse and adjustment in early adulthood. *Child Abuse and Neglect, 32*(6), 607-619. https://doi.org/10.1016/j.chiabu.2006.12.018

Fielding, S., Fayers, P. M., & Ramsay, C. R. (2009). Investigating the missing data mechanism in quality of life outcomes: A comparison of approaches. *Health Qual Life Outcomes, 7*, 57. https://doi.org/10.1186/1477-7525-7-57

Gilbert, R., Widom, C. S., Browne, K., Fergusson, D., Webb, E., & Janson, S. (2009). Burden and consequences of child maltreatment in high-income countries. *The Lancet, 373*(9657), 68-81. https://doi.org/10.1016/S0140-6736(08)61706-7

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing x: A warning against including too many in small sample research. *BMC Medical Research Methodology, 12*(1), 1-13. https://doi.org/10.1186/1471-2288-12-184

Hardt, J., & Rutter, M. (2004). Validity of adult retrospective reports of adverse childhood experiences: Review of the evidence. *Journal of Child Psychology and Psychiatry, 45*(2), 260-273. https://doi.org/10.1111/j.1469-7610.2004.00218.x

Haugaard, J. J., & Emery, R. E. (1989). Methodological issues in child sexual abuse research. *Child Abuse and Neglect, 13*(1), 89-100. https://doi.org/10.1016/0145-2134(89)90032-X

Hawkes, D., & Plewis, I. (2006). Modelling non-response in the national child development study. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(3), 479-491. https://doi.org/10.1111/j.1467-985X.2006.00401.x

Herman, D. B., Susser, E. S., Struening, E. L., & Link, B. L. (1997). Adverse childhood experiences: Are they risk factors for adult homelessness? *Am J Public Health., 87*(2), 249-255. https://doi.org/10.2105/AJPH.87.2.249

Kristman, V., Manno, M., & Côté, P. (2004). Loss to follow-up in cohort studies: How much is too much? *European Journal of Epidemiology, 19*(8), 751-760. https://doi.org/10.1023/B:EJEP.0000036568.02655.f8

Mills, R., Alati, R., Strathearn, L., & Najman, J. M. (2014). Alcohol and tobacco use among maltreated and non-maltreated adolescents in a birth cohort. *Addiction, 109*(4), 672-680. https://doi.org/10.1111/add.12447

Mostafa, T., & Wiggins, R. (2015). The impact of attrition and non-response in birth cohort studies: A need to incorporate missingness strategies. *Longitudinal and Life Course Studies, 6*(2), 16. https://doi.org/10.14301/llcs.v6i2.312

Najman, J. M., Alati, R., Bor, W., Clavarino, A., Mamun, A., McGrath, J. J., McIntyre, D., O'Callaghan, M., Scott, J., Shuttlewood, G., Williams, G. M., & Wray, N. (2015). Cohort profile update: The mater-university of queensland study of pregnancy (musp). *International Journal of Epidemiology, 44*(1), 78-78f. https://doi.org/10.1093/ije/dyu234

Norman, R. E., Byambaa, M., De, R., Butchart, A., Scott, J., & Vos, T. (2012). The long-term health consequences of child physical abuse, emotional abuse, and neglect: A systematic review and meta-analysis. *PLoS Medicine, 9*(11). https://doi.org/10.1371/journal.pmed.1001349

Prior, M., Sanson, A., Smart, D., & Oberklaid, F. (2000). *Pathways from infancy to adolescence: Australian temperament project 1983-2000*. Melbourne, Australia: Australian Institute of Family Studies.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.

Romaniuk, H., Patton, G. C., & Carlin, J. B. (2014). Multiple imputation in a longitudinal cohort study: A case study of sensitivity to imputation methods. *American Journal of Epidemiology, 180*(9), 920-932. https://doi.org/10.1093/aje/kwu224

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal, 4*(3), 227-241.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons. https://doi.org/10.1002/9780470316696

Sanson, A., & Oberklaid, F. (1985). Normative data on temperament in australian infants. *Australian Journal of Psychology, 37*(2), 185-195. https://doi.org/10.1080/00049538508256397

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Seaman, S. R., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by "missing at random"? , 257-268.

Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research, 22*(3), 278-295. https://doi.org/10.1177/0962280210395740

Smith, C. A., Ireland, T. O., Thornberry, T. P., & Elwyn, L. (2008). Childhood maltreatment and antisocial behavior: Comparison of self-reported and substantiated maltreatment. *American Journal of Orthopsychiatry, 78*(2), 173-186. https://doi.org/10.1037/0002-9432.78.2.173

Straker, L. M., Hall, G. L., Mountain, J., Howie, E. K., White, E., McArdle, N., & Eastwood, P. R. (2015). Rationale, design and methods for the 22 year follow-up of the western australian pregnancy cohort (raine) study. *BMC Public Health, 15*, 663. https://doi.org/10.1186/s12889-015-1944-6

Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research, 49*(5), 443-459. https://doi.org/10.1080/00273171.2014.931799

Vassallo, S., & Sanson, A. (Eds.). (2013). *The australian temperament project: The first 30 years*. Melbourne: The Australian Institute of Family Studies. https://doi.org/10.1037/e567282013-002

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399. https://doi.org/10.1002/sim.4067

Widom, C. S., & Maxfield, M. G. (2001). *An update on the "cycle of violence"*. Research in Brief, Washington DC: National Institute of Justice.

Wu, S. S., Ma, C.-X., Carter, R. L., Ariet, M., Feaver, E. A., Resnick, M. B., & Roth, J. (2004). Risk factors for infant maltreatment: A population-based study. *Child Abuse and Neglect, 28*(12), 1253-1264. https://doi.org/10.1016/j.chiabu.2004.07.005

Wyatt, G. E., & Peters, S. D. (1986). Methodological considerations in research on the prevalence of child sexual abuse. *Child Abuse and Neglect, 10*(2), 241-251. https://doi.org/10.1016/0145-2134(86)90085-2

An integrated and collaborative approach to developing and scripting questionnaires for longitudinal cohort studies and surveys: experience in Life Study

Suzanne Walton UCL GOS Institute of Child Health, UK
Stelios Alexandrakis UCL GOS Institute of Child Health, UK
Nicholas Gilby Ipsos MORI, UK
Nicola Firman UCL GOS Institute of Child Health, UK
Gareth Williams Ipsos MORI, UK
Duncan Peskett Ipsos MORI, UK
Peter Elias University of Warwick, UK
Carol Dezateux UCL GOS Institute of Child Health, UK
c.dezateux@ucl.ac.uk

(Received December 2016 Revised July 2017) http://dx.doi.org/10.14301/llcs.v8i4.434

Abstract

Efficient development of questionnaires for longitudinal surveys and cohort studies as computer-assisted survey instruments usually entails close collaboration between scientific and fieldwork teams. We describe a system based on the use of a Structured Query Language (SQL) database established to maximise efficiency, minimise error and ensure clear communication of requirements across teams for 'Life Study', a UK-wide cohort study designed to recruit mothers, their babies, partners and non-resident fathers, with whom further contacts were planned at the outset. The use of the SQL database enabled construction and integration of different elements of the study, initially through creating a master copy of each variable. This supported swift and accurate creation of a range of outputs enabling, for example, review and approval of successive drafts and final specifications of questionnaires, efficient implementation of changes to variables, re-use of metadata specified at the outset, reduction of ambiguities for survey programmers, and efficient and accurate automation of questionnaire scripting. The SQL database was also used to generate the syntax to transform pilot data into formats specified for data archiving and for associated publication quality questionnaires. This innovative use of an SQL database for questionnaire development and scripting, and subsequent data processing and documentation, highlights the value of this approach in improving the quality and efficiency of longitudinal surveys.

Keywords

Questionnaire design, questionnaire programming, data processing, data collection, data documentation, metadata, longitudinal studies, cohort studies, surveys and questionnaires, Structured Query Language (SQL) database

Introduction

Large-scale birth cohorts and longitudinal surveys comprise a key data resource, enabling interdisciplinary and life course research. Innovations in study design and timing of contacts can add complexity to the task of designing and programming questionnaires and may require an integrated approach for multiple respondents and contacts, especially for family-based or household designs. The aim of this research note is to report our experience of using an SQL database to develop and script questionnaires, re-using metadata specified at the outset in the context of a complex interdisciplinary cohort study for those involved in the design and administration of similar longitudinal studies. We present here an overview of methodology and advantages, rather than a technical guide.

'Life Study' comprised an innovative design of two integrated samples – a 'Pregnancy Component', purposively sampled from a small number of areas, and a nationally representative 'Birth Component', with multiple respondents and contacts in each (Dezateux, Knowles, et al., 2016; Goldstein, Sera, Elias, & Dezateux, 2017) (Table 1).

Table 1. Summary of core 'Life Study' protocol: planned contacts with participants

Participant	Pregnancy Component			Birth Component	
	Pregnancy	six months	12 months	six months	12 months
Mother	Yes	Yes	Yes	Yes	Yes*
Resident Father / Partner	Yes	No	No	Yes	No
Non-resident Father / Partner	Some	No	No	No	No

- Telephone or web-based interview

Questionnaires were developed with input from clinicians, population, social and biomedical scientists, relevant stakeholders and experts, and in consultation with UK research and policy communities. The Life Study Scientific Steering Committee was responsible for the final selection of measures and instruments.

The core protocol required design of seven questionnaires with two further questionnaires developed as a result of two funded enhancements on the maternal microbiome and on non-resident fathers (Bailey et al., 2015; Ipsos MORI, 2016). The integrated design necessitated appreciable overlap in, and harmonisation of, the questions included so that data collected in both components could be presented to the user as a single dataset. In addition, the scientific specification and operationalisation of a number of surveys had to proceed iteratively and within very tight time frames.

The UK Data Archive describes a "research data lifecycle" (UK Data Archive, 2017) comprising creating, processing, analysing, preserving, accessing and re-using data. Similarly, Banks, Calderwood, Lynn and Angel (2009) described a data production line comprising scientific direction, study design, instrument design, instrument realisation, data collection, data processing and data documentation in a report from the Survey Resources Network aimed at identifying improvements to efficiency and quality in data collection, management and processing of longitudinal surveys. This report highlighted the desirability of an efficient data production line at each stage and avoidance of duplication of work, while noting that in most longitudinal surveys the different elements of the data production line are carried out by separate organisations. Widespread practice is to specify questionnaire instruments using word processing software, typically Microsoft Word. Banks et al. (2009) state that "this can lead to ambiguities, in particular with respect to complex routing or question structures. This approach also does not facilitate the capture of structured

metadata that can then be used at other stages of the data production process." (p. 6). The authors highlighted the central role of effective capture and reuse of metadata in quality assurance, the associated efficiency gains and need for questionnaire specification tools which capture metadata in a way that is both minimally ambiguous and maximally re-usable by downstream processes. Recognising that manual transfer of metadata between applications is time-consuming and error prone, they concluded that "capture of this metadata in a machine-parseable form is thus an important aspect of gaining control over the survey process as a whole and increasing the efficiency and quality of it" (Banks et al. 2009, p. 15).

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences (DDI Alliance, 2017). While a number of DDI tools were available at the time 'Life Study' was being developed, they were not mature or straightforward to use, and after discussion with DDI developers none were considered to fully meet our requirements, especially given the complexities of developing long questionnaires with inevitable multiple revisions, for multiple sweeps and respondents simultaneously. We elected to create a database to manage development of the 'Life Study' surveys which could be used at various points in the data lifecycle to overcome the limitations of more conventional approaches to scripting highlighted by Banks et al. (Banks et al., 2009). This approach built on existing skills and expertise within the team – allowing questionnaires to be developed as rapidly and efficiently as possible – and because the metadata was collected in a structured format, this enabled future interoperability with DDI-compatible tools and metadata.

Development of the questionnaires

Requirements

Questionnaires were developed on a modular or topic basis (Table 2). Additional survey elements supported collection of consent, biosamples, and a range of measurements, assessments and observations. Each questionnaire was planned to take 30-55 minutes and included interviewer administered (Computer Assisted Personal Interviewing - CAPI) and self-complete questions (Computer Assisted Self Interviewing - CASI) with complex routing. As there were over 1,840 questions/variables in the dataset – many to be used more than once – we required a system whereby a single, master copy of each variable could be specified, updated and applied to each questionnaire, with the ability to export these in different formats.

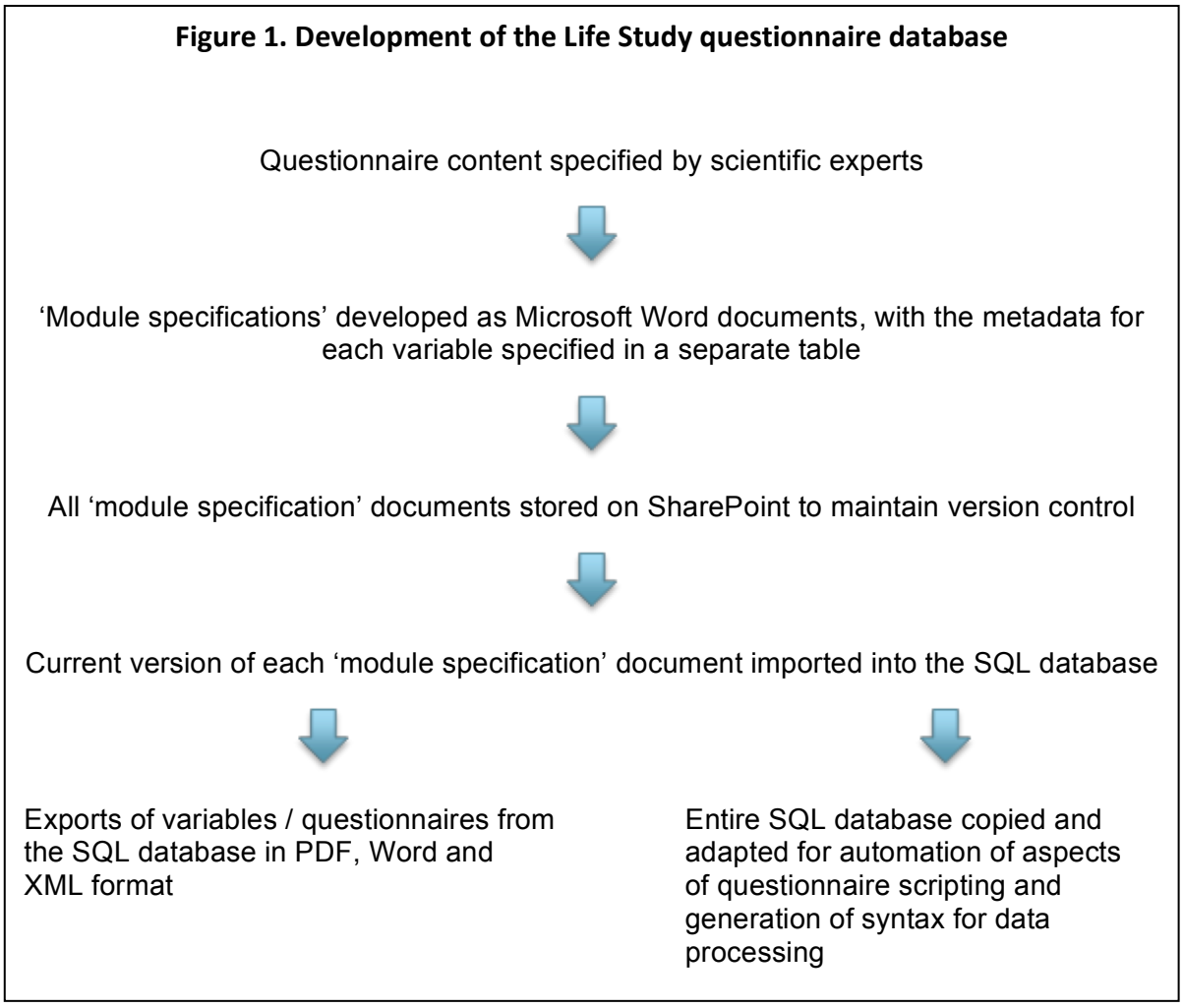
Table 2. Life study modules

1. Demographics
2. Identity
3. Parental and Family Health
4. Parental Mental Health
5. Parental Behaviour and Lifestyle
6. Parental Education
7. Parental Employment
8. Financial Situation
9. Pregnancy and Birth
10. Child Health
11. Child Development
12. Child Sleeping and Crying
13. Diet and Nutrition
14. Infections and Immunity
15. Childcare
16. Parenting
17. Family Relationships
18. Social Networks and Support
19. Housing
20. Neighbourhood
21. Environment
22. Partner Proxy
23. Consents
24. Interviewer Observations

Development of the questionnaire database

The 'Life Study' questionnaire database was developed as a client-server system comprising a database server and desktop client, using Microsoft SQL Server – a relational database using Structured Query Language (SQL). The customised application (client) was built using the Microsoft.NET framework to interact with the SQL database. The installation process entailed unzipping the client on a desktop computer with network/security permissions to connect to the SQL Server database. Questionnaires were developed and formatted so they could be imported into the SQL database (Figure 1).

