Statistical analysis of life history calendar data

Mervi Eerola* and Satu Helske†

August 22, 2012

Abstract

The life history calendar is a data-collection tool for obtaining reliable retrospective data about life events. To illustrate the analysis of such data, we compare the model-based probabilistic event history analysis and the model-free data mining method, sequence analysis. In event history analysis, we estimate instead of transition hazards the cumulative prediction probabilities of life events in the entire trajectory. In sequence analysis, we compare several dissimilarity metrics and contrast data-driven and user-defined substitution costs. As an example, we study young adults' transition to adulthood as a sequence of events in three life domains. The events define the multistate event history model and the parallel life domains in multidimensional sequence analysis. The relationship between life trajectories and excess depressive symptoms in middle-age is further studied by their joint prediction in the multistate model and by regressing the symptom scores on individual-specific cluster indices. The two approaches complement each other in life course analysis; sequence analysis can effectively find typical and atypical life patterns while event history analysis is needed for causal inquiries.

Keywords: Distance-based data; Life course analysis, Life history calendar; Multidimensional sequence analysis; Multistate model; Prediction probability

1 Introduction

Follow-up studies, which register prospective events in time, are the golden standard of reliable data collection in developmental studies and life course analysis. Yet these can be expensive and sometimes difficult to perform. Retrospective data collection is used mainly when a very large sample is required, the classical example being rare outcomes and case-control designs. Recently, however, retrospective data collection has been used in survey studies to obtain detailed information about multiple life domains and individuals' multiple activities.[1] *The life history calendar* (LHC), also called an event-history calendar, is a data-collection tool for obtaining reliable retrospective data about life events.[2] The advantage of a life history calendar is that the order and proximity of important transitions in multiple life domains can be studied at the same time. The time window of a life history calendar can be years or even an entire life-span. As a data collection tool, it encourages respondents to incorporate temporal changes as cues in the reporting of events. It has shown the ability to provide data of remarkably high quality.[1]

While life course epidemiology studies the relationship between exposure and disease, problems of special interest to psychologists and social scientists point to an understanding of individuals' behaviour and choices in their lives. These choices are often reflected in the amount of time devoted

*Department of Mathematics and Statistics, Assistentinkatu 7, 20014 University of Turku, Finland; tel: +358-2-3335437, +358-40-5622913; email: mervi.eerola@utu.fi

†Methodology Centre for Human Sciences/Department of Mathematics and Statistics, University of Jyväskylä

to different activities. Individuals also have several social roles in their lives, and in these roles they share values and resources which may form their decisions and experiences in a similar way. These links have been of interest especially in life course studies carried out by social scientists. Linking different life domains (e.g. education, family formation, health, working life) of a single individual is an effort to study the life course as an interdependent system of life processes, and makes the analysis multidimensional and dynamic at the same time. This is the focus of our article when evaluating methods for the statistical analysis of life history calendar data. We believe that the approach taken by sociologists and psychologists can be valuable also to health scientists. Variable life patterns can have effects, for example, on chronic diseases or on patients' differential response to clinical treatments.

Traditionally, life course data have been analysed by event history methods. There is a vast literature on the basic principles and on more advanced methods based on the theory of counting processes (e.g. Andersen et al.[3]). These methods are valuable when studying the time course of a few well-specified life events but when the number of states, and accordingly the number of transitions between the states, increases, joint analysis of the model especially for prediction purposes becomes rather elaborate. In this article, we compare two approaches to life course analysis: model-based probabilistic event-history analysis (EHA) and a more recent type of approach of model-free data-mining, sequence analysis (SA). The latter is well known in bioinformatics but has provided novel insight to the diversity of life trajectories and their relationship to life satisfaction and depressiveness. We emphasize the differences, but also the complementary tasks of the methods. As an example, we study young adults' pathways to adulthood and consequent depressive symptoms in middle age in a cohort established in Central Finland in 1968. The cohort members have been followed for 42 years, from age 8 until age 50.

The article is structured as follows. In Section 2, some concepts and principles of prospective and retrospective approaches to life course analysis are contrasted, the first in terms of predictive probabilities and the latter in terms of typologies of life sequences. Section 3 provides comparative analysis of the cohort data and some sensitivity analysis. Finally, Section 4 presents a methodological discussion about the different informational content of the two approaches.

2 Prospective and retrospective analysis of the life course

From a methodological point of view, the timing and order of events is of fundamental relevance in life course analysis. Events represent transitions, marking developmental stages in life, while the role and statuses accompanying such transitions feature the essential characteristics of the life course.[4] *Trajectories* are sequences of previously occupied life states, which provide a long-term view of usually one dimension of an individual's life course. *Transitions* between the states, which are of course embedded in the life trajectories, provide a short-term view of the dynamics of the life course. Historically, transitions have been more important concepts because they relate directly to important changes in life history.

Recently, more attention has been given to micro-settings and diversity of the dynamics involved in the individual's different activities, roles, and relationships. This change in scope has emphasized the analysis of whole trajectories instead of events. The role of transitions and trajectories as the basic unit of analysis is described in the next sections.

2.1 Event history analysis

Prospective analysis is based on short-term predictions of transitions in the life course. These predictions can be modified by some informative covariates $\mathbf{Z}$ which themselves may vary in time. A concise review of event history methods can be found, for example, in Andersen and Keiding.[5] Here we prefer, however, an approach based on a *marked point process* $(T, X) = \{(T_n, X_n), n \geq 1\}$. Rather

than a system of states accompanied with a transition matrix, we model the life course as a sequence of events by specifying a pair of random variables, the occurrence time T and a mark X identifying the event. An extensive overview of such models and theory is given by Arjas.[6]

Let $N_x(t) = \sum_{n\geq 1} 1\{T_n \leq t, X_n = x\}$ be a process counting x-specific events in an individual's life course such that $\sum_x N_x(t) = N(t)$ is the total number of life events by time t. Since life history calendar data is often recorded on a yearly basis, we define the discrete *event-specific hazard* in the age interval $t = 1, 2, ...$ as the conditional probability of a change in the value of N_x

$$p_x(t) = P(\Delta N_x(t) = 1 | \mathcal{F}_{t-1}^N) \tag{1}$$

given the internal history $\mathcal{F}_{t-1}^N$ of the counting process. We will denote the history of the occurrence times and marks by time t as H_t. The crude hazard that some event occurs in the interval t, regardless of which one, is the sum over the event-specific hazards, $p(t) = \sum_x p_x(t)$.