Figure 1. Development of the Life Study questionnaire database



Questionnaire specification and metadata

We specified questionnaire content in Microsoft Word for importing into an SQL database – as changes could be tracked, there were options for formatting, while spelling errors could be easily identified and corrected. We created one Microsoft Word document "module specification" for each module and, within these documents, developed a template/table containing the relevant information and metadata for each variable (See Appendix 1 - Explanation of the metadata for each variable). The tables were ordered in the sequence asked in the questionnaires. Successful import into the SQL database depended on maintaining – unaltered – the given formatting and headings of the tables within the Word document. The module specifications were populated with relevant metadata (Appendix 2 – Examples of variable tables). Fields within the table included: question name, variable label, question type (e.g. choice, date, number), and question text. The tabular structure ensured completion of all fields and a field indicating whether a question had been dropped enabled questions to be reinstated easily, if required.

Each table in the module specification documents had a field called 'universe'. These universe statements specified which questionnaires each variable was to be included in, and under what circumstances, and so were key to the functioning of the systems developed. The universe statements were based on Boolean logic and contained information on respondent types, study components, timings (sweeps) and modes of interview as well as, where relevant, conditioning based on responses to other variables. They were written in a standardised format so that the SQL database software could parse (automatically read and process) the information and export questionnaire documents and metadata accordingly. Specification of module order in conjunction with the order of tables/variables within a module and

the universe statements stipulated the routing for each questionnaire.

Version control

A single master document for each module was stored on Microsoft SharePoint. This online document management platform enabled retention of all earlier versions and version control and reduced errors arising from multiple documents and multiple users.

Importing modules into the SQL database

The module specification documents were parsed as entire documents and imported into the Microsoft SQL Server database using the desktop client. The process of importing the modules was designed to identify errors in the universe statements or other problems arising from, for example, the merging of tables. Modules already in the database were deleted before importing a newer version. As module specifications were imported, a major version was created in SharePoint, date and time were automatically recorded and a comment added to indicate successful upload.

Database exports

The modules could be exported from the database in Microsoft Word and Excel formats, and also in Extensible Markup Language (XML). Export of modules was possible either individually – or as multiple modules simultaneously – using the desktop client in a number of pre-specified layouts based on component, respondent type, sweep and mode. This allowed the questionnaires to be displayed in a variety of formats for different audiences. These summarised variable name, question text, interviewer or interviewee instructions, response types and options, and the universe statement. A more detailed format was provided for computer programmers, called 'programmer exports', which included information needed to script computerised instruments, including whether to allow "don't know" and "prefer not to answer" responses, as well as soft and hard checks.

'Life Study' staff had direct access to the Microsoft SQL Server Management Studio software, a database management user interface that was also located on the same server. This allowed them to copy the entire SQL database or individual tables and convert these into Microsoft Excel format.

Questionnaire scripting

Survey programmers at Ipsos MORI (the fieldwork partner for the 'Birth Component' pilot) were sent a copy of the entire SQL database plus its exports to enable them to computerise questionnaires using SPSS Dimensions software. They produced additional exports directly from the database to automate aspects of questionnaire programming. These allowed automatic extraction of variable names, question text, interviewer instructions, response options, and numeric ranges, thereby avoiding manual transfer of this information. Additionally, instructions such as "include timestamp here" were specified consistently throughout the module specifications, allowing the programmers to search for relevant terms and automate additional aspects of programming.

Once a computerised questionnaire had been created, Ipsos MORI checked and verified the programming using the 'programmer exports' provided by the 'Life Study' team. These documents contained all the relevant information that needed to be checked. An Excel log of any ambiguities and errors identified on checking was then passed to the 'Life Study' team, who enacted any changes required to the module specifications and sent an updated version of the SQL database back to Ipsos MORI.

Processing of raw datasets

On conclusion of the 'Birth Component' pilot, Ipsos MORI transformed raw data, held in SPSS format, using tables from the SQL database. This required addition of variable labels, "don't know" and "prefer not to answer" codes, and variable name prefixes compliant with 'Life Study' variable naming specifications. Tables from the SQL database were used to create lists of variable names, labels and prefixes by module, which were copied into Microsoft Excel where formulae were used to generate SPSS syntax. This enabled SPSS syntax files comprising thousands of lines to be generated rapidly, and updated should errors be discovered. The addition of variable name prefixes and suffixes within the datasets specified the respondent and sweep to enable the provenance of variables to be understood, and – together with the structured approach to variable naming – enabled datasets to be easily re-shaped for longitudinal analyses.

Data documentation

In October 2015, when funding for 'Life Study' was withdrawn (Dezateux, Colson, Brocklehurst, & Elias, 2016), a 'metadata export' from the SQL database was created for the purposes of data documentation. This followed the basic structure of earlier database exports and contained additional information regarding the provenance of each question. From these, questionnaire metadata (Walton et al., 2016a, 2016b, 2016c, 2016d, 2016e, 2016f, 2016g) and documents for archiving 'Birth Component' pilot data with the UK Data Service (Dezateux, 2016) were rapidly produced.

Discussion

We report here our experience of developing an innovative system based on a SQL database that provided a robust approach to delivering the scientific design of a complex multipurpose cohort study, thus supporting high quality data collection, documentation and review standards and enabling iterative review of the questionnaires by scientific advisory and steering groups to review within the timetable of the study. Strengths of this approach include reduction of ambiguities and errors during questionnaire programming by specifying metadata at the outset as part of the instrument design stage, and time savings resulting from automation of several processes. This approach also facilitated data documentation in Microsoft Word and PDF format, database exports during data processing, and generation of syntax to transform SPSS datasets collected from pilot studies.

This approach is comparable in certain respects to the relational database management system described by Olsen (Olsen, 2012) who, since the late 1980s, has been developing methods to allow for integration of software and hardware solutions to survey and questionnaire design, including approaches to data collection and management in longitudinal surveys. Olsen (2012) demonstrated the effective use of relational database management systems in achieving integration between different stages of a longitudinal survey while avoiding separate questionnaire scripting. To our knowledge, other UK cohort studies have not employed this approach.

We developed an innovative tool based on SQL for capturing metadata in a machine-parseable format for re-use at each stage along the data production line. This is a dynamic system that requires programming expertise to create, but no prior experience of SQL to use. We employed a systems architect with experience of C#, Microsoft .NET framework and SQL to carry out the initial programming work. The 'Life Study' team had no prior experience of computer programming or of working with a SQL database, and remained responsible for all day-to-day interactions with the database and communication with the programmers.

Errors were significantly reduced by having one master version for each variable – and the ability to apply it when relevant – as opposed to specifying each questionnaire separately. Any changes to the master document were implemented consistently each time that question was used. This significantly decreased the time required to enact changes. Ambiguities and instances where information was missing were also significantly reduced for the programming team, as questionnaires were developed in a very structured format and were highly specified.

Exports from the SQL database allowed the modules to be produced swiftly as a professional publication with a standardised format. Each module was created as a separate document – however, it was possible to create entire questionnaires by combining PDF or Word documents with modules in the correct order. Exports were created in a standardised and consistent format for all steering and scientific group meetings and this facilitated these discussions and decisions. These summary formats were also used for ethics applications and for sharing with scientific experts, fieldwork agency staff and interviewers, other 'Life Study' team members and collaborators, and, ultimately, for publication. In creating these metadata documents from the SQL database, the 'Life Study' team had complete control over their content, format and style. The advantages of this approach were that questionnaire metadata were available before a fieldwork partner was even appointed, and were not constrained by the software used to generate the computerised scripts by a specific fieldwork partner. This approach also makes it easier to change fieldwork partner during a longitudinal study (as scripting is more efficient and less labour intensive, thus reducing incumbency advantage) and therefore improves tendering competitions.

Survey programmers at Ipsos MORI initially considered working directly with the XML outputs, but, after discussions, chose to work with the SQL database directly. The SQL database structure was flexible enough to allow the programmers to tailor the export mechanisms to their requirements. After about a week of development time, they were able to produce exports from the SQL database that extracted variable names, question text, interviewer instructions and response options. This reduced questionnaire scripting time as well as typographical and other errors, compared to typical practice where all aspects of the computerised script are programmed manually. Additionally, the more laborious aspects of questionnaire programming were removed, freeing up programmer time to concentrate on more complex tasks such as programming of routing.

Use of the programmer exports to check the scripted questionnaires reduced errors in programming, as well as time taken to check and correct programming. Similarly, using the SQL database to create syntax to process the raw SPSS datasets made it much easier to identify and correct errors and significantly reduced the time needed for this task.

In addition to aspects already covered, Ipsos MORI experienced time savings in scripting as a result of, firstly, not having to create survey metadata – whereas conventionally this is created 'de novo' with each study. Secondly, having a consistent structure to all the modules made collating them into one script much easier and, thirdly, there were fewer amendments required after checking. Time savings were also made, in that it was easy to make changes to question texts so that scripting created for one component of the study could be re-used for the other component of the study. The time required in programming the export process from the SQL database was more than recovered and, had the study continued, additional time savings were anticipated at each stage.

We had planned to produce 'Life Study' metadata for researchers to interrogate in an interactive electronic format such as provided by UK Biobank (Biobank UK, 2016) and the Health Survey for England (UK Data Service, 2017), either by using the SQL database directly or via the XML exports. We also anticipate our methodology would enable rapid import of survey metadata into cross-cohort search engines (CLOSER, 2017). 'Life Study' was closed before these could be developed, however exports from the SQL database allowed rapid production of metadata in PDF format for final reporting and data archiving.

In summary, the approach described in this research note enabled a complex survey with multiple sweeps and multiple respondent types to be developed by multidisciplinary teams in an efficient manner. It allowed automation of aspects of computerised survey scripting, saving time and reducing errors, and enabled rapid production of a dataset and questionnaire metadata for researchers to use.

As far as we are aware, 'Life Study' is the first UK cohort study to use a SQL questionnaire database in this way, using metadata specified at the outset at various points along the data production line. While many studies will not have the same complexity as the initial sweeps of 'Life Study', we would encourage others embarking on longitudinal studies, especially those where variables are re-used between sweeps, to consider adopting this novel approach.

Acknowledgements

This work was supported by the Economic and Social Research Council [Grant numbers ES/J007501/1, ES/L002507/1, ES/L002353/1, ES/L012871/1, ES/N007549/1].

The 'Life Study' team would like to thank Daniel Wallis and Dr John Godfrey from Tessella Technology and Consulting (www.tessella.com), and Robert Ireland for their assistance with programming developments and modifications of the SQL database after the initial development and programming work.

We also acknowledge the assistance of NatCen Social Research, the fieldwork partner appointed to carry out the 'Life Study' 'Pregnancy Component' pilot, for assistance with initial development of the module specification templates.

References

Bailey, S., Townsend, C., Rodgers, A., Dent, H., Mallet, C., Tsaliki, E., . . . Field, N. (2015). 15. Acceptability of collection of multiple bio-samples to birth cohort participants: implications for large studies. In Abstracts of the UK Molecular Epidemiology Group Winter Meeting on Metabonomics in Molecular Epidemiology. Imperial College London, London, UK. November 28, 2014. *Mutagenesis, 30*(3), 459-466. https://doi.org/10.1093/mutage/gev012

Banks, R., Calderwood, L., Lynn, P., & Angel, G. (2009). *A Feasibility Study to Investigate Integrated Survey Data Collection, Fieldwork Management and Survey Data Processing Systems for Longitudinal Studies*. Retrieved from http://www.researchcatalogue.esrc.ac.uk/grants/RES-234-25-0003/outputs/read/76474946-fba7-4a57-bfa6-b05a2414f661

Biobank UK. (2016). UK Biobank - Data Showcase. Retrieved from http://biobank.ctsu.ox.ac.uk/crystal/

CLOSER. (2017). About CLOSER Discovery. Retrieved from http://www.closer.ac.uk/data-resources/closer-search-platform/

DDI Alliance. (2017). Document, Discover and Interoperate. Retrieved from https://www.ddialliance.org/

Dezateux, C. (2016). *Life Study: Birth Component Pilot Study Sample, 2015: Secure Access [data collection]. UK Data Service. SN: 8072*. http://dx.doi.org/10.5255/UKDA-SN-8072-1

Dezateux, C., Colson, D., Brocklehurst, P., & Elias, P. (2016). *Life after Life Study: Report of a Scientific Meeting held at The Royal College of Physicians 14th January 2016*. https://doi.org/10.14324/000.rp.1485681

Dezateux, C., Knowles, R., Brocklehurst, P., Elias, P., Burgess, S., Colson, D., . . . Walton, S. (2016). *Life Study Scientific Protocol*. https://doi.org/10.14324/000.rp.1485668

Goldstein, H., Sera, F., Elias, P., & Dezateux, C. (2017). Integrating area-based and national samples in birth cohort studies: the case of Life Study. *Longitudinal and Life Course Studies 8*(3). http://dx.doi.org/10.14301/llcs.v8i3.439

Ipsos MORI. (2016). *Life Study: Qualitative work with lone mothers: Exploring options for contacting non-resident fathers*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485696/

Olsen, R. (2012). Infrastructure for Survey Data Processing in Urban and Planning Studies. In Carlos Nunes Silva (Ed.), *Online Research Methods in Urban and Planning Studies: Design and Outcomes* (pp. 17-36). https://doi.org/10.4018/978-1-4666-0074-4.ch002

UK Data Archive. (2017). Create and manage data - research data lifecycle. Retrieved from http://www.data-archive.ac.uk/create-manage/life-cycle

UK Data Service. (2017). Health Survey for England 2014 Retrieved from http://nesstar.ukdataservice.ac.uk/webview/index.jsp?v=2&mode=documentation&submode=abstract&study=http://nesstar.ukdataservice.ac.uk:80/obj/fStudy/7919&top=yes

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016a). *Life Study Birth Component: Mother questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485694/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016b). *Life Study Birth Component: Non-resident Father questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485692/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016c). *Life Study Birth Component: Partner questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485693/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016d). *Life Study Pregnancy Component: 6 month visit, Mother questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485677/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016e). *Life Study Pregnancy Component: 12 month visit Mother questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485708/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016f). *Life Study Pregnancy Component: Mother pregnancy visit questionnaire*. Retrieved from UCL Discovery, London UK: http://discovery.ucl.ac.uk/1485674/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016g). *Life Study Pregnancy Component: Partner questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485676/

Appendix 1 – Explanation of the metadata for each variable

Field	Explanation
Question-Name	A unique variable name, consistent with the Life Study variable naming document
Variable-Label	Variable label to appear in the Life Study dataset
Allow-DK/REF	Indicates whether responses of "Don't know" (DK) and "Prefer not to answer" (REF) were permitted Yes = allow both Allow refusal only = allow "Prefer not to answer" but not "Don't know" No = don't allow either
Type	Choice: multiple = respondent can choose multiple response options Choice: single = respondent can only choose only one response option Control: = administrative e.g. introduction text. Open: XX = open text with fields limited to XX number of characters Number: A, B, C..D = number field, where A is the number of decimal places allowed, B is the maximum number of characters that can be entered, C..D is the range (lower and upper values) permitted Date: dd/mm/yyyy = date with day, month and year Date: mm/yyyy = month and year Date: yyyy = year
Source	Where the question originated from e.g. name of cohort study where used previously
Text	Wording of the question text to appear on the screen ^[xxxxxx] = text fill
TextBC	Text for Birth Component where this differs from the Pregnancy Component.
Interviewer-Instruction	UPPER-CASE TEXT signifies further instruction or information for interviewers in the CAPI sections USE CARD XX = there is a showcard for the interviewer to refer to. XX is the reference name/number of that card
Programmer-Instruction	Text fill instruction, including where to get the text fill information from and in which sweeps to implement the text fill.