The likelihood contribution of an individual's life history can be interpreted as a product of a sequence of multinomial trials over the intervals. Since $\Delta N_x(t)$ can only have the value of 1 or 0 in a short interval t, the outcome of the multinomial trial within each interval can be read from its value. This determines which one of the x-components of N contributes to the likelihood. For a generic individual, the likelihood contribution by time t is

$$L(t) = \prod_{s\leq t} \prod_x p_x(s)^{\Delta N_x(s)} (1 - p(s))^{1-\Delta N(s)}. \tag{2}$$

While the hazard gives a very short-term prediction of the life course, the *prediction process* associated with a marked point process gives a long-term prediction of some random event related to (T, X) for the whole observed trajectory.[7–10] We can then view the prediction process as the conditional distribution of that random event given the history H_t. The prediction probabilities are again functions of event-specific hazards, so modelling the hazards brings external explanatory information to the prospective analysis of the whole life trajectory.

In Section 3.2, we shall consider in detail the specification and estimation of prediction probabilities in a multistate model. For a tutorial on event history analysis and prediction probabilities, we refer to Putter et al.[10] In the next section, we contrast the model-based predictions of life events, extended to the whole observed trajectory, with the model-free approach of sequence analysis. Since it is still less familiar than event history analysis to health scientists, we give a more extensive overview of its basic principles.

2.2 Life sequence analysis

A completely different approach is taken in sequence analysis (SA), originally used in bioinformatics to organize, classify, and parse protein and DNA sequence data.[11] In the 1980s, data mining methods were developed to analyse molecular sequences as texts (e.g. TGACT = Thymine-Guanine-Adenine-Cytosine-Thymine). Comparing sequences corresponds to comparing amino acids in protein sequences or nucleotides in DNA sequences at each position. The goal is to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. This is accomplished by aligning the sequences pairwise. Gaps are inserted between the elements so that identical or similar characters are aligned in successive columns. Mismatches between the sequences can have biological interpretations as point mutations and gaps as insertion or deletion mutations introduced in one or both lineages since their divergence from a common ancestor.

In the life course setting, sequence analysis was first introduced by the social scientist A. Abbott.[12] He criticized the event-oriented method as being unable to reveal life patterns when focusing only on isolated events. Aligning life sequences provided correspondences with similar life patterns, while mismatches and gaps corresponded to differential timing and/or a lack of certain life events or

Table 1: Basic differences of sequence analysis and event history analysis.

Method	Sequence analysis	Event-history analysis
Unit of analysis	sequence	event
Basic tool	distance matrix	transition rate
Direction of inference	retrospective	prospective
Mode of inference	static, unconditional	dynamic, conditional
Type of inference	alignment of sequences	comparison of rates
Aim of inference	population-level	individual-level

episodes. Studying trajectories as the basic units allowed them to be interpreted as connected series of experiences or summaries of lives, not isolated events.[13]

While event-history analysis models the risk of life events with explanatory covariates, sequence analysis aims at forming typologies of life trajectories based on their similarity and characterizing them by means of covariates. To assess similarity, pairwise distances of the sequences are first calculated. The distance matrix is then used as data for clustering to find similar life patterns. Table 1 summarizes the basic differences of the two methods. We notice that, from a statistical point of view, they have in many respects completely different approaches. One can expect that they also provide different types of information about the life course.

2.2.1 Probabilistic sequence analysis

We start with reviewing a probabilistic approach to SA to more clearly contrast the prospective and retrospective probabilistic life course analyses, and then focus on the non-probabilistic sequence analysis that has been used exclusively in life sequence analysis to date. We follow closely Durbin et al.[11] in the probabilistic SA presentation.

Sequence alignment depends on a scoring model, on the algorithm for optimizing the scoring, and on statistical methods to evaluate the goodness of the results. In a probabilistic scoring model, the *substitution score* measures the relatedness of sequences in the observed data with the expected case, where matching occurs only randomly at each position. The log odds ratio of the scoring models for the whole sequences compares the log of observed and "expected by chance" models.

Consider two sequences, x and y with lengths m_x and m_y. Let x_i be the symbol of ith site of x and y_j be the symbol of the jth site of y. In the case of DNA sequences, the symbols are elements of $\{A, T, C, G\}$ so there are $K = 4$ symbols. We want to assign a score to the alignment that measures the relative likelihood that the sequences are related as opposed to being unrelated. The unrelated scoring model assumes that a symbol, say a, occurs independently with frequency q_a. For sequences of equal length, the unrelated or random model is then of the form

$$P(x, y|R) = \prod_i q_{x_i} \prod_j q_{y_j} \tag{3}$$

whereas the related or match model M is the product of joint probabilities for the whole alignment

$$P(x, y|M) = \prod_i p_{x_i y_i}. \tag{4}$$

Here p_{ab} is the joint probability of elements a and b occurring as an aligned pair. The ratio of the models is the odds ratio

$$\frac{P(x, y|M)}{P(x, y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}}. \tag{5}$$

To have an additive score model, we take a logarithm of the odds ratio which can further be written as

$$log\frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}} = \sum_i log\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} = \sum_i s(x_i, y_i). \tag{6}$$

The substitution costs $s(a, b)$ for each aligned pair of elements can be arranged in a $K \times K$ matrix which gives a statement about the probability of ab occurring jointly. The probabilities p and q are based on biological theory.

An optimal *alignment algorithm* to minimize the total cost of dissimilarity (or maximize the score of similarity) is based on dynamic programming. Optimal matching (OM) computes generalized Levenshtein distances[14] by minimizing the cost of elementary operations: substitution, insertion, or deletion of an element. Insertions or deletions are jointly called *indels*. The cost of a gap (one or more conjoined indels) is often set as the length of the gap, but opening and extending gaps can also be given different weights. OM quantifies the effort needed to transform one sequence to another. Example 1 illustrates a possible alignment of two sequences and the OM operations needed to compute the cost.

Example 1

sequence 1: AAAABBBB
sequence 2: AAA–BBCC

The alignment above contains five matching elements, two mismatches, and a gap of length 1. For defining the cost of this alignment, sequence 2 is transformed to sequence 1 using an insertion of an element A and two substitutions of an element C with B (shown bold).

AAAABBBB → AAAABBBB → AAAABBBB
AAABBCC → AAA**A**BBCC → AAAABB**BB**

The cost of the alignment is the sum of the costs of the operations. Transforming sequence 1 would lead to exactly the same result. The best possible alignment with the lowest cost is found using dynamic programming.

A global alignment algorithm is, for example, the Needleman–Wunsch algorithm.[15] The idea is to build up an optimal alignment, using previous solutions for optimal alignments of smaller subsequences. To find the alignment with the lowest score, a matrix D is allocated. The value $D(i, j)$ is the score of the best alignment between the initial segments $x_{1...i}$ and $y_{1...j}$ and can be built recursively. First $D(0, 0) = 0$ is initialized. The matrix is then filled from top left to bottom right with

$$D(i,j) = \min \begin{cases} D(i-1, j-1) + s(x_i, y_j) \\ D(i-1, j) - o \\ D(i, j-1) - o, \end{cases} \tag{7}$$

where $s(x_i, y_j)$ is the cost of a substitution and o of an indel. In the first row x_i is aligned with y_j; in the second row x_i is aligned with a gap in y; and in the third row y_j is aligned with a gap in x. The best score up to (i, j) will be the smallest of these. The equation is applied repeatedly until the matrix is filled. The value in the final cell $D(m_x, m_y)$ is the best score for an alignment of x with y.

The significance of a particular alignment M can be assessed, for example, by Bayesian model comparison. The posterior of alignment M is

$$P(M|x, y) = \frac{P(x, y|M)P(M)}{P(x, y)}, \tag{8}$$

where $P(M)$ is the prior probability of M and $P(x, y|M)$ the likelihood of data given alignment M. The Bayes factor of the odds ratio is then

$$log\left(\frac{P(x,y|M)}{P(x,y|R)}\right) + log\left(\frac{P(M)}{P(R)}\right) \tag{9}$$

where R is the random model.

2.2.2 Non-probabilistic sequence analysis

In life sequence applications, only non-probabilistic sequence analysis has been used to date. These methods are either based on sequence editing and pairwise alignment of sequences as in the probabilistic case, or on counting common sequence attributes (non-alignment methods).

Substitution costs. The most important difference is that in pairwise alignment the substitution cost $s(x_i, y_j)$ is not a log odds ratio as in probabilistic SA but rather a given constant, defined by the analyst. At least three alternatives have been used. The first derives costs from *substantive theory* that often suggests some order between the states. Different substantive questions have of course different interpretations for the similarity of states. In social science applications, a theory-based cost matrix is often preferred because the timing of events and the similarity of states are considered conceptually separate issues (e.g. Halpin[16]).

Subjectivity in cost definition can be reduced by *data-driven costs* which are inversely proportional to transition frequencies from state A to B and B to A.[17,18] The time-independent cost of substituting A to B is then

$$2 - p(A, B) - p(B, A)$$

where $p(A, B)$ is the estimated proportion of transitions from A to B. The substitution cost is therefore symmetric.

A third alternative is to calculate pairwise distances from some theory-driven *prototypes.*[19]

Sequence alignment. When the cost matrix is defined, pairwise distances between the sequences are calculated as in probabilistic SA. Optimal matching (OM) algorithm, described in the previous section, has been used most often in life sequence analysis. However, changing the order of states with insertions and deletions (indels) have been criticized for warping the time in an unnatural way.[20,21] A generalization of the Hamming distance[22] is a special case of optimal matching where indels are not used, and thus only states at the same position (time) are aligned.

The assumption of independent positions within a sequence may be a reasonable approximation to reality in bioinformatics but unrealistic in life course analysis. Most of the criticism of life sequence analysis has been directed at the ignorance of time order, which is so fundamental in prospective analysis (e.g. Wu[23]). In the probabilistic setting, it would be natural to model a life sequence as a Markov chain and generalize the independent elements (iid) assumption by assuming homogeneity or non-homogeneity of the chain.

In non-probabilistic SA, several ad-hoc alternatives for *duration-dependence* have been proposed to account for the length of time spells in sequence comparison. Stovel et al.[24] used a decay-function, which depends on a specific index period. Halpin[16] suggested a variant of OM that makes substitutions and indels cheaper for long spells than short spells. Marteau[25] used time-warping, which locally compresses or expands the time-scale to minimize the distance to the other sequence. Lesnard[21] proposed a time-dependent cost matrix for each time unit, depending on the neighbouring states (dynamic Hamming distance).

Non-alignment metrics. Elzinga[26–28] has taken a completely different approach, based on combinatorial methods, which does not require any cost matrix. The distance between two sequences is

generally defined as

$$d(x, y) = A(x, x) + A(y, y) - 2A(x, y), \tag{10}$$

where A is some sequence attribute. Some natural attributes are shown in table 2. As an illustration, the distance between the two sequences in example 1, now based on the LCS metric (the length of the longest common subsequence), is shown in example 2.

Example 2

sequence 1: **AAA**A**BB**BB
sequence 2: **AAABB**CC

The longest common subsequence of sequences 1 and 2 is AAABB of length 5, so the distance based on the LCS metric is $8 + 7 - 2 \times 5 = 5$.

While being intuitively meaningful and more objective than user-defined cost matrices, these distance criteria usually produce quite different results compared to alignment methods and have not been used often in real applications. Table 2 summarizes differences of some distance metrics used in life sequence analysis.

Censoring. A common problem in life history data are censored observations which in sequence analysis amounts to sequences of uneven length. The assumption of uninformative censoring in EHA is closely related to prediction; the predictions of observable participants are assumed to also be valid for the censored cases. In SA, the problem is how an incomplete observation window for some individuals affects the distance values. The solution is either to simply use shorter sequences or to extend the state space with a new "missing" state. Table 2 summarizes how censoring is handled with different metrics.