	Loops - some variables are to be asked multiple times. The maximum number of loops is specified. Exclusive codes - some 'Choice: multiple' questions will have a response option which is exclusive and cannot be selected alongside other options. Include timestamp here = variable where time stamp required
Showcard	Yes = Showcard exists if the question is administered as CAPI. If there is no showcard, this field will remain empty If the showcard is different from the Options (then what is to appear on the showcard is given in this field).
Options	Table with response options
Help	Text for a help screen
Universe	Universe statement for the Pregnancy Component. RespType = 1 or 2 1 = Mother 2 = Partner Sweep = 1, 2 or 3 1 = Pregnancy Component pregnancy visit 2 = Pregnancy Component 6 month visit 3 = Pregnancy Component 12 month visit ModeType = 1, 2, 3 or 4 1 = CAPI (Computer-assisted personal interviewing) 2 = CASI (Computer-assisted self interviewing) 3 = PAPI (Pen and paper interview) 4 = Pre-visit paper questionnaire <> Is used to mean is not equal to @ sweep x means information to be fed forward from another sweep
UniverseBC6	Universe Statement for the Birth Component, 6 month visit. RespType = 1, 2 or 4 1 = Mother 2 = Partner 4 = Non-resident Father Sweep = 4 4 = Birth Component 6 month visit ModeType = 1, 2 or 4 1 = CAPI (Computer-assisted personal interviewing) 2 = CASI (Computer-assisted self interviewing) 4 = Pre-visit or post-visit paper questionnaire
UniverseBC12	Universe Statement for the Birth Component, 12 month visit. RespType = 1

	1 = Mother Sweep = 5 5 = Birth Component 12 month visit ModeType = 4, 5 or 6 4 = Pre-visit or post-visit paper questionnaire 5 = CATI (Computer-assisted telephone interviewing) 6 = WASI (Web-assisted self interviewing) <> Is used to mean is not equal to @ sweep x means information to be fed forward from another sweep
Soft-Check	Asks the participant to review their answer, as it exceeds the expected range.
Hard-Check	Indicates that an answer is not feasible. The answer needs to be changed before proceeding to the next question.
Flags	Who: This variable name prefix denotes who the variable refers to: m = mother p = partner c = child s = sibling g = grandparent h = household n = not applicable What: This variable name prefix denotes the type of data the variable is collecting: o = original d = derived y = physical measurement c = observations and assessments p = proxy report q = pre-complete question b = biological sample a = administrative data Multiple: yes This flag appears for questions which need to be asked separately for each child of a multiple birth.
Drop	Yes = This variable has been dropped and is no longer included in any of the questionnaires

Appendix 2 – Examples of variable tables within the module specification templates

Question-Name BabyDoB
Variable-Label Date of birth of cohort baby
Allow-DK/REF No
Type Date: DD/MM/YYYY
Source Millennium Cohort Study (MCS) First Survey
Text And what is ^[Cohort baby name]'s date of birth?
Interviewer-Instruction
Programmer-Instruction Include timestamp here Loop for each cohort baby as given at given at MultPreg @ sweep 1 or if sweep = 4,5 then loop for each cohort baby as given at NumBaby Textfill ^[Cohort baby name] from BabyName
Showcard
Options
Help

Universe
 	
IF (RespType = 1) // Mother
AND (Sweep = 2) // 6 months
AND (ModeType = 1) // CAPI

UniverseBC6
 	
`IF (RespType = 1) // Mother`
`AND (Sweep = 4) // 6 months`
`AND (ModeType = 1) // CAPI`

Soft-Check
 	
`If NumBaby>1 and babies are born on different dates:`
 	
`"INTERVIEWER. THIS BABY WAS BORN ON A DIFFERENT DAY TO THE PREVIOUS BABY, PLEASE CHECK IF THIS IS CORRECT."`

Hard-Check
 	
`IF date of birth is after date of interview: 'INTERVIEWER: This date is in the future. Please change!'`
 	
`If date of birth is before 01/07/2014: "Answer '^[date entered]' is not in range '01/07/2014 – 31/12/9999'."`

Flags Who: c What: o

Question-Name `ActSing`
Variable-Label `Activities – sing`
Use
Allow-DK/REF `Allow refusal only`
Type `Choice: single`
Source `Adapted from ALSPAC - Children of the children of the 90's (COCO90's)`

Text
 	
`How often do you do these activities with ^[Cohort baby name]…`
 	
`… Sing to ^[him/her]?`

Interviewer-Instruction

Programmer-Instruction

```
Textfill for cohort baby's name and sex
```

Showcard

Options

1	Every day
2	Several times a week - 2 to 6 times
3	Once a week
4	Less than once a week
5	Not at all

Help

Universe

```
IF ((RespType = 1) // Mother
And (ModeType = 2) // CASI
And (Sweep = 2)) // 6 month

OR ((RespType = 1) // Mother
And (ModeType = 4) // pre-visit
And (Sweep = 3)) // 12 month
```

UniverseBC6

```
IF ((RespType = 1 or 2) // Mother or Partner
And (ModeType = 2) // CASI
And (Sweep = 4)) // 6 month

OR ((RespType = 4) // Non-resident partner
And (ModeType = 2) // CASI
And (Sweep = 4) // 6 month
And (FrqSeChd = 1..7)) // Sees cohort baby
```

UniverseBC12

```
IF (RespType = 1) // Mother
And (ModeType = 5 or 6) // CATI or WASI
And (Sweep = 5) // 12 month
```

Soft-Check

Hard-Check

Flags

```
Multiple: yes
Who: m
What: o
```

The biology of inequalities in health: the LIFEPATH project

(Received February 2017 Revised August 2017) http://dx.doi.org/10.14301/llcs.v8i4.448

Paolo Vineis Imperial College London, UK
p.vineis@imperial.ac.uk
Mauricio Avendano-Pabon King's College London, UK
Henrique Barros University of Porto, Porto, Portugal
Marc Chadeau-Hyam Imperial College London, UK
Giuseppe Costa Department of Clinical and Biological Sciences, Turin University Medical School, Italy
Michaela Dijmarescu Imperial College London, UK
Cyrille Delpierre UMR 1027 INSERM – University Paul Sabatier, France
Angelo D'Errico ASL TO3, Italy
Silvia Fraga University of Porto, Portugal
Graham Giles Cancer Council Victoria, Melbourne, Australia
Marcel Goldberg UMS 011 Inserm – UVSQ, France
Marie Zins UMS 011 Inserm – UVSQ, France
Michelle Kelly-Irving UMR 1027 INSERM – University Paul Sabatier, France
Mika Kivimaki University College London, UK
Thierry Lang UMS 011 Inserm - UVSQ, France
Richard Layte Trinity College Dublin, Ireland
Johan P. Mackenbach Erasmus MC, University Medical Center Rotterdam, Netherlands.
Michael Marmot University College London, UK
Cathal McCrory Trinity College Dublin, Ireland
Cristian Carmeli Lausanne University Hospital, Switzerland
Roger L. Milne Cancer Council Victoria, Australia
Peter Muennig Columbia University, US
Wilma Nusselder Erasmus MC, University Medical Center Rotterdam, Netherlands.
Silvia Polidoro Human Genetics Foundation, Italy
Fulvio Ricceri ASL TO3, Italy
Oliver Robinson Imperial College London, UK
Silvia Stringhini Lausanne University Hospital, Switzerland
The LIFEPATH Consortium [i]

Abstract

Socioeconomic differences in health have been consistently observed worldwide. Physical health deteriorates more rapidly with age among men and women with lower socioeconomic status (SES) than among those with higher SES. The biological processes underlying these differences are best understood by adopting a life course approach. In this paper we introduce the pan-European LIFEPATH project which uses multiple cohorts – including biomarker data – to investigate ageing as a phenomenon with two broad stages across life: build-up and decline. The 'build-up' stage, from conception and early intra-uterine life to late adolescence or early twenties, is characterised by rapid successions of developmentally and socially sensitive periods. The second stage, starting in early adulthood, is a period of 'decline' from maximum attained health to loss of function, overt disease and death.

LIFEPATH adopts a study design that integrates social science and public health approaches with biology (including molecular epidemiology), using well-characterised population cohorts and omics measurements (particularly epigenomics). LIFEPATH includes information and biological samples from 17 cohorts, including several with extensive phenotyping and repeat biological samples, and a very large cohort (1 million individuals) without biological samples (WHIP, from Italy). The countries that are covered by the cohorts are France, Italy, Portugal, Ireland, UK, Finland, Switzerland and Australia. These cohorts are only a small proportion of all cohorts available in Europe, but we have chosen them for the combination of good measures of socioeconomic status, risk factors for non-communicable diseases (NCDs) and biomarkers already measured (or availability of blood samples for further testing). The majority of cohorts include 'hard' outcomes (diabetes, cancer, Cardiovascular Disease (CVD), total mortality), and the extensively phenotyped cohorts also include several measurements of the functional components of healthy ageing, including frailty, impaired vision, cognitive function, renal and brain function, osteoporosis, sleep disturbances and mental health. All age groups are represented with two birth cohorts, one cohort of adolescents and several cohorts encompassing young adults (age 18 and above). Furthermore, there is a strong representation of elderly subjects in seven cohorts.

The specific objectives of the project are: (a) to show that healthy ageing is an achievable goal for society; (b) to improve the understanding of the mechanisms through which healthy ageing pathways diverge by SES, by investigating life course biological pathways using omic technologies; (c) to examine the consequences of the current economic recession on health and the biology of ageing (and the consequent increase in social inequalities); (d) to provide updated, relevant and innovative evidence for healthy ageing policies (particularly 'health in all policies') using both observational studies and an experimental approach based on a reanalysis of data from a 'conditional cash transfer' randomised experiment in New York and new data collected as part of an earned income tax credit randomised experiment in Atlanta and New York. To achieve these objectives, data are used from three categories of studies: 1. national census-based follow-up data to obtain mortality by socioeconomic status; 2. cohorts with intense phenotyping and repeat biological samples; 3. large cohorts with biological samples. With these objectives and methodologies, LIFEPATH seeks to provide updated, relevant and innovative evidence to underpin future policies and strategies for the promotion of healthy ageing, targeted disease prevention and clinical interventions that address the issue of social disparities in ageing and the social determinants of health.

The present paper describes the design and some initial results of LIFEPATH as an example of the integration of social and biological sciences to provide evidence for public health policies.

Keywords

Social inequalities, socioeconomic status, healthy ageing, life-course, omics, biology

Aims in brief of LIFEPATH

LIFEPATH is a large consortium of cohort studies with the following aims:

I. To demonstrate that healthy ageing is highly variable in society, due to multiple environmental, behavioural and social circumstances that affect individual life trajectories.
II. To improve the understanding of the mechanisms through which healthy ageing pathways diverge by social circumstances, by investigating life course biological pathways using biomarkers and omic technologies.
III. To provide evidence on the reversibility of the poorer ageing trajectories experienced by individuals exposed to the strongest adversities; and to analyse the health consequences of the current economic recession in Europe (i.e. changes in social and economic circumstances).
IV. To provide updated, relevant and innovative evidence to inform future policies.

This paper describes the context, aims, objectives, design and early achievements of LIFEPATH.

Context for the project

Rationale

The striking difference in healthy ageing, quality of life and life expectancy observed between individuals from different socioeconomic groups is a major societal challenge that Europe is currently facing. Healthy ageing is an achievable goal in society as it is already experienced by individuals in the highest socioeconomic groups. Individuals with high socioeconomic status (SES) experience much better health and healthier ageing than groups with low SES. Healthy ageing is also strongly related to cumulative exposure to harmful events during the life course, and particularly in sensitive or critical periods. There is strong evidence that (a) the risk of disease is influenced by early exposures, including *in utero* exposures; (b) the life course is characterised by *critical* periods (during which changes in exposure have long term effects on disease risks) and *sensitive* periods (during which an exposure has stronger effects on development and, hence, disease risk than at other times).

LIFEPATH is based on a few assumptions that we describe below and have been addressed through the design choices we have made.

Challenges

1. The healthy ageing model

Socioeconomic differences in health are striking and are seen in all European countries (Mackenbach et al., 2008). Physical health deteriorates more rapidly with age for men and women from lower socioeconomic groups than among those from higher socioeconomic backgrounds. In addition, there is a large amount of evidence that this holds true also for mental health (see for instance Melchior et al., 2013).

This creates a progressively increasing health differential between social groups such that the average physical health of a 70 year old man or woman with high SES is similar to that of a person with low SES (occupational grade) who is eight years younger (Chandola, Ferrie, Sacker & Marmot, 2007). These differences, and the biological processes underlying them, cannot be understood without adopting a life course approach. In the project we describe below we use the revised Strachan-Sheikh (2004) model of life-course functioning (Blane, Kelly-Irving, d'Errico, Bartley & Montgomery, 2013; Kuh, 2007, 2014;), to describe ageing over a lifetime. This model presents ageing as a phenomenon with two broad stages: build-up and decline. The 'build-up' stage, from conception and early intra-uterine life to late adolescence or early twenties, is characterised by rapid successions of developmentally and socially sensitive periods. This stage strongly determines subsequent ageing trajectories as it influences the maximum attained level of health. The second stage, starting in early adulthood, is a period of 'decline' from maximum attained health to loss of function, overt disease and death. Lifestyle and exposures during the second stage can influence the rate at which functioning is lost (**Figure 1**- dashed line). We are aware that this working scheme is a simplification and does not incorporate the complex relationships that underlie ageing processes, but we use it as a working tool to organise our life course observations.

2. The complex causal network underlying SES and health

There is much evidence that low SES across the life course is associated with poor ageing

trajectories and early death (Chetty et al, 2016; House, Lantz & Herd, 2005; Mackenbach, J. et al., 2008; Stringhini, et al., 2010). There is evidence that the impact of SES on ageing is partly mediated by risk factors for non-communicable diseases (NCD), but this is not the only pathway. Also, the features of low SES in high-income countries have dramatically changed over the last century. In Europe, low SES is often associated with food abundance (but of poor quality), lack of physical exercise, and psychosocial stress (including sleep deprivation and mental problems) (Dugravot et al., 2010; Kivimäki et al., 2002; Kivimäki et al., 2012; Kivimäki et al., 2015;Stringhini et al., 2010; Stringhini et al., 2012). This adds to the traditional stressors, such as physical fatigue related to occupation and housework. Analyses in the Gazel cohort, which is part of Lifepath (Kivimaki, et al., 2008; Platts et al., 2016) show that demanding jobs were associated with fewer life years spent in good health.

SES is linked to many determinants of health, including: 1) access to and use of medical care; 2) access to health information; 3) patterns of unhealthy behaviours (smoking, heavy drinking, unhealthy diet, physical inactivity, drug use); 4) exposure to environmental and occupational hazards; 5) exposure to stressful life events; 6) access to resources mediating the physiological consequences of stress (social relationships and support and cultural capital); 7) early life adverse experiences; 8) age and time-related susceptibilities (Abel, 2008; James, Nelson, Ralph & Leather, 1997; Kelly-Irving, et al., 2011; Lantz, House, Mero & Williams, 2005; Lynch, Kaplan & Shema, 1997; Mitchell, Blane & Bartley, 2002; Siegrist & Marmot, 2004; van Doorslaer, Masseria, Koolman & the OECD Health Equity research Group, 2006). The Whitehall studies (part of LIFEPATH) have shown that the SES gradient in unhealthy ageing and CVD in particular is not completely explained by traditional risk factors, even when they are reduced to the lowest levels for the whole population (Kivimäki et al., 2008; Marmot, Shipley, Hemingway, Head & Brunner, 2008). In addition, global changes in the economy are accompanied by rising social inequalities. Instability of markets and lack of growth may lead to instability of health indicators and of ageing trajectories as well. The issue of social inequalities (and their health impact) is increasingly being recognised as a global challenge and priority by the global scientific and policy community, as demonstrated by the publication of an issue of the *Science* magazine devoted to socioeconomic differentials, their projections, and their impact on health (Chin & Culotta, 2014; Piketty & Saez, 2014).

3. Role of behavioural risk factors

Though epidemiology has extensively investigated chemical, physical and behavioural risk factors (such as smoking, diet, alcohol, physical exercise, and occupational factors), these still explain only a fraction of the burden of NCD's and of common disabilities. For example, a large fraction of breast, colon and prostate cancers is not explained by known risk factors, and the causes of common conditions such as cognitive impairment are largely unknown. The United Nations 25x25 strategy for NCDs (aiming to reduce mortality from NCDs by 25% by the year 2025) lists diabetes/obesity, alcohol, physical exercise, tobacco, raised blood pressure and salt as the main targets of preventive programmes. The strategy is limited by (a) referring exclusively to known risk factors, (b) excluding highly prevalent conditions that limit individual functioning, such as cognitive decline and musculoskeletal disorders, and (c) not accounting for the social structure that underpins the distribution of known risk factors and the effectiveness of preventive strategies and policies.

A considerable amount of evidence shows that socioeconomic factors are as important as lifestyle-related risk factors, and chemical and physical agents, in determining healthy ageing (House et al., 2005; Kelly-Irving et al., 2013; Stringhini et al., 2010; Stringhini et al., 2017) Low SES is one of the strongest predictors of healthy ageing in Europe, where, on average, disability-free life expectancy at age 65 is 4.5 years shorter in the lowest SES group than in the highest group (Majer, Nusselder, Mackenbach & Kunst, 2011). Differences in the actual life expectancy to age 65 in the US are even larger because social conditions predict a higher likelihood of premature death due to violence or unintentional injury (Muennig, Fiscella, Tancredi & Franks, 2010)

4. Biological pathways linking determinants to healthy ageing

What is missing in linking overarching determinants of SES with health and poor ageing is an understanding of the intermediate mechanisms

and pathways that relate low SES with deterioration of organic parameters. For example, research on immune markers in the Whitehall II study has shown that glucocorticoids and inflammation may in part explain how the body mediates the effects of low SES thus leading to disease, and this is partly independent of common known risk factors (Stringhini et al., 2013). More generally, recent studies have shown that SES can influence the global physiological dysregulation across the life course, measured using allostatic load, a measure of biological multisystem wastage (Barboza Solís et al., 2015; Merkin, Karlamangla, Roux, Shrager & Seeman, 2014;).