Multiple life domains. In the case of one life domain only, the alignment procedure is straightforward and most problems are related to the choice of the distance metric and the definition of the substitution costs. Multiple interdependent life domains complicate analysis in that not only does the state space grow rapidly, but the meaningfulness of the substitution costs also becomes more of an issue. For non-alignment metrics no methods for multi-domain sequences have been proposed to date. For alignment methods at least two approaches have been suggested.

In the *extended alphabet* approach, the letter corresponding to a particular state is replaced by a combination of letters (e.g. being simultaneously in states A, C, G, and J is denoted by ACGJ).[13,17,29] This can extend the state space rapidly. A conceptual problem is that the same cost matrix is applied to all states although it is not straightforward what a substitution of one state with another means in this approach. Gauthier et al.[30] define instead a separate cost matrix for each life domain $c = 1, \ldots, C$ and take an *average* of the costs at each position. If $s_c(x_i, y_j)$ is the cost for aligning x_i with y_j for the life-domain c, the average substitution (or indel) cost is calculated as

$$s(x_i, y_j) = \frac{\sum_{c=1}^{C} s_c(x_i, y_j)}{C}. \tag{11}$$

Typology of sequences. Once the distance matrix has been obtained with some of the alternative metrics in table 2, the goal is to find a typology of sequences by means of clustering methods. In life course studies, the differences between sequences should somehow be related to the timing of events, lengths of episodes determined by onset events, and the complete lack of some events or episodes.

Several alternatives are again available. In life sequence applications, Ward's agglomerative algorithm[31] is most commonly used because it tends to produce more equal-sized clusters than other

clustering algorithms, and this has been preferable for interpretation purposes. At each step, the algorithm combines the two clusters that minimize the within-cluster variability. No unique typology may exist if several pairs of sequences have the same distance value (i.e. there are ties) because a random start of the clustering algorithm can lead to different clustering results.

To determine the optimal number of clusters, generalizations of the usual goodness-of-fit statistics, coefficient of determination R^2 and F-test for non-Euclidian metrics have been used .[32] The sums of squares

$$SS = \frac{1}{n}\sum_{x=1}^{n}\sum_{y=x+1}^{n} d(x,y) \tag{12}$$

are now based on the chosen dissimilarity criterion $d(x, y)$ between sequences x and y. The pseudo R^2 and pseudo F-test, although defined as usual as the ratio of the between and within sum of squares, and that multiplied with the ratio of the degrees of freedom, can now have a different interpretation than in the Euclidean metric.

3 Application to life history calendar data

3.1 The JYLS Study

We illustrate the differences of the prospective and retrospective approaches with the Jyväskylä Longitudinal Study of Personality and Social Development (JYLS), ongoing in Finland. The participants, born in 1959, have been followed from age 8 to 50.[33] In 1968, twelve randomly selected second-grade classes in Jyväskylä, Central Finland, were chosen for the study. All of the pupils participated, so the initial attrition was zero. The original sample consisted of 173 girls and 196 boys. During the follow-up, no systematic attrition has been found.[34,35]

A life history calendar was used to retrospectively collect information about partnership status, children, studies, and work, as well as other important life events. The occurrence, timing, and duration of the transitions were recorded annually from age 15 to age 42 (in 2001[36]) and from age 42 to age 50 (in 2009) during interviews in which 275 participants gave reports based on memory and visual aids provided by the LHC-sheet. The information collected with the LHC was complemented using other sources of information, such as life situation questionnaires and interviews at ages 27, 36, and 42.

Figure 1: A section of the first life history calendar of the JYLS study.

Year													
Marriage/cohab. Age	15	16	17	18	19	20	21	22	23	24	25	...	42
Partner(s)													
Children	15	16	17	18	19	20	21	22	23	24	25	...	42
First child													
Second child													
⋮													
Other parenthood													
Education	15	16	17	18	19	20	21	22	23	24	25	...	42
Type of education													
Work	15	16	17	18	19	20	21	22	23	24	25	...	42
Fulltime work													
⋮													

Table 2: Comparison of non-probabilistic sequence analysis metrics.

	Alignment			Either[a]	Non-alignment		
Method	**Optimal matching**	**Hamming**	**Dynamic Hamming**	**Longest common subsequence**	**Longest common prefix/postfix**	**Number of common subsequences**	**Number of matching subsequences**
Operations/ attributes	Substitution, indels	Substitution	Time-varying substitution	Indels/subsequence	Substring	Subsequences	
Cost definition	substitutions: user-defined, transition probabilities, prototypes; higher indel costs favour substitutions, lower indels (in OM)			Constant	Not relevant		
Computing	Dynamic programming	Sum of substitutions		Dynamic programming	Direct comparison	Dynamic programming	
Principle of similarity	Most common states				Exact prefix/postfix	Same order of states (ignoring repetition)	Same order of states (counting repetition)
Sequences of uneven length	Insert/delete elements	Add missing states		No action			
Multidimensional sequences	Possible				Not yet possible		

[a]The longest common subsequence metric can be seen as either an alignment metric (using only indels) or a non-alignment metric (with the length of the longest common subsequence as the attribute).

We compared event history methods and sequence analysis in a setting where the dynamics of three inter-dependent life domains – partnership formation, parenthood, and employment – are studied in parallel. In EHA, we specified a multistate model for the event-specific transitions and in SA we specified domain-specific cost matrices. As a more substantive question, we studied the relationship between different life paths and excess depressive symptoms in middle age. These were assessed at age 42 using a shortened version of General Behavior Inventory (GBI).[37,38]

3.2 A multistate model

We note first that all events (partnership formation, child births, and career events) can be repeated several times in an individual's life course, making some simplification necessary. We limited the state space to the first transitions in each domain. In particular, we defined "employment" as the year when the person definitively had entered working life. The timings of initial partnership (either marriage or cohabitation) and parenthood are usually easily defined, but the onset of steady employment requires some thought. We defined it as the year which was followed by two subsequent years of employment. Studying and working in the same year was coded either in accordance with the subject's individual situation. The hazards for these events are shown in figure 2, while a histogram of GBI depression scores at age 42 is presented in figure 3.

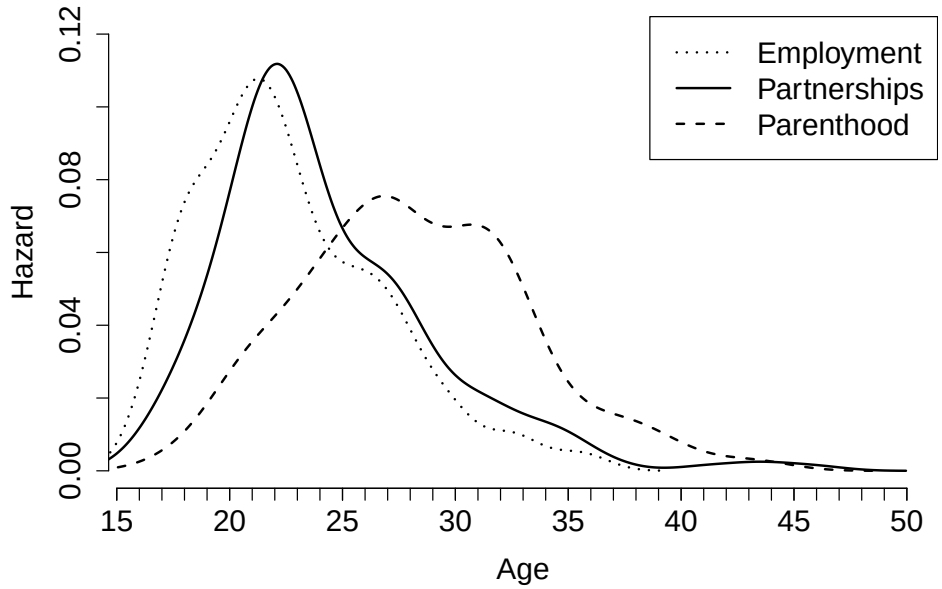

Figure 2: JYLS data: smoothed hazards of initial partnership, parenthood, and employment by age.

We were interested in how the timing of initial partnership and steady employment affect the joint prediction of remaining childless and having excess depressive symptoms at age 42. Excess depressive symptoms was defined as a higher than median GBI score value ($GBI_{med} = 1.44$).

For the sake of simplicity, we excluded cases who had become a parent before the prediction time (age 20) and also one case who had incomplete information on the transitions. This led to a sample size of 260 cases. Figure 4 shows the possible transitions between the states.

The events of interest were denoted by W=entering working life, P=forming an initial partnership, and C=becoming a parent, and their occurrence times by T_W, T_P, and T_C, respectively. The time interval of the LHC recordings was one year and we denote this interval by t, where $t = 20, \ldots\ldots, 42$. Because of the coarse data, it was possible for two or all three events of interest to occur within the same year. Since in that case we do not know the order of events, we simply multiply the discrete hazards in that year in the prediction formulae.

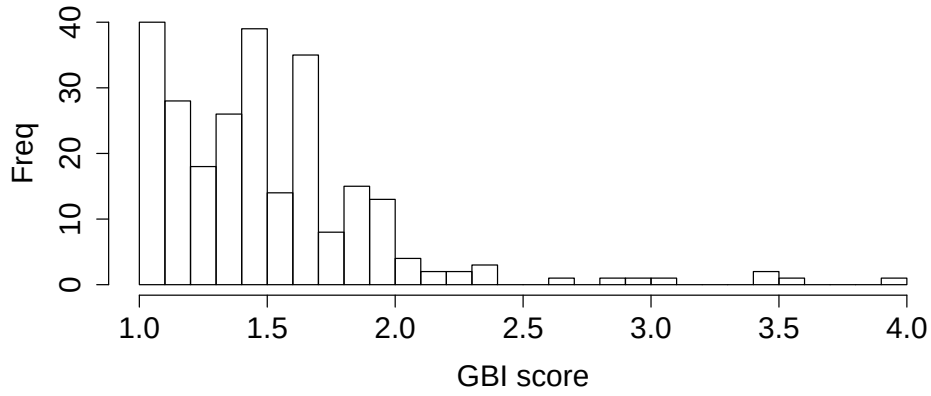

Figure 3: JYLS data: histogram of GBI depression scores at age 42.

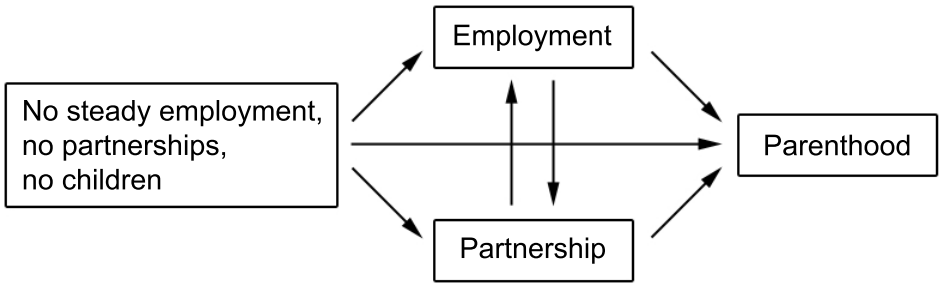

Figure 4: JYLS data: the multistate event history model.

Event-specific hazards. The discrete hazard of entering working life (W) at age t when neither an initial partnership (P) nor parenthood (C) has yet occurred, can be written in the general form as

$$p_W(t) \quad = \quad P(T_W = t | T_W \geq t, T_P \geq t, T_C \geq t). \tag{13}$$

Since any of the events P, W, or C can occur first, a similar hazard model can be defined for initial partnership and parenthood. If both P and W have already occurred at times $w \lesseqgtr v < t$, the conditional hazard of having a first child at age t is then

$$p_{C|WP}(t|v,w) \quad = \quad P(T_C = t | T_W = v, T_P = w, T_C \geq t). \tag{14}$$

Other conditional hazards are defined in an obvious way.

We used piecewise constant logistic hazard models where

$$p_x(t) = (1 + exp(-\boldsymbol{\beta}'\mathbf{Z}_t))^{-1} \tag{15}$$

is the discrete hazard of event x. The effect of the preceding events was modelled with time-dependent covariates which were simple indicators because the sample size did not allow for more complicated modelling. For example, in the hazard $p_{P|W}(t|v)$, the covariate $Z_t(W) \equiv 1$, $t \geq v$, when W occurred at v, whereas in $p_P(t)$ the covariate $Z_t(W)$ was not defined. Although possible, no other covariates were used in the models. Men and women were both included in the final model because no apparent differences in the effects of timing of partnership and work on the response event were found in separate analyses.