A research approach based on intermediate biomarkers is more powerful in identifying the links between SES and healthy ageing than one based on traditional risk factors alone because (a) it may explain subtle chronic effects acting throughout the life course (like chronic stress) that are not easily captured by questionnaire-based epidemiology; (b) it allows tracing signals that start in early life through to health effects in later life; (c) it provides an approach to the discovery of new pathways and causes of disease, particularly through the new omic technologies (Box 1). By analysing intermediate biomarkers potentially involved in various diseases, this approach is likely to reveal some common (currently unknown) roots of many NCDs (multi-morbidity), thus improving our ability to implement successful interventions, with a wide range of actions

---

Box 1. What are omics technologies?

Omic technologies allow to measures the whole set of compounds in a certain compartment (e.g. proteomics is about all proteins). These are the definitions of the omics available or measured *de novo* in LIFEPATH:

Epigenomics: The analysis of epigenetic changes in DNA, histones, and chromatin that regulate gene expression. Epigenetic changes are changes other than changes in DNA sequence that are involved in gene silencing.

Metabolomics: The scientific study of small molecules (metabolites) that are created from chemicals that originate inside the body (endogenously) or outside the body (exogenously). For purposes of the present report, metabolomics is assumed to include exogenous chemicals found in biological systems in their unmetabolised forms.

Proteomics: The analysis of the proteins produced by cells, tissues, or organisms. Analysis is conducted to understand the location, abundance, and post-translational modification of proteins in a biological sample.

Transcriptomics: Qualitative and quantitative analysis of the transcriptome, that is, the set of transcripts (mRNAs, noncoding RNAs, and miRNAs) that is present in a biological sample.

Research strategy

Aims

The specific objectives of LIFEPATH are:

- To demonstrate that healthy ageing is highly variable in society, due to multiple environmental, behavioural and social circumstances that affect individual life trajectories.
- To improve the understanding of the mechanisms through which healthy ageing pathways diverge by social circumstances, by investigating life course biological pathways using omic technologies.
- To provide evidence on the reversibility of the poorer ageing trajectories experienced by individuals exposed to the strongest adversities, by using an experimental approach (randomised experiments of an earned income tax credit programme in New York and Atlanta and a "conditional cash transfer" experiment for poverty reduction in New York City carried out by

MDRC); and to analyse the health consequences of the current economic recession in Europe (i.e. changes in social and economic circumstances).
- To provide updated, relevant and innovative evidence to inform future policies.

The project objectives will be accomplished by using different data sources:

- Europe-wide and national surveys (updated to 2014), including EU-27;
- Longitudinal cohorts (across Europe) with intense phenotyping and repeat biological samples;
- Other large cohorts with biological samples
- A large registry dataset with over a million individuals and very detailed information on work trajectories and health;
- A randomised experiment on conditional cash transfer for poverty reduction in New York City.
- A randomised experiment of an earned income tax credit programme in New York and Atlanta
- Publicly available cohort data.

The geographic location of the studies is shown in figure 2. A detailed description of the cohorts and corresponding datasets is presented in Table 1 (not including Europe-wide national surveys). Data are harmonised and integrated to conceptualise healthy ageing as a composite outcome at different stages of life, resulting from life-course environmental, behavioural and social determinants.

Cohorts

LIFEPATH includes information and biological samples from eight longitudinal population-based cohorts with extensive phenotyping and repeat biological samples, nine large longitudinal population-based cohorts with biological samples, and a very large cohort without biological samples (WHIP, from Italy). The countries that are covered by the cohorts are France, Italy, Portugal, Ireland, UK, Finland, Switzerland and Australia. These cohorts are only a small proportion of all cohorts available in Europe, but we have chosen them for the combination of good measures of socioeconomic status, risk factors for NCD and biomarkers already measured (or availability of blood samples for further testing). The majority of cohorts include 'hard' outcomes (diabetes, cancer, cardiovascular disease - CVD, total mortality), and the extensively phenotyped cohorts also include several measurements of the functional components of healthy ageing, including frailty, impaired vision, cognitive function, renal and brain function, osteoporosis, sleep disturbances and mental health. All age groups are represented with two birth cohorts, one cohort of adolescents and several cohorts encompassing young adults (age 18 and above). Furthermore, there is a strong representation of elderly subjects in seven cohorts (Table 1). This age structure allows us to address the *first challenge*, the life course perspective (model described in Figure 1).

Experimental studies to test policy strategies

Data from cohorts are complemented by data from the evaluation of two randomised experiments in the United States carried out by MDRC in New York: (a) the *Opportunity NYC–Family Rewards study*, a randomised experiment of conditional cash transfers (CCT) to help families break the cycle of poverty, and the first CCT program in a high-income country. The program was evaluated through a randomised controlled trial involving approximately 4,800 families and 11,000 children, half of whom could receive the cash rewards if they met the required conditions, and half who were assigned to a control group that could not receive the rewards. Data include extensive measures of social and economic conditions as well as extensive health assessments, which the present project plans to complement with additional biological measures. (b) The *Paycheck Plus*, a randomised experiment of tax credits to help adults without dependent children break the cycle of poverty. The Earned Income Tax Credit is a federal program in which low-income individuals and families receive an annual credit that varies according to their working income. The program was evaluated through a randomised controlled trial involving approximately 10,000 participants, with control participants assigned to the current maximum $500 annual return and experimental participants assigned to a maximum $2000 return. Data include extensive measures of social and economic conditions and we will collect data on serum cholesterol, C-reactive protein, glycosylated haemoglobin, blood pressure,

abdominal circumference, height, and weight after the intervention. These experimental studies are relevant to the *second challenge*, the complexity of causal relationships and the demonstration of reversibility of some effects.

Outcomes and definition of healthy ageing

Indicators of healthy ageing are being developed using the wealth of individual information collected in the cohorts participating in LIFEPATH (Table 1) as well as from publicly available datasets. Four 'har indicators of healthy ageing are being used (incident diabetes, cancer, cardiovascular disease, and death), and several functional and physiological indicators are also available (blood pressure, anthropometry, frailty, physical strength, walking speed and grip strength, eye macular health, sleep, cognitive function including dementia, mental health, bone health, renal function, urinary electrolytes and cardiometabolic measures). The World Health Organisation increasingly emphasises the need to shift from a definition of health based on hard indicators like mortality and diseases, to ability and disability indicators. In addition to being analysed separately, in LIFEPATH these measures will be incorporated into composite indicators of (un)healthy ageing. The contribution of each outcome to the composite indicators will be weighted based on the severity of the outcome (with premature death having the highest weight). To achieve this, we will incorporate methods and estimates from other projects such as the Global Burden of Diseases Collaboration (Lim et al., 2012), taking into account life stages and gender**.** Because the GBD system is controversial, we will also test other systems of weighting such as those based on mobility, impairment of usual activities, pain/discomfort and anxiety/depression derived from a systematic review of the literature. Early-life health events (such as sleep disturbances and changes in blood pressure) will be separated from typically late-life problems like frailty, osteoporosis, macular and cognitive impairment.

In LIFEPATH we have operationally defined healthy ageing with two components. First, through a general statement about healthy ageing and second, through recommendations about variable selection and choice of analyses across the life course using longitudinal datasets; these are provided below:

- Statement on Healthy Ageing: "Healthy ageing is the optimal state of performance and wellbeing capable for each particular phase of the lifecourse that can be expected in a society, across social and cultural groups of a population".

Key to variable selection is that the component variables be a) appropriate to the life course stage, and b) the wellbeing variables capture something of perception and lived experience.

Risk factors and healthy ageing

One of the leading hypotheses of LIFEPATH is that SES operates partially through an unequal distribution of conventional risk factors for poor health across SES strata, but that there is also an additional effect of SES on healthy ageing that is not explained by these conventional risk factors (*third challenge* above). Therefore, having good information on a number of risk factors for disease is key to ensure the success of the project. Also, crucial is the ability to encompass the whole trajectory that links SES with healthy ageing by modelling risk factors (and omics, see below) with proper analytic tools and study designs. These include mediation analysis, structural equation models, randomised experiments and natural experiments. Once the data are harmonised across cohorts, LIFEPATH will assess:

(a) the relationship between SES measures and risk factors;

(b) the relationship between individual risk factors for poor health and their combinations with multiple diseases and (composite) indicators of healthy ageing;

(c) the intermediate role of risk factors in the relationship of SES with healthy ageing (including interaction between risk factors);

(d) the proportion of the variance in healthy ageing that remains unexplained by measured risk factors.

Biological embedding and omic biomarkers over the life span

Differentiating the analysis of ageing in terms of the two stages of build-up and decline allows an appropriate and detailed exploration of ageing mechanisms, and the use of iterative biological measures from cohorts at different ages. For example, an analysis of the first 'build-up' stage may use as its embodiment measure a biological indicator collected in childhood or early adulthood; whereas an analysis of the 'decline' stage may treat mid-life biological indicators as mediating variables, and measures of physical functioning in later life or

diseases as an outcome. This dynamic approach to life course analyses is central in LIFEPATH and will be made possible by the integrated use of longitudinal cohorts capturing different life stages with repeat measures and biological samples.

By combining existing information from birth and adult longitudinal studies, LIFEPATH is organised into three main interacting streams, illustrated in Figure 3:

(a) ***an outer layer of the project: overarching DETERMINANTS***.

The focus is on healthy ageing as a continuous process in life ("life-path") and on SES differentials as key determinants, based on the well-documented assumption that SES over time is one of the strongest predictors (and also a summary measure) of the development and preservation of good health and of later-life diseases and disabilities;

(b) ***a middle layer of the project: RISK FACTORS***.

The project studies the influences SES has on healthy ageing and life expectancy via behavioural, occupational, nutritional, environmental and other modifiable risk factors for poor health, from childhood to adulthood;

(c) ***an inner layer of the project: BIOLOGICAL PATHWAYS and human social genomics.***

Omics data (in particular epigenomics, metabolomics and transcriptomics), is already available in several cohorts, and is being used to follow and compare the lifetime ageing pathways of individuals in different SES groups or with different exposure to risk factors from across the life-course. This will allow us to reconstruct biologically-embedded life-course ageing trajectories through the integration of (i) early- and later-life SES and risk factors, and (ii) early- and later-life omic measurements and markers of inflammation or immune response.

The scheme in Figure 3 explains the different components and outcomes of the project, as well as the types of biomarkers that can be developed as intermediate steps between socioeconomic status (SES), risk factors and unhealthy ageing: (a) epigenetic markers, both short-term (and amenable to modification), and long-term (irreversible changes that can be attributed to early-life and late-life exposures); (b) markers of stress and HPA-axis dysregulation; (c) markers of inflammation and immune response (particularly important in cardiometabolic disease and cancer); (d) markers of neural function and structure.

The following biomarkers are already available or will be measured with existing funds: methylome (n=4,250); inflammation markers (n=61,000); metabolomics (n=23,000). Most cohorts have C - reactive protein (CRP), and many have cytokines measurements.

New methylome analyses will be performed for 2,500 new subjects and new transcriptomic analyses for about 600 subjects. Peripheral blood is a convenient source for epigenetic testing, but cellular heterogeneity can confound or mask the results (Adalsteinsson et al,. 2012; Consortium TEP, 2012; Liu et al., 2013), because epigenetic signatures differ from one cell subtype to another (Reinius et al., 2012). Moreover, the composition of the blood cells can change under a plethora of pathophysiological conditions, and this can also be influenced by SES. The importance of measuring epigenetic patterns across specific cell subtypes also depends on disease status, and subtle differences are likely more important for multifactorial traits. For these reasons, the first analyses of the methylome will be performed in two specific blood cell types in 500 Whitehall II subjects. We will then validate the findings in PBMC in cohorts with cell counts available. Also, the biomarker stock will be enriched by measuring IL6 and other cytokines in 6,600 additional subjects in Whitehall II (funded by the US NIH) and in Skipogh, in which also the methylome and RNA sequencing will be measured (n=750).

In some cohorts we also have *genetic data* (Table 1). This allows three types of analyses: first, we can measure the relative importance of genetic predisposition to disease vs environmental variables including SES; second, we can look at gene-environment interactions with our non-genetic variables including SES (low-penetrance susceptibility); third, we can apply the concept of Mendelian Randomisation, i.e. identify gene variants that are associated with intermediate factors that connect SES with health outcomes, and reinforce causal reasoning (given that no genetic variants have been clearly identified that directly influence SES itself). Genotyping will be performed (or is already available) through other funds for 17,000 subjects from Airwave, 6,600 from Whitehall II, 1,100 from Skipogh and 6,000 from Colaus. The Cardiometabochip (200K SNPs) will be used (200K

SNP) in all of them except in Colaus (500K SNP Affymetrix chip).

Statistical Methods

This project is working towards developing novel statistical models integrating the different features determining healthy life trajectories. Statistical models are designed to (a) test different conceptual pathways to (un)healthy ageing, integrating information from risk factors, SES, omics and outcomes that contribute to healthy ageing; (b) provide estimates of the weight of different causes, distal/fundamental or proximal, to enable the simulation of different scenarios of public health policies or interventions which might improve healthy living.

An outline of the different aspects of statistical analyses is provided in Table 2. Statistical methods are also being developed to investigate the association between SES and healthy ageing, between SES and risk factors, and between risk factors and healthy ageing indices as conceptualised above. Additional methodological efforts will rely on the development of omics profiling techniques integrating data arising from different platforms, and mechanistic models investigating the interplay of these different sets of markers.

Omic signatures – As summarised in Table 3, numerous analyses rely on the identification, from high dimensional omics profiles, of potential omic signatures of SES-related exposures/patterns.

Biological mechanisms – The full exploration of biological mechanisms involved in the SES-affected regulatory cascades relies on in-depth exploration of the correlation structures and the potential effect mediations among the candidate markers found within and across the different platforms.

Time-related patterns – In order to exploit longitudinal data available in LIFEPATH, specific time-related patterns are sought, which are predictive of the quality of ageing.

Conceptual model of ageing and predictive score – At the end of the process, LIFEPATH will construct a holistic conceptual model of ageing by confronting it in an iterative way to data from cohorts involved in the project. This model will integrate the three layers (*outer*: determinants; *intermediate*: risk factors; *inner*: biomarkers and omics), so that plausible evidence-based pathways encompassing the layers can be built.

Age and sex/gender disparities

From birth onwards, a sex-based mortality divide is present, with males having a greater mortality risk than females. Conversely, women have longer life-expectancies but live with higher rates of morbidity and disability. These differences vary in their nature across the lifespan. Sex and gender are central to any research into the processes of ageing, healthy or otherwise. The difficult task in disentangling sex and gender has been highlighted in epidemiology by Krieger (2003) who advocates more complex formulations such as "biologic expressions of gender" and "gendered expression of biology". Here, we conceptualise sex/ gender "as a domain of complex phenomena *that are simultaneously biological and social*, rather than a domain in which the social and biological "overlap"" (Springer et al, 2012). All age groups are included in LIFEPATH, with two birth cohorts, one cohort in adolescents, several cohorts encompassing young adults (age 18 and above), and representation of elderly subjects in seven cohorts.

Policy aims

LIFEPATH is devoted to develop policy implications of the scientific findings. Early work in this project will develop conceptual models of the intermediate mechanisms and pathways through which SES gradients influence organic parameters. These models will be empirically examined using data from the study countries to produce evidence of the importance of different pathways in explaining SES inequalities in chronic disease occurrence. The LIFEPATH project takes the innovative step of combining these results with the existing scientific and policy literatures, with the aim of developing a health impact assessment framework (HIA) which will allow the research team to simulate the impact of different policy choices on the overall occurrence of disease and its distribution across different SES groups. A number of HIA models for specific disease groups are already available. These will be further developed to capture healthy ageing using the results from the LIFEPATH project. In particular, a model using Markov chains has been developed by the Erasmus group (DynamoOHIA) for application to routine data. The model can be expanded and enriched with socioeconomic variables from the LIFEPATH cohorts and it can be tested among the different European cohorts.

Early achievements so far

We describe here a few early achievements of the project, that support some of the basic assumptions and respond to the first challenges we have identified in LIFEPATH.

SES variables and harmonisation

The LIFEPATH consortium brings together data from early life and adult European cohorts with intense phenotyping and repeat biological samples, other large cohorts with biological samples and a large registry dataset with over a million individuals and very rich information on work trajectories and health (D'Errico et al, 2017). To merge and analyse together data from the different LIFEPATH cohorts, information on occupational class, education, father's occupational class and income has been harmonised. The harmonisation was performed in the initial phases of the project for each of the early life cohorts and adult cohorts participating in the study (Table 1). Detailed codebooks have been created and given that SES indicators relevant for adults differ from those important to characterise children SES, the cohorts have been harmonised using different methods.