Prediction probabilities. In the multistate model the possible paths of not having children within the prediction interval depend on the occurrence times of initial partnership and steady employment.

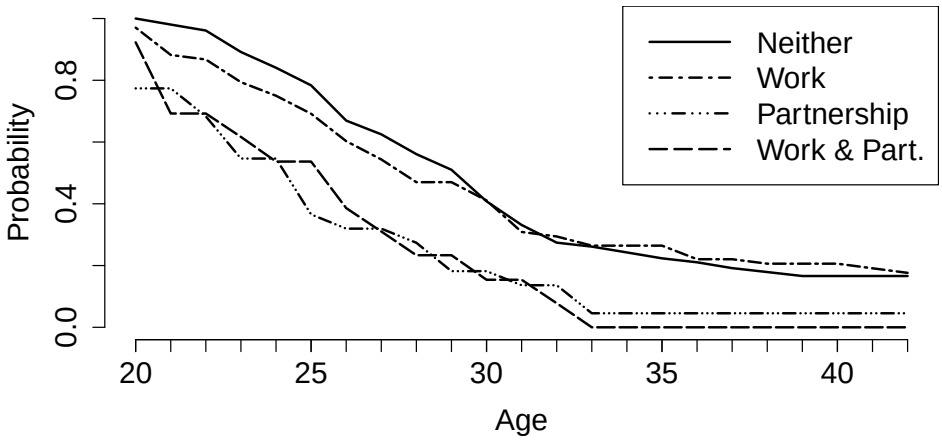

Figure 5: Survival probabilities of not having children for individuals who have or have not entered working life or initial partnership by the prediction time, age 20. "Neither" corresponds to no initial partnership nor employment by age 20.

The most complicated situation is when nothing has yet happened by the prediction time t. In this case, we must account for all possible timings of partnership and employment. We then have the prediction

$$
\begin{aligned}
&P(T_C > u | T_W > t, T_P > t, T_C > t) = \\
&\quad \prod_{s=t+1}^{u} (1 - p_W(s) - p_P(s) - p_C(s)) \\
&\quad + \sum_{s=t+1}^{u} \prod_{r=t+1}^{s-1} (1 - p_W(r) - p_P(r) - p_C(r)) p_W(s) \\
&\quad \times P(T_C > u | T_W = s, T_P > s, T_C > s) \\
&\quad + \sum_{s=t+1}^{u} \prod_{r=t+1}^{s-1} (1 - p_W(r) - p_P(r) - p_C(r)) p_P(s) \\
&\quad \times P(T_C > u | T_W > s, T_P = s, T_C > s) \\
&\quad - \sum_{s=t+1}^{u} \prod_{r=t+1}^{s-1} (1 - p_W(r) - p_P(r) - p_C(r)) p_W(s) p_P(s) \\
&\quad \times P(T_C > u | T_W = s, T_P = s, T_C > s). \qquad (16)
\end{aligned}
$$

The last sum accounts for the paths in which P and W occur within the same year and their order is unknown.

The other paths are special cases of (16). In particular, when initial partnership (P) and entering working life (W) have occurred by the prediction time t, the prediction is simply, for $0 < v \lessgtr w < t < u$,

$$
P(T_C > u | T_W = v, T_P = w, T_C \geq t) = \prod_{s=t+1}^{u} (1 - p_{C|WP}(s|v, w)). \qquad (17)
$$

The prediction probability is a function of the prediction time t and the prediction interval $I = (t, u]$ and its realizations depend on the history H. By letting one of them be variable and fixing the values of the other two, we can obtain different views of the life course dynamics. In figure 5, we obtain the usual survival probability $S(u)$ of not having children by age u when fixing the prediction

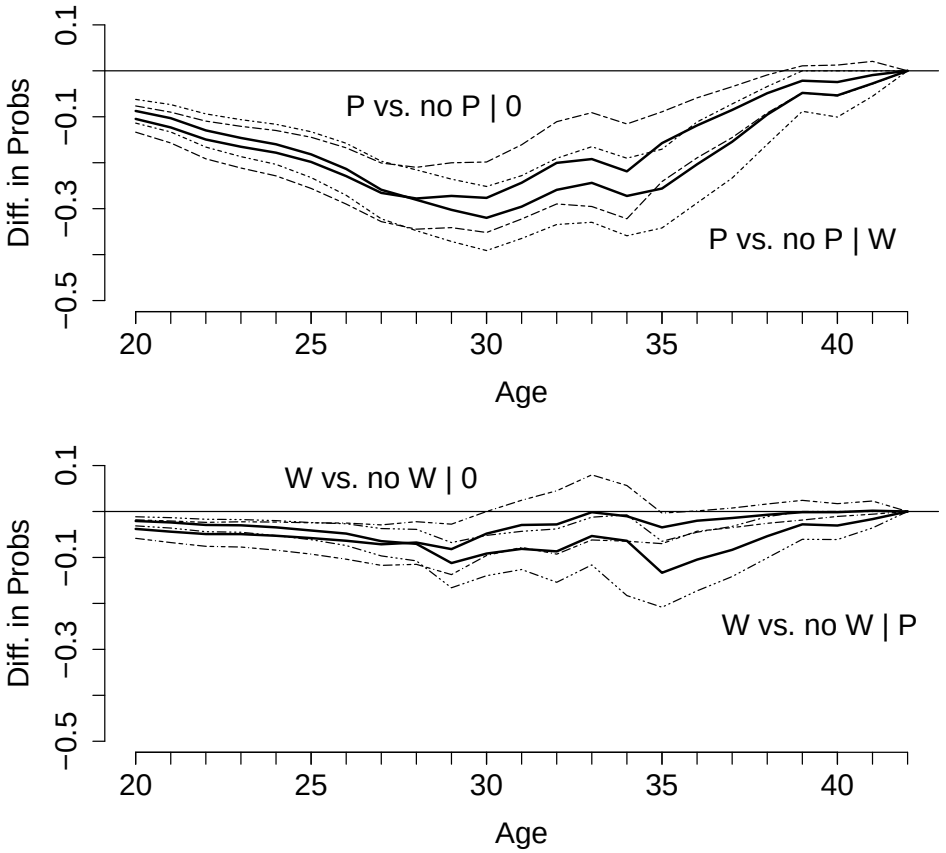

Figure 6: Innovation gains in predicting no children by age $u = 42$ from observing employment (W) and initial partnership (P) at the prediction time $t = 20, ..., 42$, given that nothing/the other one has occurred previously (0 = nothing has yet happened). The confidence intervals are based on 5000 bootstrap samples of the data.