Amongst men, the proposed three-level classification of occupational class based on the European Socio-economic Classification (ESeC) – higher, intermediate and lower professions – and education appear not to differ substantially from more detailed classifications in discriminating between main social strata, and in predicting differences in mortality between them (D'Errico et al., 2017). In spite of differences in recruitment among the cohorts, especially in terms of time, age, gender composition and type of sample, variability in the distribution of the socioeconomic indicators in the different study populations was relatively low. SES indicators categories observed in the different cohorts were also roughly comparable in their distribution to those observed in other studies conducted in Europe, including several conducted on representative samples of the general population. Despite the categories employed, both education and occupational class differed somehow from those used in the present study and the study populations did not share the same characteristics of those included in LIFEPATH. A shift toward higher occupational classes was noted between fathers' and subjects' occupations in most cohorts, which seem to reflect an increasing trend of social improvement in these populations.

Strong differences in mortality between genders were observed for all SES indicators, with much higher and significant associations in males than females, although characterised by variable strength in the different cohorts (Figure 4).

SES as a risk factor of physiological wear and tear

Understanding how human environments affect our health by 'getting under the skin' and penetrating the cells, organs and physiological systems of our bodies is a key tenet in public health research. In LIFEPATH we examine the idea that early life socioeconomic position can be biologically embodied, potentially leading to the production of health inequalities across population groups. Allostatic load (AL), a composite measure of overall physiological wear-and-tear, could allow for a better understanding of the potential biological pathways playing a role in the construction of the social gradient in adult health. We have investigated the factors mediating the link between two components of parental socioeconomic position, maternal education (ME) and parental occupation (PO), and AL at 44 years. Data was used from 7,573 members of the 1958 British birth cohort follow-up to age 44. AL was constructed using 14 biomarkers representing four physiological systems. We assessed the contribution of financial/materialist, psychological/psychosocial, educational, and health behaviours/BMI pathways over the life course, in mediating the associations between ME, PO and AL. ME and PO were mediated by three pathways: educational, material/financial, and health behaviours for both men and women. A better understanding of embodiment processes leading to disease development may contribute to developing adapted public policies aiming to reduce health inequalities (Barboza Solís et al., 2016). In the same cohort we also showed that AL was associated with subsequent health status.

SES as a risk factor of high grade inflammation

To explore further potential biological embedding and the consequences of socioeconomic position experiences from early life to adulthood, we investigated how socioeconomic position indicators at different points across the life course may be related to a combination of 28 inflammation markers. Using blood-derived inflammation profiles measured by a multiplex array in 268 participants from the Italian

component of the European Prospective Investigation into Cancer and Nutrition cohort, we evaluated the association between early life, young adulthood and later adulthood socioeconomic position with each inflammatory marker separately, or by combining them into an inflammatory score. We identified an increased inflammatory burden in participants whose father had a manual occupation, through increased plasma levels of CSF3 (G-CSF; $\beta = 0.29$; $P = 0.002$), and an increased inflammatory score ($\beta = 1.96$; $P = 0.029$). Social mobility was subsequently modelled by the interaction between father's occupation and the highest household occupation, revealing a significant difference between 'stable Non-manual' profiles over the life course versus 'Manual to Non-manual' profiles ($\beta = 2.38$, $P = 0.023$). Low socioeconomic position in childhood is associated with modest increase in adult inflammatory burden; however, the analysis of social mobility suggests a stronger effect of an upward social mobility over the life course.

Outstanding issues

LIFEPATH is a large and ambitious undertaking, and it will try to address some of the main challenges in the field, but it is foreseen that some issues will remain outstanding:

- Measurement error is inherent in SES but also in biomarkers and omics, and this is likely to lead to blurring of the causal pathways
- In spite of the development of statistical methods for mediation analysis, the complex (multi-layered) nature of the relationships between SES, risk factors, biomarkers and outcomes makes it unlikely that a complete picture can be developed. This includes the heterogeneity of the cohorts involved.
- How far our inferences can be extended to other populations with a different distribution of risk factors, SES composition and life expectancy is unknown.

Access to LIFEPATH resources

LIFEPATH aims to become an open-access source of data for biosocial research. The central database contains harmonised variables for socioeconomic status, risk factors, biomarkers and outcomes. Access will be regulated by the Steering Committee and rules for access will be made available to interested researchers via the Principal Investigator.

Conclusions

The marriage between biology and social sciences

Accelerated ageing and many human diseases result from a complex interaction between social and biological factors. Risky behaviours, occupational and environmental exposures and psychosocial stress are rooted in societal structures, have historical foundations, and lead to alterations in physiological states (such as allostatic load) that are a prelude to overt disease. The study of the interplay of social and biological factors in LIFEPATH, poses several challenges:

Conceptual challenges: the definition of socioeconomic differentials depends on different theoretical constructs that put variable emphasis on status, job position (social class), prestige, and other factors. Similar challenges are encountered in biological modelling, in relation to the relevance of certain markers (such as DNA methylation or metabolomics) to the exploration of ageing trajectories.

Practical challenges, such as the availability and harmonisation of good and well measured descriptors of SES across the cohorts. The same challenges are encountered in biomarker science, where measurements are affected by 'nuisance parameters' (such as batch) and confounders, and also need harmonisation.

A major challenge is to avoid reductionism in the interpretation of the findings. We seek to establish causal relations across a spectrum that encompasses societal structures, inter-individual relationships, socioeconomic position, exposure to risk factors, and the underlying layers of biological mechanisms (that in turn involve whole individuals, tissues, cells and molecules). Causal models have been proposed separately in social sciences and in biological sciences, and they tend to be very different. Our ambition is to integrate them not only conceptually but also operationally and technically in the use of statistically modelling).

Acknowledgements

This project has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 633666.

References

Abel, T. (*2008).* Cultural capital and social inequality in health. *Journal of Epidemiology and Community Health*. *62*:e13. https://doi.org/10.1136/jech.2007.066159

Adalsteinsson, B. T., Gudnason, H., Aspelund, T., Harris, TB, Launer, LJ., Eiriksdottir, G., Smith, A.V. & Gudnason, V. *(2012).* Heterogeneity in white blood cells has potential to confound DNA methylation measurements. *PLoS ONE; 7*(10), e46705. https://doi.org/10.1371/journal.pone.0046705

Alves, L., Silva, S., Severo, M., Costa, D., Pina, M.F., Barros, H. & Azevedo, A.. (*2013).* Association between neighbourhood deprivation and fruits and vegetables consumption and leisure-time physical activity: a cross-sectional multilevel analysis*. BMC Public Health, 13*, 1103. https://doi.org/10.1186/1471-2458-13-1103

Avendano, M., Glymour, M.M., Banks, J., & Mackenbach, J.P. (*2009).* Health Disadvantage in US Adults Aged 50 to 74 years: a comparison of the health of rich and poor Americans with that of Europeans. *American Journal of Public Health, 99,* 540-548. https://doi.org/10.2105/AJPH.2008.139469

Barboza, S., Kelly-Irving, M., Fantin, R., Darnaudery, M., Torrisani, J., Lang, T. & Delpierre, C. *(2015)*. Adverse childhood experiences and physiological wear-and-tear in midlife: Findings from the 1958 British birth cohort*. Proceedingof the National Acadamy of Sciences of the USA,* E738-E746. https://doi.org/10.1073/pnas.1417325112

Barboza Solís, C., Fantin, R., Castagné, R., Lang, T., Delpierre, C & Kelly-Irving, M. *(2016).* Mediating pathways between parental socio-economic position and allostatic load in mid-life: Findings from the 1958 British birth cohort. *Social Science and Medicine 165,* 19-27. https://doi.org/10.1016/j.socscimed.2016.07.031

Barboza Solís, C., Fantin, R., Kelly-Irving, M. & Delpierre. C. *(2016). Physiological wear-and-tear and later subjective health in mid-life: Findings from the 1958 British birth cohort. Psychoneuroendocrinology 18*, 74, 24-33. https://doi.org/10.1016/j.psyneuen.2016.08.018

Blane, D., Kelly-Irving, M., d'Errico, A., Bartley, M. & Montgomery, S. *(2013)* Social-biological transitions: how does the social become biological? *Longitudinal and Life Course Studies*.

Borghol, N., Suderman, M., Mccardle, W. & Szyf, M. *(2012).* Associations with early-life socio-economic position in adult DNA methylation*. International Journal of Epidemiology, 41*(1), 62-74. https://doi.org/10.1093/ije/dyr147

Chadeau-Hyam, M., Vermeluen, R.C., Hebels, D.G., Castagne, R., Camapnella, G., Portengen, L., Kelly, R.S., Bergdahl, I.A., Melin, B., Hallmans, G., Palli, D., Krogh, V., Tumino, R., Sacerdote, C., Panico, S., de Kok, T.M., Smith, M.T., Kleinjans, J.C., Vineis, P., Kyrtopoulos, S. & EnviroGenoMarkers project consortium. *(2014).* Pre-diagnostic Transcriptomics Markers of chronic lymphocyctic leukemia reveal perturbations 10 years before diagnosis*. Annals of Oncology, 25*(5),1065-72. https://doi.org/10.1093/annonc/mdu056

Chandola, T., Ferrie, J., Sacker, A. & Marmot, M. *(2007).* Social inequalities in self-reported health in early old age: follow-up of prospective cohort study. *BMJ, 12*, 334(7601), 990. https://doi.org/10.1136/bmj.39167.439792.55

Chin, G. & Culotta, E. *(2014)* The science of inequality. What the numbers tell us. Introduction. *Science, 344*(6186), 818-21. https://doi.org/10.1126/science.344.6186.818

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Beregron, A. & Cutler, D. *(2016).* The association between income and life expectancy in the United States, 2001-2014. *JAMA, 315*(16), 1750-66. https://doi.org/10.1001/jama.2016.4226

Correia, S., Rodrigues, T. & Barros, H. *(2014).* Socioeconomic variations in female fertility impairment: a study in a cohort of Portuguese mothers. *BMJ Open 4*(1), e003985. https://doi.org/10.1136/bmjopen-2013-003985

d'Errico A, Ricceri F, Stringhini S, Carmeli C, Kivimaki M, Bartley M… Vineis P; LIFEPATH Consortium. (2017). Socioeconomic indicators in epidemiologic research: A practical example from the LIFEPATH study. *PLoS One. 12*(5):e0178071. https://doi.org/10.1371/journal.pone.0178071

Dartois, L., Fagherazzi, G., Boutron-Ruault, M.C., Mesrine, S. & Clavel-Chapelon, F. *(2014).* Association between five lifestyle habits and cancer risk: results from the E3N cohort. *Cancer Prevention Research (Phila)*;7(5), 516-25. https://doi.org/10.1158/1940-6207.CAPR-13-0325

Dugravot, A., Sabia, S., Stringhini, S., Kivimaki, M., Westerlund, H., Vahtera, J., Gueguen, A., Zins, M., Goldberg, M., Nabi, H. & Singh-Manoux, A. *(2010).* Do socioeconomic factors shape weight and obesity trajectories over the transition from midlife to old age? Results from the French GAZEL cohort study. *American Journal of Clinical Nutrition, 92*(1),16-23. https://doi.org/10.3945/ajcn.2010.29223

Eikemo, T.A., Hoffmann, R., Kulik, M.C., Kulhánová, I., Toch-Marquardt, M., Menvielle, G., Looman, C., Jasilionis, D., Martikainen, P., Lundberg, O., Mackenbach, J.P. & EURO-GBD-SE Consortium. *(2014).* How can inequalities in mortality be reduced? A quantitative analysis of 6 risk factors in 21 European populations. *PLoS One, 9*(11), e110952. https://doi.org/10.1371/journal.pone.0110952

ENCODE Project Consortium. *(2012).* An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57–74. https://doi.org/10.1038/nature11247

Gallo, V., MAckenbach, J.P., Ezzati, M., Menvielle, G., Kunst, A.E., Rohrmann, S., Kaaks, R., Teucher, B., Boeing, H., Bergmann, M.M., Tionneland, A., Dalton, S.O., Overvad, K., Redondo. M.L., Aguado, A., Daponte, A., Arriola, L., Navarro, C., Gurrea, A.B., Khaw, K.T., Wareham, N., Ket, T., NAska, A., Tricopoulou, A., Trichopoulos, D., Masala, G., Panico, S., Contiero, P., Tumino, R., Bueno-de-Mesquita, H.B., Siersema, P.D., Peeters, P.P., Zackrisson, S., Almguist, M., Eriksson, S., Hallmans, G., Skeie, G., Braaten, T., Lund, E., Illner, A.K., Mouw, T., Riboli, E. & Vineis, P. *(2012).* Social inequalities and mortality in Europe – results from a large multi-national cohort. *PLoS One, 7*(7):e39013. https://doi.org/10.1371/journal.pone.0039013

Glymour, M., Avendano, M, & Kawachi, I. *(2014).* Socioeconomic Status and Health. In L.F. Berkman, I. Kawachi, & M. Glymour (Eds.), *Social Epidemiology, 2nd Edition* pp. 17-62. New York: Oxford University Press. https://doi.org/10.1093/med/9780195377903.003.0002

Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. *(2010).* Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ,* 341:c6624. https://doi.org/10.1136/bmj.c6624

Horvath, S. *(2013).* DNA methylation age of human tissues and cell types. *Genome Biology, 14*(10), R115. https://doi.org/10.1186/gb-2013-14-10-r115

House, J.S., Lantz, P.M. & Herd, P. *(2005).* Continuity and change in the Social Stratification in Aging and Health over the Life Course: Evidence from a Longitudinal Study from 1986 to 2001/2002 (Americans' Changing Lives Study*). Journal Gerontoogy Series B Psychological sciences and Social Sciences, 2*, 15-26. https://doi.org/10.1093/geronb/60.Special_Issue_2.S15

James, W.P., Nelson, M., Ralph, A. & Leather, S. *(1997).* Socioeconomic determinants of health. The contribution of nutrition to inequalities in health. *BMJ, 314*, 1545-9. https://doi.org/10.1136/bmj.314.7093.1545

Kelly-Irving, M., Lepage, B., Dedieu, D., Bartley, M., Blane, D., Grosclaude, P., Lang, T. & Delpierre, C. *(2013).* Adverse childhood experiences and premature all-cause mortality. *European Journal of Epidemiology, 28*(9), 721-34.

Kelly-Irving, M., Delpierre, C., Schieber, A.C., Lepage, B., Rolland, C., Afrite, A., Pascal, J., Cases, C., Lombrail, P. & Lang, T. *(2011).* Do general practitioners overestimate the health of their patients with lower education? *Social Sciences & Medicine, 73*, 1416-21. https://doi.org/10.1007/s10654-013-9832-9

Kickbusch, I.S. *(2001).* Health literacy: addressing the health and education divide. *Health Promotion International*, *16*, 289-97. https://doi.org/10.1016/j.socscimed.2011.07.031

Kivimäki, M., Leino-Ajas, P., Luukkonen, R., Riikimaki, H., Vahtera, J. & Kirjonen, J. *(2002).* Work stress and risk of cardiovascular mortality : prospective cohort study of industrial employees. *BMJ*, *325*(7369), 857. https://doi.org/10.1136/bmj.325.7369.857

Kivimäki, M., Shipley, M.J., Ferrie, J.E., Singh-Manoux, A., Batty, G.D., Chandola, T., Marmot, M.G. & Davey Smith, G. *(2008).* Best-practice interventions to reduce socioeconomic inequalities of coronary heart disease mortality in UK: a prospective occupational cohort study. *Lancet, 372*(9650), 1648-54. https://doi.org/10.1016/S0140-6736(08)61688-8

Kivimaki, M. et al. *(2008).* Sickness absence as a prognostic marker for common chronic conditions: analysis of mortality in the GAZEL study. *Occupational and Environmental Medicine, 65*, 820-826. https://doi.org/10.1136/oem.2007.038398

Kivimäki, M., Head, J., Ferrie, J.E., Singh-Manoux, A., Westerlund, H., Vahtera, J., Leclerc, A., Melchior, M., Chevalier, A., Alexanderson, K., Zins, M. & Goldberg, M. *(2012).* Job strain as a risk factor for coronary heart disease: a collaborative meta-analysis of individual participant data. *Lancet*, *380,* 1491-7. https://doi.org/10.1016/S0140-6736(12)60994-5

Kivimäki, M., Jokela, M., Nyberg, S.T., Singh-Manoux, A., Fransson, E.I., Alfredsson, L., Bjorner, J., Borritz, M., Burr, H., Casini, A., Clays, E., De Bacquer, D., Dragano, N., Erbel, R., Geuskens, G.A., Hamer, M., Hooftman, W.E., Houtman, I.L., Jockel, K.H., Kittel, F., Knutsson, A., Koskenvuo, M., Lunau, T., Madsen, E.E.H., Nielsen, M.L., Nordin, M., Oksanen, T., Pejtersen, J.H., Pentti, J., Rugulies, R., Salo, P., Shipley, M.J., Siegrist, J., Steptoe, A., Suominen, S.B., Theorell, T., Vahtera, J., Westerholm, P.J.M., Westerlund, H., O'Reilly, D., Kumari, M., Batty, D.G., Ferrie, J.E. & Virtanen, M. *(2015).* Long working hours and risk of coronary heart disease and stoke: a systematic review and meta-analysis of published and unpublished data for 603 838 individuals. *Lancet*, *386*, 1739-46. https://doi.org/10.1016/S0140-6736(15)60295-1