time at $t = 20$ and history at H_{20} and letting the prediction interval vary with ages $u = 21, ..., 42$. We notice that half of those who had formed initial partnership already by age 20, had children by age 25, whereas the effect of employment by age 20 had a much smaller effect on early parenthood compared to those who had neither formed initial partnership nor entered working life at that age.

Factual and counterfactual predictions. Instead of fixing the prediction time t we now identify it with the variable occurrence time $t = 20, ..., 42$ of either initial partnership or employment. When comparing these predictions at age $u = 42$, we obtain a visual representation of the effect of timing of initial partnership P and employment W on the prediction of no parenthood by age 42. In figure 6, we consider the difference of the two predictions:

$$P(T_C > u | T_P = t, T_W > t, T_C > t) - P(T_C > u | T_P > t, T_W > t, T_C > t). \tag{18}$$

This is the *innovation gain* from observing initial partnership at age $t = 20, ..., 42$ related to the prediction of not having children by age 42, given no steady employment by age t. If a person actually forms an initial partnership at age t, the first probability is a *factual* prediction of not having children by age u, given the history, and the second probability is a *counterfactual* prediction of the same event.

At all ages, both initial partnership and employment decreased the probability of remaining childless compared to the situation where neither has occurred yet. The 95% confidence limits show, however, that the timing of steady employment had a significant effect only if it occurred before age 30 if no partnership had been formed yet. Initial partnership around ages 28 to 31 decreased the

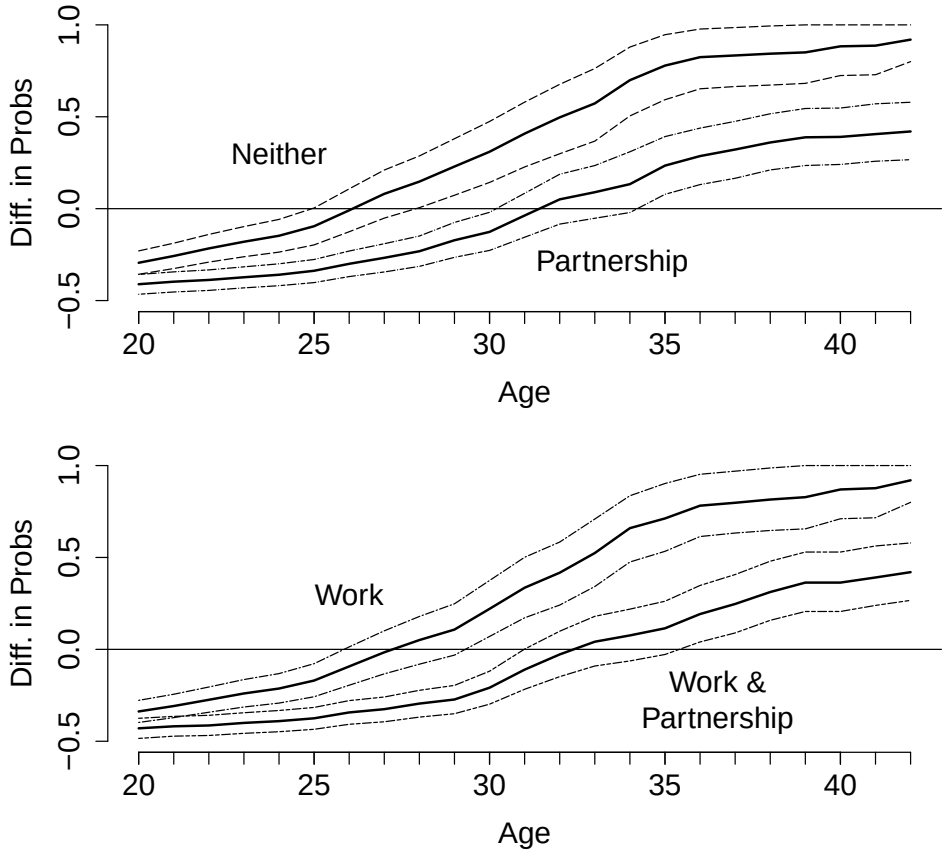

Figure 7: Difference in the joint prediction probabilities of excess depressive symptoms and having children versus not having children by age 42, given that employment or/and initial partnership have occurred at the prediction time $t = 20, ..., 42$. "Neither" corresponds to no partnership or employment (yet) at the time of prediction. The confidence intervals are based on 5000 bootstrap samples of the data.

prediction of remaining childless the most, but had a significant effect at any age. It should be noted that, while controlling for the history effect, the size of the innovation gain from observing initial partnership depends on the length of the remaining prediction interval.

Joint prediction probability. Finally, to evaluate the relationship between possible histories of family formation and employment with depressive symptoms (D) in middle age, we compared the joint prediction of parenthood/no parenthood and excess depressive symptoms at age 42, given the history of partnership and employment. For the case of having children, we then have

$$P(T_C \leq 42, D_{42} > d^* | H_t) = P(D_{42} > d^* | T_C \leq 42, H_t)(1 - P(T_C > 42 | H_t)) \tag{19}$$

with obvious changes for the case of no children.

The first probability on the right is evaluated only at age $u = 42$, so it only affects the last terms at time $u = 42$ in the prediction formulae. It is the cross-sectional logistic probability for a higher than median GBI score d^* at age $u = 42$, depending on family formation and employment

$$p_D(42) = logit(P(D_{42} > d^* | \mathbf{Z}_{42})) = \alpha + \beta_1 Z_{42}(W) + \beta_2 Z_{42}(P) + \beta_3 Z_{42}(C) \tag{20}$$

where $Z_{42}(C) = 1$ for the case when $T_C \leq 42$ and $Z_{42}(C) = 0$ for the case when $T_C > 42$. Since all these covariates were indicators, the occurrence times did not make a difference.