Krieger, N., Chen, J.T., Waterman, P.D., Rehkopf, D.H. & Subramanian, S.V. *(2003).* Race/ ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures – the public health disparities geocoding project. *American Journal of Public Health*, *93*(10),1655-71. https://doi.org/10.2105/AJPH.93.10.1655

Kuh, D.*(2007).* A life course approach to healthy ageing, frailty, and capability. *Journal of Gerontology Series A, Biological Sciences and Medical Sciences*, *62,* 717-21. https://doi.org/10.1093/gerona/62.7.717

Kuh, D., Karunananthan, S., Bergman, H. & Cooper, R. *(2014).* A life course approach to healthy ageing: maintaining physical capability. *Proceedings of the Nutrition Society, 73*(2), 237-248. Oxford, Oxford University Press. https://doi.org/10.1017/S0029665113003923

Lam, L.L., Emberly, E., Fraser, H.B., Neumann, S.M., Chen, E., Miller, G.E. & Kobor, M.S. *(2012).* Factors underlying variable DNA methylation in a human community cohort. *Proceedings of the National Academy of Sciences of the USA*, *109*, Suppl 2: 17253-60. https://doi.org/10.1073/pnas.1121249109

Lantz, P.M., House, J.S., Mero, R.P. & Williams, D.R. *(2005).* Stress, life events, and socioeconomic disparities in health: results from the American's Changing Lives Study. *Journal of Health & Social Behavior*; *46,* 274-88. https://doi.org/10.1177/002214650504600305

Lantz, P.M., House, J.S., Lepkowski, J.M., Williams, D.R., Mero, R.P. & Chen, J. *(1998).* Socioeconomics factors, health behaviours, and mortality: results from a nationally representative prospective study of US adults. *Journal of the American Medical Association*, *279*(21),1703-8. https://doi.org/10.1001/jama.279.21.1703

Larsen, P. S., Kamper-Jorgensen, M., Adamson, A., Barros, H., Bonde, J. P., Brescianini, S., Brophy, S., Casas, M., Charles, M.A., Devereux, G., Eggesbo, M., Fantini, M.P., Frey, U., Gehring, U., Grazuleviciene, R., Henriksen, T.B., Hertz-Picciotto, I., Heude, B., Hryhorczuk, D.O., Inskip, H., Jaddoe, V.W., Lawlor, D.A., Ludvigsson, J., Kelleher, C., Kiess, W., Koletzo, B., Kuehni, C.E., Kull, I., Kyhll, H.B., Magnus, P., Momas, I., Murray, D., Pekkanen, K., Porta, D., Poulsen, G., Richiardi, L., Roeleveld, N., Skovgaard, A.M., Sram, R.J., Strandberg-Larsen, K., Thijs, C., Van Eijsden, M., Wright, J., Vrijheid, M. & Andersen, A.M. *(2013).* Pregnancy and birth cohort resources in europe: a large opportunity for aetiological child health research. *Paediatric and Perinatal Epidemiology, 27*(4), 393-414. https://doi.org/10.1111/ppe.12060

Leist, A.K., Hessel, P. & Avendano, M. *(2014).* Do economic recessions during early and mid-adulthood influence cognitive function in older age? *Journal of Epidemiology and Community Health; 68*(2): 151-158. https://doi.org/10.1136/jech-2013-202843

Leombruni, R., Richiardi, M., Demaria, M. & Costa G. Life expectancy, strenuous work and pension system's fairness. First evidence from the Work Histories Italian Panel. *Epidemiologia & Prevenzione., 2010 Jul-Aug;34*(4), 150-8.

Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Ezzati, M et al. *(2012).* A comparative risk assessment of burden disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, *380*, 2224-60. https://doi.org/10.1016/S0140-6736(12)61766-8

Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekstrom, T.J. & Feinberg, A.P. *(2013).* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology, 31*, 142–147. https://doi.org/10.1038/nbt.2487

Loscalzo, J. & Barabasi, A.L. *(2011).* Systems biology and the future of medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 3*(6), 619-27. https://doi.org/10.1002/wsbm.144

Lynch, J.W., Smith, G.D., Kaplan, G.A. & House, J.S. *(2000).* Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions. *BMJ;320*(7243):1200-4. https://doi.org/10.1136/bmj.320.7243.1200

Lynch, J.W., Kaplan, G.A. & Shema, S.J. *(1997).* Cumulative impact of sustained economic hardship on physical, cognitive, psychological, and social functioning. *N Engl J Med;337*:1889-95. https://doi.org/10.1056/NEJM199712253372606

Lynch J.W, Kaplan G.A.& Salonen J.T. *(1997).* Why do poor people behave poorly? Variation in adult health behaviours and psychosocial characteristics by stages of the socioeconomic lifecourse. *Soc Sci Med*; 44:809-819. https://doi.org/10.1016/S0277-9536(96)00191-8

Mackenbach, J., Stirbu, I., Roskam, A.J.R., Schaap, M.M., Menvielle, G., Leinsalu, M. & Kunst, A.E. *(2008).* Socioeconomic inequalities in health in 22 European Countries. *N Engl J Med*; *358*(23):2468-81. https://doi.org/10.1056/NEJMsa0707519

Mäki, N.E.,Martikainen, P.T., Eikemo, T., Menvielle, G., Lundberg, O. & Mackenbach, J. *(2014).* The potential for reducing differences in life expectancy between educational groups in five European countries: the effects of obesity, physical inactivity and smoking. *J Epidemiol Community Health.; 68*(7):635-40. https://doi.org/10.1136/jech-2013-203501

Majer, I.M., Nusselder, W.J., Mackenbach, J.P. & Kunst, A.E. *(2011).* Socioeconomic inequalities in life and health expectancies around official retirement age in 10 Western-European countries. *J Epidemiol Community Health; 65*(11):972-9. https://doi.org/10.1136/jech.2010.111492

Marmot, M.G., Shipley, M.J., Hemingway, H., Head, J. & Brunner E.J. *(2008).* Biological and behavioural explanations of social inequalities in coronary heart disease: the Whitehall II study. *Diabetologia;51*:1980-8. https://doi.org/10.1007/s00125-008-1144-3

McEwen, B.S. & Stellar, E. *(1993).* Stress and the individual. Mechanisms leading to disease. *Arch Intern Med;153*(18):2093-101. https://doi.org/10.1001/archinte.1993.00410180039004

McGuinness, D., Mcglynn, L.M., Johnson, P.C.D. & Shiels, P.G. *(2012).* Socio-economic status is associated with epigenetic differences in the pSoBid cohort. *Int J Epidemiol;41*(1): 151-60. https://doi.org/10.1093/ije/dyr215

Melchior, M., Chastang, J.F., Head, J., Goldberg, M., Zins, M., Nabi, H. & Younes, N. *(2013).* Socioeconomic position predicts long-term depression trajectory: a 13-year follow-up of the GAZEL Cohort Study. *Mol Psychiatry; 18:*112–121. https://doi.org/10.1038/mp.2011.116

Merkin, S.S., Karlamangla, A., Roux, A.V., Shrager, S. & Seeman, T.E. *(2014).* Life course socioeconomic status and longitudinal accumulation of allostatic load in adulthood: multi-ethnic study of atherosclerosis. *Am J Public Health;104*(4):e48-55. https://doi.org/10.2105/AJPH.2013.301841

Miller GE, Chen, E. & Parker, K.J. *(2011).* Psychological stress in childhood and susceptibility to chronic diseases of aging : moving towards a model of behavioral and biological mechanisms. *Psychol Bull;137*(6): 959-97. https://doi.org/10.1037/a0024768

Mitchell, R., Blane, D. & Bartley, M. *(2002).* Elevated risk of high blood pressure: climate and the inverse housing law. *Int J Epidemiol*;31:831-8. https://doi.org/10.1093/ije/31.4.831

Mosca, I., Bhuachalla, B.N. & Kenny, R.A. *(2013).* Explaining Significant differences in subjective and objective measures of cardiovascular health: evidence for the socioeconomic gradient in a population-based study. *BMC Cardiovasc Disord*;13:64. https://doi.org/10.1186/1471-2261-13-64

Muennig, P., Fiscella, K., Tancredi, D. & Franks, P. *(2010).* The relative health burden of selected social and behavioral risk factors in the United States: Implications for Policy. *Am J Public Health*: 1758-1764. https://doi.org/10.2105/AJPH.2009.165019

Pearce, N., Ebrahim, S., McKee, M., Lamptey, P., Barreto, M.L., Matheson, D., Walls, H., Foliaki, S., Miranda, J., Chimeddamba, O., Marcos, L.G., Haines, A. & Vineis, P. *(2014).* The road to 25x25 :how cna the five-target strategy reach its goal? *Lancet Global Health(in press)*. https://doi.org/10.1016/S2214-109X(14)70015-4

Piketty, T. & Saez, E. *(2014).* Inequality in the long run. *Science*; 344:838-42. https://doi.org/10.1126/science.1251936

Platts, L.G., Head, J., Stenholm, S., Singh Chungkham, H., Goldberg, M. & Zins, M. *(2016).* Physical occupational exposures and healthy life expectancy in a French occupational cohort. *Occup Environ Med.*

Ponte, B., Prujim, M., Ackermann, D., Vuistiner, P., Eisenberger, U., Guessous, I., Rousson, V., Mohaupt, M.G., Alwan, H., Ehret, G., Pechere-Bertschi, A., Paccaud, F., Staessen, J.A., Vogt, B., Burnier, M., Martin, P.Y. & Bochud, M. *(2014).* Reference values and factors associated with renal resistive index in a family-based population study. *Hypertension;63*(1):136-42. https://doi.org/10.1161/HYPERTENSIONAHA.113.02321

Ramos, E. & Barros, H. *(2007).* Family and school determinants of overweight in 13-year-old Portuguese adolescents. *Acta Paediatr.;96*(2):281-6. https://doi.org/10.1111/j.1651-2227.2007.00107.x

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlen, S.E., Greco, D., Soderhall, C., Scheynius, A. & Kere, J. *(2012).* Differential DNA Methylation in Purified Human Blood Cells : implications for cell lineage and studies on disease susceptibility. *PLoS ONE. 7*(7), e41361. https://doi.org/10.1371/journal.pone.0041361

Siegrist, J. & Marmot, M. *(2004).* Health inequalities and the psychosocial environment – two scientific challenges. *Soc Sci Med;58*:1463-73. https://doi.org/10.1016/S0277-9536(03)00349-6

Springer, K.W., Hankivsky, O. & Bates, L.M. *(2012).* Gender and health: relational, intersectional, and biosocial approaches. *Soc Sci Med;74*(11):1661-6. https://doi.org/10.1016/j.socscimed.2012.03.001

Strachan, D. & Sheikh, A. In D. Kuh & Y. Ben-Shlomo. *(2002).* A Life Course Approach to Chronic Disease Epidemiology. *2nd edition (pp. 240-259). Oxford: Oxford University Press*.

Stringhini, S., Sabia, S., Shipley, M., Brunner, E., Nabi, H., Kivimaki, M. & Singh-Manoux, A. *(2010).* Association of socioeconomic position with health behaviors and mortality. *JAMA;303*(12):1159-66. https://doi.org/10.1001/jama.2010.297

Stringhini, S., Dugravot, A., Shipley, M., Goldberg, M., Zins, M., Kivimäki, M., Marmot, M., Sabia, S. & Singh-Manoux A. *(2011).* Health behaviours, socioeconomic status, and mortality: further analyses of the British Whitehall II and the French GAZEL prospective cohorts. *PLoS Med;8*(2):e1000419. https://doi.org/10.1371/journal.pmed.1000419

Stringhini, S., Tabak, A.g., Akbaraly, T.N., Sabia, S., Shipley, M.J., Marmot, M.G., Brunner, E.J., Batty, G.D., Bovet, P. & Kivimaki, M. *(2012).* Contribution of modifiable risk factors to social inequalities in type 2 diabetes: prospective Whitehall II cohort study. *BMJ;345*:e5452. https://doi.org/10.1136/bmj.e5452

Stringhini, S., Batty, G.D., Bovet, P., Shipley, M.J., Marmot, M.G., Kumari, M., Tabak, A.G. & Kivimaki, M. *(2013).* Association of lifecourse socioeconomic status and chronic inflammation and type 2 diabetes risk: the Whitehall II prospective cohort study. *PLoS Med;10*(7):e1001479. https://doi.org/10.1371/journal.pmed.1001479

Stringhini, S., Carmeli, C., Jokela, M., Avendano, M., Muennig, P., Guida, F., Ricceri, F., d'Errico, A., Barros, H., Bochud, M., Chadeau-Hyam, M., Clavel-Chapelon, F., Costa, G., Delpierre, C., Fraga, S., Goldberg, M., Giles, G.G., Krogh, V., Kelly-Irving, M., Layte, R., Lasserre, A.M., Marmot, M.G., Preisig, M., Shipley, M.J., Vollenweider, P., Zins, M., Kawachi, I., Steptoe, A., Mackenbach, J.P., Vineis, P., Kivimaki, M. *(2017).* Socioeconomic status and the 25x25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women. *The Lancet;* 389(10075): 1229-1237. https://doi.org/10.1016/S0140-6736(16)32380-7

Tehranifar, P., Wu, H.U., Fan, X., Flom, J.S., Ferris, J.S., Cho, Y.H., Gonzalez, K., Santella, R.M. & Terry, M.B. *(2012).* Early life socioeconomic factors and genomic DNA methylation in mid-life. *Epigenetics;8*(1):23-7. https://doi.org/10.4161/epi.22989

Tung, J., Barreiro, L.B., Johnson, Z.P., Hansen, K.D., Michopoulos, V., Toufexis, D., Michelini, K., Wilson, M.E. & Gilad, Y. *(2012).* Social environment is associated with gene regulatory variation in the rhesus macaque immune system. *Proc Natl Acad Sci U S A ; 109*(17): 6490-5. https://doi.org/10.1073/pnas.1202734109

van der Pol, M. & Cairns, J.A. *(2011).* Negative and zero time preference for health. *Health Econ*;20:917-29. https://doi.org/10.1002/hec.1655

van Doorslaer, E., Masseria, C., Koolman, X. & the OECD Health Equity Research Group. *(2006).* Inequalities in access to medical care by income in developed countries. *CMAJ;174*:177-83. https://doi.org/10.1503/cmaj.050584

Vineis, P. & Wild, C.P. *(2013).* Global cancer patterns: causes and prevention. *Lancet. .* pii: S0140-6736(13)62224-2.

Westerlund, H., Kivimaki, M., Singh-Manoux, A., Melchior, M., Ferrie, J.E., Pentti, J., Jokela, M., Leineweber, C., Goldberg, M., Zins, M. & Vahtera, J. *(2009).* Self-rated health before retirement in France (GAZEL): a cohort study. *Lancet;374*(9705):1889-96. https://doi.org/10.1016/S0140-6736(09)61570-1

Zins, M., Bonenfant, S., Carton, M., Coeuret-Pellicer, M., Gueguen, A., Gourmelen, J., Nchtigal, M., Ozguler, A., Quesnot, A., Ribet, C., Rodrigues, G., Serrano, A., Sitta, R., Brigand, A., Henny, J. & Goldberg, M. *(2010).* The CONSTANCES cohort: an open epidemiological laboratory. *BMC Public Health*;10:479. https://doi.org/10.1186/1471-2458-10-479

Endnotes

i Harri Alenius, Mauricio Avendano, Henrique Barros, Murielle Bochud, Cristian Carmeli, Luca Carra, Raphaele Castagne, Marc Chadeau-Hyam, Francoise Clavel-Chapelon, Giuseppe Costa, Emilie Courtin, Michaela Dijmarescu, Cyrille Delpierre, Angelo D'Errico, Pierre-Antoine Dugue, Paul Elliott, Silvia Fraga, Valerie Gares, Graham Giles, Marcel Goldberg, Dario Greco, Allison Hodge, Michelle Kelly-Irving, Piia Karisola, Mika Kivimaki, Vittorio Krogh, Thierry Lang, Richard Layte, Benoit Lepage, Johan Mackenbach, Michael Marmot, Cathal McCrory, Roger L. Milne, Peter Muennig, Wilma Nusselder, Salvatore Panico, Dusan Petrovic, Silvia Polidoro, Martin Preisig, Olli Raitakari, Ana Isabel Ribeiro, Fulvio Ricceri, Oliver Robinson, Jose Rubio Valverde, Carlotta Sacerdote, Roberto Satolli, Gianluca Severi, Terrence Simmons, Silvia Stringhini, Rosario Tumino, Anne-Clare Vergnaud, Paolo Vineis, Petter Vollenweider, Marie Zins.

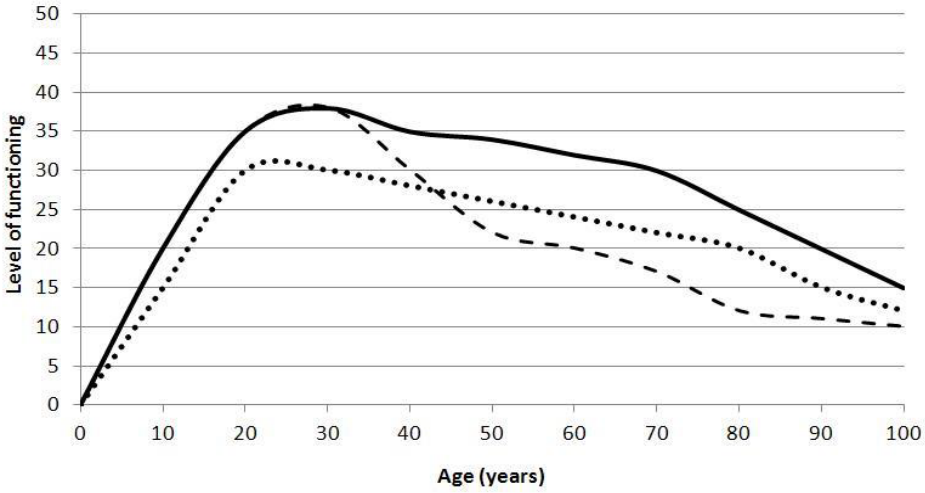


Figure 1. The “build-up” and “decline” model.

The "Build-up" and " Decline " Model of life course functioning. The solid line shows the optimal level of life course functioning. Dotted line shows sub-optimal development during the ‘build-up' phase while the dashed line shows increased rate of loss of function during the ‘decline’ phase

Table 1. LIFEPATH: cohorts included in the consortium. Study design/cohort description

Cohort	Study design	Outcomes and clinical tests	Already available/funded biomarkers
Whitehall II (UK) N=10,308 Kivimaki, M., Stringhini, S. et al, PLoS Med 2013	The Whitehall II study was established in 1985 to examine the socioeconomic gradient in health among 10,308 London-based civil servants (6,895 men and 3,413 women) aged 35–55. Baseline examination (phase 1) took place during 1985–1988, and involved a clinical examination and a self-administered questionnaire containing sections on demographic characteristics, health, lifestyle factors, work characteristics, social support, and life events.	Mortality, cancers, CVD events. BP, diabetes, anthropometry, physical functioning, cognitive function, mental health, hospitalisations.	Genotype (metabochip, cardiochip); metabolomics (NMR)(n=6,600); repeat data on lipids, glucose, insulin, inflammatory markers (CRP, IL6), cortisol (saliva, hair)(n=6,000-10,000). Subset: Brain MRI
TILDA (Ireland) N = 8,175 Layte, R., Kearney et al. Int. J. Epidemiol. 2011 40 (4): 877-884 Cronin et al., J Am Geriatr Soc 2013 61 Suppl 2:S269-78	The Irish Longitudinal Study on Ageing (TILDA) is a large prospective cohort study examining the social, economic, and health circumstances of 8,175 community-dwelling older adults (3,744 men, 4,431 women) aged 50 years and older at the time of recruitment and resident in the Republic of Ireland. Baseline examination (phase 1) took place during 2009-2011 and participants are followed up biennially. There were three components to the survey. Respondents completed a computer-assisted personal interview (n = 8,175) and a separate self-completion paper and pencil module (n = 6,915) which collected information that was considered sensitive. All participants were invited to undergo a separate health	BP, heart rate variability, pulse wave velocity, cerebral perfusion, balance, gait speed, osteoporosis, renal function, cognitive function, physical function, accelerometry, muscle mass, muscle strength, retinal imaging, macular pigment density, sensory performance, anthropometry mental health, quality of life, hospitalisation, mortality, cancers, diabetes, CVD events,	Lipids (N=5,800). Funded: Vitamin B12, Vitamin D, CRP, Creatinine, Folate, HbA1c (N=5,800). Subset: Brain MRI

	assessment at one of two national centers using trained nursing staff (n=5,897). A total of 5036 respondents completed the health-centre based assessment and a further 861 respondents completed a home-based assessment which involved a reduced set of tests. A more detailed exposition of study design, sample selection and protocol is available elsewhere (Whelan & Savva, 2013).	medical history, medications.	
Generation 21 (Portugal) **N=8,647** Barros, H., Correia, S. et al. BMJ Open 2013 Larsen PS et al. Paediatr Perinat Epidemiol 2013	Generation 21 (G 21) comprises a cohort of 8,647 newborns recruited in 2005–2006 in the Porto Metropolitan Area, in northern Portugal. Recruitment occurred at the 5 public maternity units, which are responsible for 95% of all births in the region (remaining births occurred at private hospitals)[21]. During the hospital stay, women delivering live births were invited to participate, and 92% of mothers agreed. All who agreed were invited to be re-evaluated at child's 4 years of age (2009-2011) , at 7 (2012-2014) and then again at 10 years (2015-2016).	Self-reported health status (health conditions), Tetrapolar Bioimpedance, ECG , Blood Pressure, Spirometry , DEXA, cognitive test.	CRP, lipids, glucose, insulin (4yrs n=1,530; 7yrs n=4,500)
The Airwave Health Monitoring Study (UK) **N=45,596** Paul Elliott http://www.police-health.org.uk/	Airwave Health Monitoring Study is an ongoing long term epidemiological occupational cohort study open to all police forces in UK. The study's current main goal is to investigate any possible impacts of TETRA on health by looking at TETRA exposure and subsequent health amongst police officers and staff. The study was launched in 2004, and since its inception, 28 out of 54	Cancer, death, hospitalisations. Sickness absences, blood pressure, ECG, arterial stiffness, anthropometry, cognitive test on sub-samples.	Blood count, haemoglobin, urea, creatinine, Gamma GT, lipids, glucose, HBA1c, C-peptide, CRP, fibrinogen, prothrombin time (N=35,000) Genotyping (N=17,000) Metabolome (N=3,000)

	police forces have agreed to participate. The cohort aims to have 60,000 participants by 2018. By December 2010, the cohort had recruited 42,057 participants out of which 34,957 have undergone a health screening that includes extensive lifestyle and questionnaire data, clinical measurements and collection of biological samples.		
EPIPORTO (Portugal) N=2,485 Barros, H., Alves, L. et al. BMC Public Health 2013	The EPIPorto is a general adult population-based cohort established with the initial aim of evaluating the major determinants of cardiovascular health. For this purpose, 2,485 (949 men and 1,536 women) adult dwellers in Porto, aged 18 years or over, were recruited between 1999 and 2003 using random digit dialling. The follow-up studies were performed for the whole sample in 2005-2008, and in 2014-2015. At all waves, information was collected using questionnaires administered by trained interviewers, self-administered questionnaires and objective measurements were made, including physical examination and blood tests (Ramos et al, 2004; Pereira et al, 2012).	Mortality, cancers, CVD events. BP, diabetes, anthropometry, ECG, Musculoskeletal conditions.	CRP, lipids, glucose, insulin, fibrinogen
SKIPOGH study Switzerland (3 centres) N=1,128 Bochud, M., Ponte, B. et al. Hypertension 2014	The Swiss Kidney Project on Genes in Hypertension is a longitudinal family-based study, following the standardised EPOGH (European Project on Genes in Hypertension) protocol. Baseline examination was conducted between 2009 and 2013. Three-year follow-up examination started in 2013 and is currently ongoing (expected to be finished in 2015). The aim of the study is	Deaths, CVD events, diabetes, hypertension, 24h BP, anthropometry, ECG, arterial properties, retinal imaging, renal function, chronic kidney disease, kidney morphology, cardiac imaging (3D); hand-grip	Methylome (Illumina 450K) (N=250). Transcriptomics (N=250). Inflammation markers, CRP, lipids, fasting glucose, insulin. Urine Na, K, urea, steroid metabolites (N=1,100). Cardiometabochip (200K SNPs)(N=1,100).

	to explore the role of genes and kidney haemodynamics in blood pressure regulation and kidney function in the general population. From December 2009 to March 2013, adult participants were recruited in two regions (Berne and Geneva) and one city (Lausanne) of Switzerland. A random sample of the inhabitants was drawn using different strategies. Inclusion criteria were (1) having a minimum age of 18 years; (2) being of European ancestry; (3) having ≥one and ideally three first-degree family members willing to participate; and (4) providing a written informed consent. Pregnant or breastfeeding women were not included. The general participation rate was 27.1%. At baseline, we collected data on cardiovascular and metabolic risk factors as well as on the prevalence of kidney and cardiovascular diseases. During follow-up, we are collecting data on new kidney and cardiovascular events. The primary endpoints are fatal and non-fatal strokes, ischaemic heart disease, heart failure and chronic kidney disease. The 1,128 participants (537 men and 591 women) belong to 272 nuclear families.	strength.	
COLAUS Study N=6,733 Vollenweider, P., Waeber, G., Firmann, M. et al. BMC	The CoLaus (COhorte LAUSannoise) is an ongoing prospective study assessing the clinical, biological and genetic determinants of cardiovascular disease in the city of Lausanne, Switzerland (Firmann et al., 2008). The initial survey was conducted between 2003 and 2006 and	Deaths, CVD events, diabetes, hypertension, BP, anthropometry, physical and cognitive functioning, FFQ, PAFQ.	Inflammation markers (N=6,300), CRP, lipids, fasting glucose, insulin, vitamins, genotype (500 K Affymetrix chip technology)

Cardiovascular Disorders 2008	included 6,733 participants aged between 35 and 75 years; the first follow-up survey was conducted 5.5 years afterwards and included 5,064 participants. In each survey, data on socio-economic status, lifestyle, mental status and cardiovascular risk factors is collected by questionnaire or clinical examination.	N=2,000 polysomnography (Hypnolaus) N=3,500 psychiatric examination (Psycholaus)	
Growing Up in Ireland (Ireland) **N = 19,702** Layte, R., Williams, J. et al, Dept. of Children and Youth Affairs (2009)	Longitudinal child cohort established in 2007/2008. The study takes place over seven years and follows the progress of two groups of children: 1)containing 11,134 participants aged 9 months at baseline with three study waves (9 months, 3 years and 5 years) and 2) containing 8,568 participants aged 9 years at baseline (9 years, 13 years). The dataset gathered contains multiple SES measures of parents at each wave and economic strain and anthropometric measures at each wave of child and parents.	Chronic illness of parents and children and whether confirmed by physician. Acute illness for infants at baseline. Child psychological adjustment, cognitive function, parent/child relationship, child self-concept.	Measured heights and weights for parents and children at each wave. Measured head circumference for 9 month cohort at baseline.

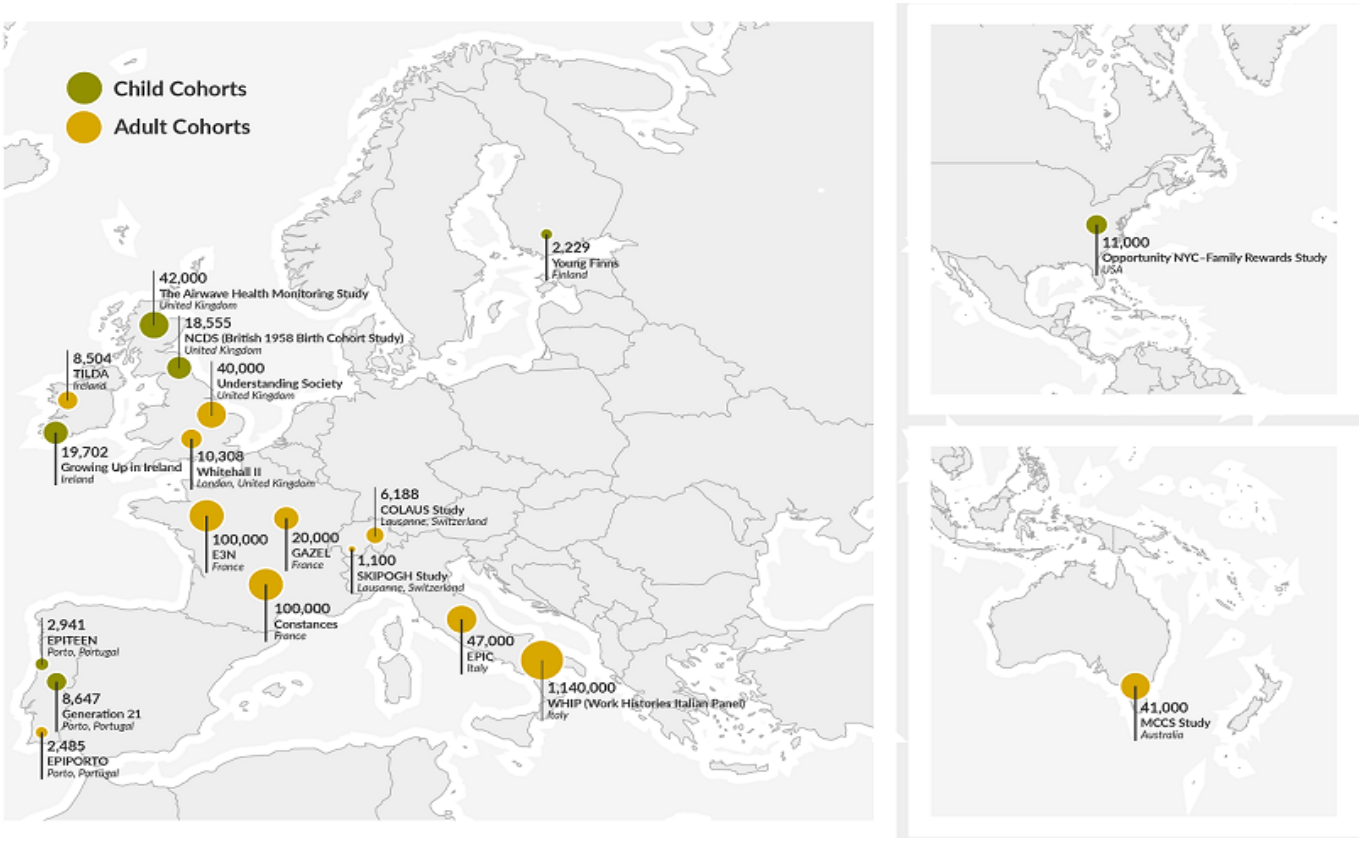


Figure 2. Map of the adult and child cohorts participating in LIFEPATH

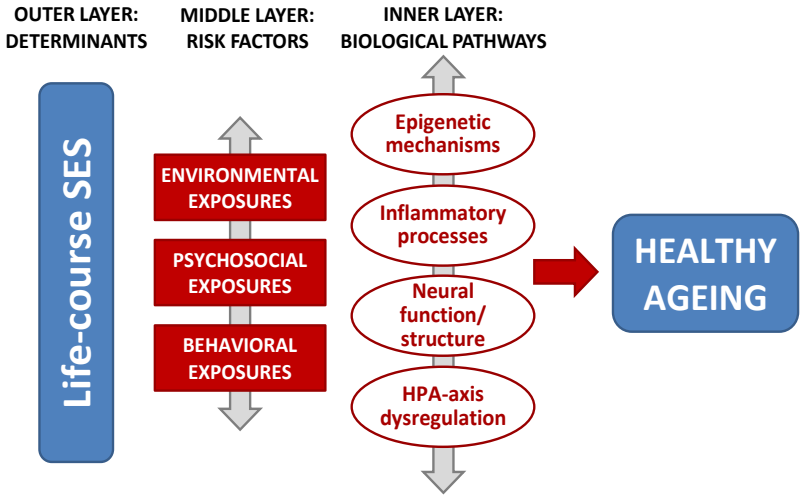
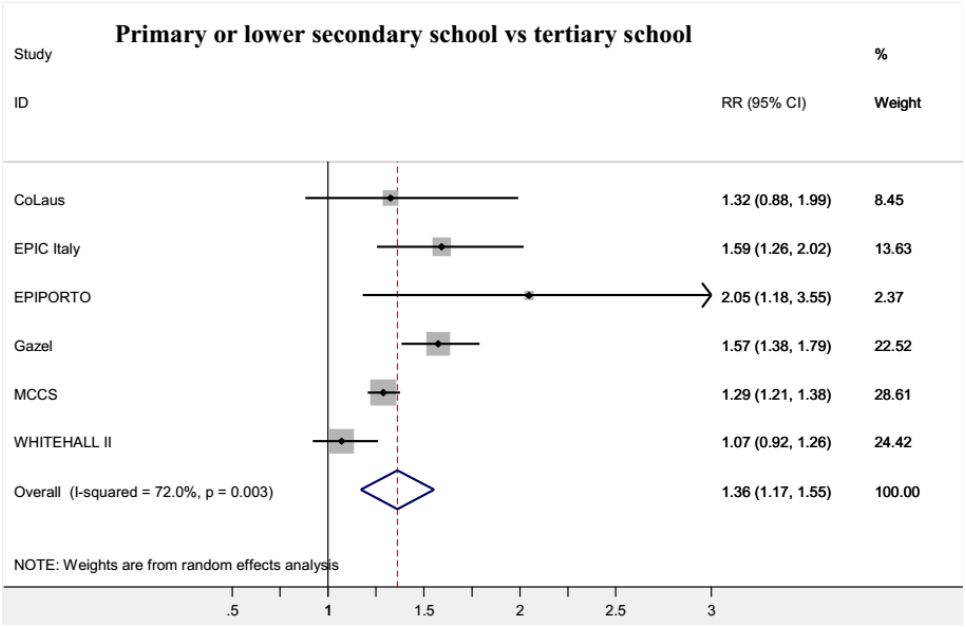


Figure 3. The role of biomarkers in biological pathways leading from SES to healthy ageing

Table 2. Statistical analyses involved in LIFEPATH

Task	Aim/ Data	Expected outcome relevant to LIFEPATH
SES and exposure variables vs health outcomes	**Aim**: Harmonise and analyse different indicators of SES available in the cohorts; build a composite indicator of SES; analyse exposure variables. Harmonise and analyse health outcomes and build indicators of healthy ageing	Analysis of all cohorts yielding estimates of: - Association between SES measures and healthy ageing variables - Association between SES variables and risk factors - Adjustment of SES-outcome relationships by risk factors
Identification of internal markers of SES	**Aim:** Identify in the untargeted –omic profiles which candidates are correlated to SES	List of putative internal markers of SES (methodologically validated)
Implementation of Life-course disease risk models	**Aim:** Identify markers of SES/exposures also relating to the healthy ageing outcomes, and elucidate how their effect is mediated by critical stages in life (early vs late)	Quantified predictive abilities of the validated markers. Estimates of the age-related susceptibility functions, and identification of potential critical age ranges at which SES/exposures are influential with regard to the disease risk.
Burden of disease calculations SES-based risk score	**Aim:** Estimate the Burden of disease using refined risk estimates based on SES, external exposures and omics. **Aim:** Develop a prototype for a SES-based score for the prediction of unhealthy ageing.	Updated estimates of the burden of disease for selected exposures. To assess whether the additional predictive ability associated with SES measures can be incorporated into policies and clinical work.

MEN



WOMEN

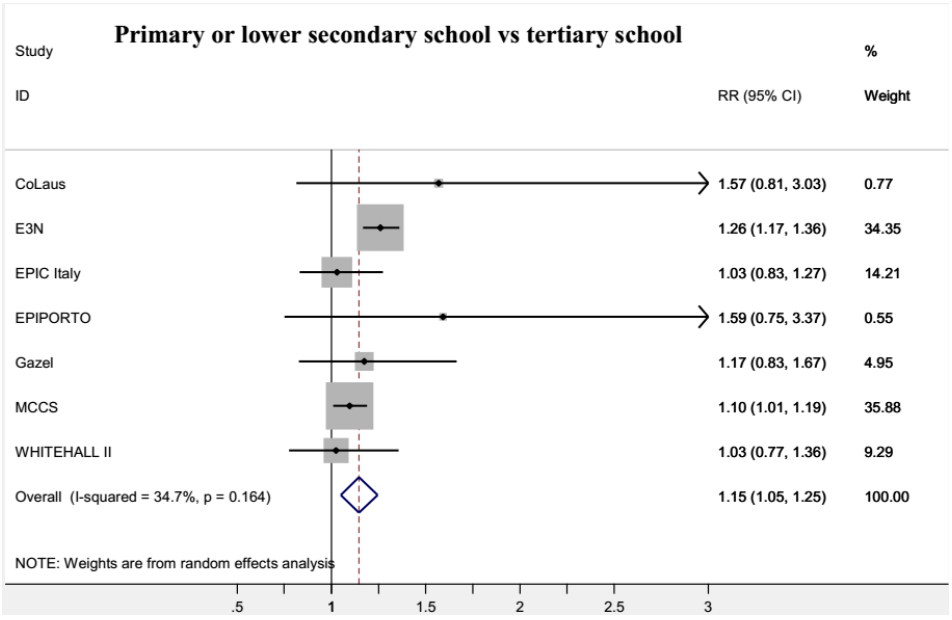


Figure 4. Meta-analysis of the association of mortality with education, by gender (d'Errico et al., PLOS ONE, *in press*)

Appendix

Other large longitudinal cohorts with biological samples (n=202,555).

Cohort	Study Design	Outcomes and clinical trials	Already available/ funded biomarkers
EPIC- Italy **N=47,749** Vineis, P., Gallo, V. et al, PLoS One 2012	The European Prospective Investigation into Cancer and Nutrition (EPIC) is a large European study on diet and cancer. The Italian component of EPIC EPIC-Italy) recruited 47,749 adult volunteers (men and women) at five centres: Varese and Turin in northern Italy, Florence in central Italy and Naples and Ragusa in southern Italy. All participants signed an informed consent form and completed two questionnaires: one about dietary habits (food-frequency) and one about lifestyle, with information on education, socioeconomic status, occupation, history of previous illnesses and surgery, lifetime tobacco use and alcohol consumption and physical activity. EPIC database records were linked to cancer and regional mortality registries after EPIC database quality control. All EPIC-Italy centres except Naples are covered by population-based cancer registries. In Naples, follow-up information was collected from electronic hospital discharge records and also by periodic personal contact with participants. We are including in LIFEPATH centres from EPIC Italy (N=34,148) with the exclusion of Florence.	All cancers, CVD, diabetes, BP.	Methylome (Illumina 450K) for >1,000 subjects. Inflammation markers, CRP, lipids, in subsets

E3N (France) **N=98,995** Clavel-Chapelon, F., Dartois et al, 2014	The E3N study (Étude Epidémiologique auprès de femmes de la Mutuelle Générale de l'Education Nationale) is a prospective cohort of 98,995 women aged 40-65 years at recruitment in 1990 and it constitutes the French component of EPIC. The study was established to produce a large mass of data relevant for the identification of environmental and molecular causes of cancer and other chronic diseases, and to contribute to the development of effective public health strategies. Data on residence, education, early life events, exposures, lifestyle factors and life events were collected at baseline and follow-up questionnaires were sent every 2-3 years with more than 80% response. Mortality data were obtained by means of regular record linkage with various French mortality databases.	Self-reported health outcomes at each questionnaire**.** All cancers (verified), CVD, diabetes, asthma, osteoporosis, depression, migraine	Metabolomics data on 800 breast cancer cases, 800 controls. B-vitamins 1,000 samples, 1,800 samples, Fatty acids 1,700 samples
NCDS (British 1958 Birth Cohort Study) **N= 18,555** Goodman, A., Hypponen et al, 2005	The National Child Development Study (NCDS) is a birth cohort established in 1958, which follows the lives of over 17,000 people born in England, Scotland and Wales in a single week in 1958. The participants to the study are followed up at ages 7, 11, 16, 23, 33, 42, 46 and 50. A biomedical survey (for 9,377 cohort participants) was conducted when participants were aged 44-46 years (N=18,555).The datasets contains prospective SES adult life and early life.	Self-reported health outcomes at each sweep (CVD, Cancer etc). All-cause mortality from register to 2008. Biomedical data from measurements (BP; lung function; BMI)	Biomarkers from biomedical survey at age 44-46y (lipids, IGE, HbA1c, IGF, CRP, Fibrinogen, Von Willebrand). Subset: metabolomics data (N=8,000).
Understanding society	*Understanding Society* is an innovative UK household	Self-reported health data on	Blood sample, Saliva sample

N=40,000 households Buck, N., Shiue et al, 2014	longitudinal panel study established in 2008 (population-based households sample). This study captures information about the 21st century UK life, looking specifically at its participants' social and economic circumstances, attitudes, behaviours and health. Information is also collected on all new household members. The data contains information related to SES adult and early life.	20,000 individuals. Biomedical data from nurse visit (6,104).	(details of analytes not released yet)
Gazel (France) **N= 20,625** Goldberg, M., Zins, M., Westerlund, H. et al, Lancet 2009	The GAZEL study was established in 1989 among employees of the French national gas and electricity company, Electricité de France-Gaz de France (EDF-GDF) (Goldberg et al., 2007, 2015). At baseline (1989), 20,625 employees (15,011 men and 5,614 women), aged 35–50, gave consent to participate. The study design consists of an annual questionnaire used to collect data on health, lifestyle, individual, familial, social and occupational factors, and life events.	Mortality, cancers, CVD events. Diabetes, asthma, depression, sleep disturbances, migraine, quality of life, cognitive and physical functioning. Systematic individual linkage to health insurance, hospital discharge, retirement and mortality national database.	Blood samples (serum, plasma, buffy-coat) for about 5,000 subjects
Constances **N (by 2017)= 200,000** **May 2016 N: 100,000** Zins, M., Goldberg, M. and Berkman, L., et al, BMC Public Health. 2010	The CONSTANCES cohort was established in late 2012 (Zins et al., 2015). It was designed as a randomly selected representative sample of French adults aged 18-69 years at inception; 200,000 subjects will be included over a five-year period. At enrolment the participants fill questionnaires collecting data on health, lifestyle, individual, familial, social and occupational factors, and life events and benefits from a comprehensive health examination. The follow-up includes a yearly self-	Mortality, cancers, CVD events. Diabetes, asthma, depression, sleep disturbances, migraine, quality of life, cognitive and physical functioning. Systematic individual linkage to health insurance, hospital discharge, retirement (occupational history) and mortality national	Blood count, glucose, lipids, creatinine, GGT (N=35,000)

	administered questionnaire, a health examination every 5 years and an annual linkage to social and health national databases.	databases.	
MCCS study (Cancer Council Victoria, Melbourne) **N=41,514** Giles, G., Hodge, A. et al, 2013	The Melbourne Collaborative Cohort Study is a prospective cohort study of 41,514 participants (24,469 women) living in Melbourne, Australia. Caucasian volunteers aged between 40 and 69 years were recruited in randomly selected census districts. At baseline (1990–1994), participants attended clinics where demographic, lifestyle and dietary information were collected and anthropometric measurements were performed.	Verified cancer diagnoses, deaths with cause of death, CVD, diabetes, obesity (through physical measures at second visit)	Methylome (Illumina 450K) for 3,000 subjects. Inflammation markers for 500 controls, fatty acids for 1,000 controls. Glucose and lipids for 23,000.
EPITEEN (Portugal) **Henrique Barros** **N=2,942** Ramos, E. et al. ActaPaediatr 2007	Epidemiological Health Investigation of Teenagers in Porto is a random sample of 2,942 adolescents enrolled at public and private schools in Porto, Portugal. The first wave was completed in 2003-2004 when participants were aged 13 years. The second wave in 2007-2008 (17 years), the third wave in 2011-2013 (21 years) and the fourth wave in 2014-2015 (24 years).	Self-reported health outcomes (diseases diagnosis), BP, Anthropometry, Spirometry, Bioelectrical impedance , Bone health (forearm dual-energy X-ray absorptiometry), ECG (21yrs)	CRP, lipids, glucose, insulin
The Cardiovascular Risk in Young Finns Study (Finland) **Olli Raitakari** **N=3,596** Raitakari, O. et al. JAMA 2003	The Cardiovascular Risk in Young Finns Study (YSF) collected data on 4,320 children and adolescents aged 3, 6, 9, 12, 15 and 18 years at baseline who were randomly chosen from the population register of five Finnish cities with universities with medical schools to produce a representative sample of Finnish children. Of those invited (N=4,320), 3,596 (83%) participated in the first cross-	Morbidity, mortality, self-reported health outcomes, BP, Anthropometry, Cognitive functioning	CRP, lipids, glucose, insulin, GWAS, metabolomics

sectional study in 1980. Between 1980 and 1992, these cohorts were followed up at 3-year intervals. The follow-up field studies were performed for the whole study population in 1983 and 1986, when 2,991 (83.2%) and 2,799 (78.3%) subjects participated. In 1989 and 1992 questionnaires were sent to every individual, but only a subset of individuals had their height measured in these survey years so we omit these observations from the analysis. In 2001, 2,283 subjects (63.5%) of the original cohort participated in clinical examinations. A detailed description of the sample can be found elsewhere[22]. For the purposes of this analysis, we treat YFS as 6 different cohorts representing 3 year age bands (3, 6, 9, 12, 15, and 18 years of age upon recruitment into the study).

In addition to the cohorts shown in the table we will also use a large cohort called **Work History Panel (WHIP)**. WHIP is based on a sample of individual-level data from the Social Security Administration archives in Italy, covering almost 8% of all Italian workers employed in the private sector in 1985-2010. Unlike all other cohorts in Table 1 it does not have biological samples, but it provides very rich information on income, pensions, unemployment benefits, disability indemnities, workplace and job contracts, linked to hospital and mortality follow-up (see box below).

Cohort	Study design	Outcomes and clinical tests	Already available/funded biomarkers
Work History Panel (WHIP) - Health Italy - N=1,364,922 Costa, G., Leombruni, R. et al, Epidemiol & Prev 2010	The Work History Panel is based on a sample of individual-level data from the Social Security Administration archives in Italy, covering almost 8% of all Italian workers employed in the private sector in 1985-2010. Unlike all other cohorts in LIFEPATH, it does not have biological samples, but it provides very rich information on income, pensions, unemployment benefits, disability indemnities, workplace and job contracts, linked to hospital and mortality follow-up.	Mortality and hospitalisation by cause (2001-2012).	Not available.

AUTHOR GUIDELINES SUMMARY

Submission of Papers

All papers, written in the English language, should be submitted via the *LLCS* website as a Microsoft Word file. If there is a good reason why this is not possible, authors are requested to contact crandall@slls.org.uk **before** submitting the paper. All subsequent processes involving the author are carried out electronically via the website.

Preparation of Texts

Length. The paper should normally be **approximately 5,000 words**, with longer papers also accepted up to a maximum of 7,000 words. This word count excludes tables, figures and bibliography.

Font and line spacing. Please use Calibri (or similar sans serif) font, 12pt, with 1.5 line spacing in all main text, single space in figure or table captions.

Page layout. All text should be justified to the left hand margin (all margins of at least 2.5cm) with no indents at the start of paragraphs and a line space separating paragraphs. All headings and sub-headings used within the body of the paper should also be left justified. Please do **NOT** use automatic text formatting.

Weblinks. To help our readers to look up cited references or other information available on the web, authors should ensure that all such references are activated.

DOIs. Do NOT include DOIs in the bibliography – they are added during editing as resolvable URLs.

Ensuring an anonymous (blind) review. Please submit papers with a full detailed title page. Once a paper has been submitted to *LLCS* via the website, it will be 'anonymised' by the removal of all author name(s) and institution names on the title page, and any identifying electronic document properties will also be removed. Authors do not need to remove their name(s) from the main text or references but any reference to their work or themselves should be expressed in the third person.

Abstract. The abstract (no subheads or paragraphs) should be no more than 250 words (not part of the main word count).

Keywords. These should be included just below the author list (minimum 3 maximum 10).

Abbreviations. Words to be abbreviated should be spelt out in full the first time they appear in the text with the abbreviations in brackets. Thereafter the abbreviation should be used.

References. Please use the APA 6th edition and refer to examples in the full Guidelines.

Authors not complying with these reference guidelines will be asked to make all the necessary alterations themselves, if their paper is accepted for publication.

Notes. As a general rule, supplementary notes should be avoided, but if thought to be essential, they should *not* appear on the page as Footnotes, but instead, be included as Endnotes.

Supplementary material. Supplementary material may be uploaded with the submission, and if the paper is published, will be visible to the reader via a link on the RHS of the screen.

Submissions containing graphs, tables, illustrations or mathematics. All graphs, tables and illustrations should be embedded in the submitted text, and have clear, self-explanatory titles and captions. Mathematical expressions should be created in Word 2003 (or a compatible package) with equation editor 3.0, unless the author has good reason to use other software, in which case please contact crandall@slls.org.uk. All biological measures should be reported in SI units, as appropriate, followed, in the text, by traditional units in parentheses.

Author citation. If the paper is accepted for publication, a version of the paper with all authors cited in full on the title page will be used. Only individuals who have contributed substantially to the production of the paper should be included.

Copy editing. All accepted manuscripts are subject to copy editing, with reference back to author with suggested edits and queries for response.

Proofs. The corresponding author will be asked to view a layout proof of the article on the website and respond with final amendments within three days.

(Full Author Guidelines at: http://www.llcsjournal.org/index.php/llcs/about/submissions#authorGuidelines**)**

Open Journal System

The ***LLCS*** journal is produced using the Open Journal System, which is part of the Public Knowledge Project. OJS is open source software made freely available to journals worldwide, for the purpose of making open access publishing a viable option for more journals. Already, more than 8,000 journals around the world use this software.

INTRODUCTION

PAPERS

RESEARCH NOTE

STUDY PROFILE

LLCS Journal can be accessed online at www.llcsjournal.org

Published by the Society for Longitudinal and Life Course Studies
info@slls.org.uk www.slls.org.uk