Figure 7 shows the differences in the joint prediction probabilities of having children versus not having children by age 42 and excess depressive symptoms at that age, given initial partnership or employment at the time of prediction. This analysis provides "limiting" ages for increasingly higher prediction of excess depressive symptoms in middle age and remaining childless, compared to having children. We find that if initial partnership is formed later than at age 31, the difference of these joint probabilities becomes positive and increasing. For steady employment but no initial partnership, this age limit is about 27 years. For those who have no initial partnership nor steady employment at the prediction time, this limit is reached already at age 26. By age 34, the prediction of excess depressive symptoms and no children is already about 80% higher than the prediction of excess predictive symptoms and children.

This analysis shows that, having estimated the event-specific hazards, we can evaluate joint predictions of events related to both dynamic and non-dynamic parts of a multistate model. Including explanatory covariates in the hazard models (which we did not do), would allow to compare predictions of hypothetical individuals with different histories and characteristics.

3.3 Multidimensional sequence analysis

In sequence analysis, instead of transitions we studied the distribution of individuals in the states year by year. This difference corresponds to annually evaluating the prevalence of the states instead of incidence. We define the state space of partnership, parenthood, and career histories from age 15 to 50 as shown in table 3.

Table 3: Life domains and respective states for three-domain sequence analysis.

Life domain	States
Partnership	single, in partnership, divorced/separated/widowed
Parenthood	no children, has children (biological, adopted, foster)
Career	studying, working, other (unemployed, out of labour force)

Unlike in the EHA example, we did not restrict the analysis to the first events but used all events for the three life domains. This state space results in 18 possible state combinations for each year. The transition matrix was sparse, but in non-probabilistic sequence analysis and with domain-specific costs this is not a serious issue.

In our data, sequence lengths vary because of the two data collection phases and small differences in ages: 215 participants have sequences of length 36, 14 participants of length 35, and 46 participants of length 28.

Dissimilarity criteria. We compared six dissimilarity metrics suitable for multidimensional sequence analysis. They were based on different definitions of the substitution costs. In optimal matching (OM) and Hamming distance, we used either user-defined or data-driven substitution costs. In dynamic Hamming distance, they were based on estimated transition probabilities taking into account the neighbouring states of the previous and the following year.[21] The LCS criterion (the length of the longest common subsequence) corresponds to OM with the specific choice of substitution cost 2 and indel cost 1.

Since Hamming distance does not allow indels, censored positions were replaced by a "missing" state. The effect of different costs for missing states was investigated by defining costs 0, 0.5, or 1 times the largest substitution cost. With larger costs, the sequences with missing states tend to form their own uninformative cluster. Thus, using no cost at all resulted in the best results.

Table 4: Partnership, parenthood and career-related substitution costs based on theory (user-defined) or transition probabilities.

	User-defined				Transition probabilities			
	→ S	→ P	→ D	→ *	→ S	→ P	→ D	→ *
Single (S) →	0	2	3	0	0	1.89	2.00	0
Partnership (P) →	2	0	1	0	1.89	0	1.80	0
Divorced/sep. (D) →	3	1	0	0	2.00	1.80	0	0
Missing (*) →	0	0	0	0	0	0	0	0

	User-defined			Transition probabilities		
	→ N	→ C	→ *	→ N	→ C	→ *
No children (N) →	0	3	0	0	1.94	0
Has children (C) →	3	0	0	1.94	0	0
Missing (*) →	0	0	0	0	0	0

	User-defined				Transition probabilities			
	→ S	→ W	→ O	→ *	→ S	→ W	→ O	→ *
Studying (S) →	0	3	1.5	0	0	1.77	1.87	0
Working (W) →	3	0	1.5	0	1.77	0	1.67	0
Other (O) →	1.5	1.5	0	0	1.87	1.67	0	0
Missing (*) →	0	0	0	0	0	0	0	0

Combined substitution costs. The substitution cost matrix was defined separately for each three domain and then averaged. The domain-specific costs for OM and Hamming distance are shown in table 4. For dynamic Hamming, time-specific costs resulted in 36 distinct substitution cost matrices (not shown here). It should be noted that the absolute numbers in user-defined substitution costs have no meaning since the information is only relative. In our application, the states "single" and "divorced" were the most distant because forming a partnership was regarded as one step in the developmental process to adulthood. In another study, these could be interpreted as similar, as both indicate a state of "living without a partner". Compared to the transition-based costs, this is the main difference in the partnership domain. For career domain, transitions from states "studying" to "other" (or vice versa) were the least common (highest cost in the cost matrix based on transition probabilities), but in the user-defined matrix the corresponding cost was set relatively low, due to the versatile nature of the state "other". The indel costs were set to half of the largest substitution cost, making them equally costly. For averaging, the costs in each matrix were scaled to have the same range in order to give equal weight to each life domain.

Typology of sequences. Ward's agglomerative clustering was used to find a typology of life sequences, applying the six dissimilarity criteria for solutions starting from 2 to 15 clusters. Based on dendrograms, the goodness-of-fit statistics, and interpretability of the clusters, an eight-cluster solution was chosen.

The goodness-of-fit statistics in table 5 for the chosen eight cluster solutions suggested that clustering based on the Hamming distance with theory-based substitution costs fits the data best. It covered around 45% of sequence variation ($F = 31.56$) and resulted in interpretable clusters where all three life domains were well represented. In comparison, the second best criterion, dynamic

Table 5: Goodness-of-fit statistics for eight cluster solutions obtained with six distance measures based on transition probabilities or user-defined costs.

Dissimilarity measure	Pseudo R^2	Pseudo F
Hamming distance (user-defined)	0.453	31.56
Dynamic Hamming distance (trans. prob.)	0.433	29.18
Optimal matching (user-defined)	0.406	26.09
Hamming distance (trans. prob)	0.395	24.93
Optimal matching (trans. prob)	0.369	22.33
Length of longest common subsequence	0.358	21.23

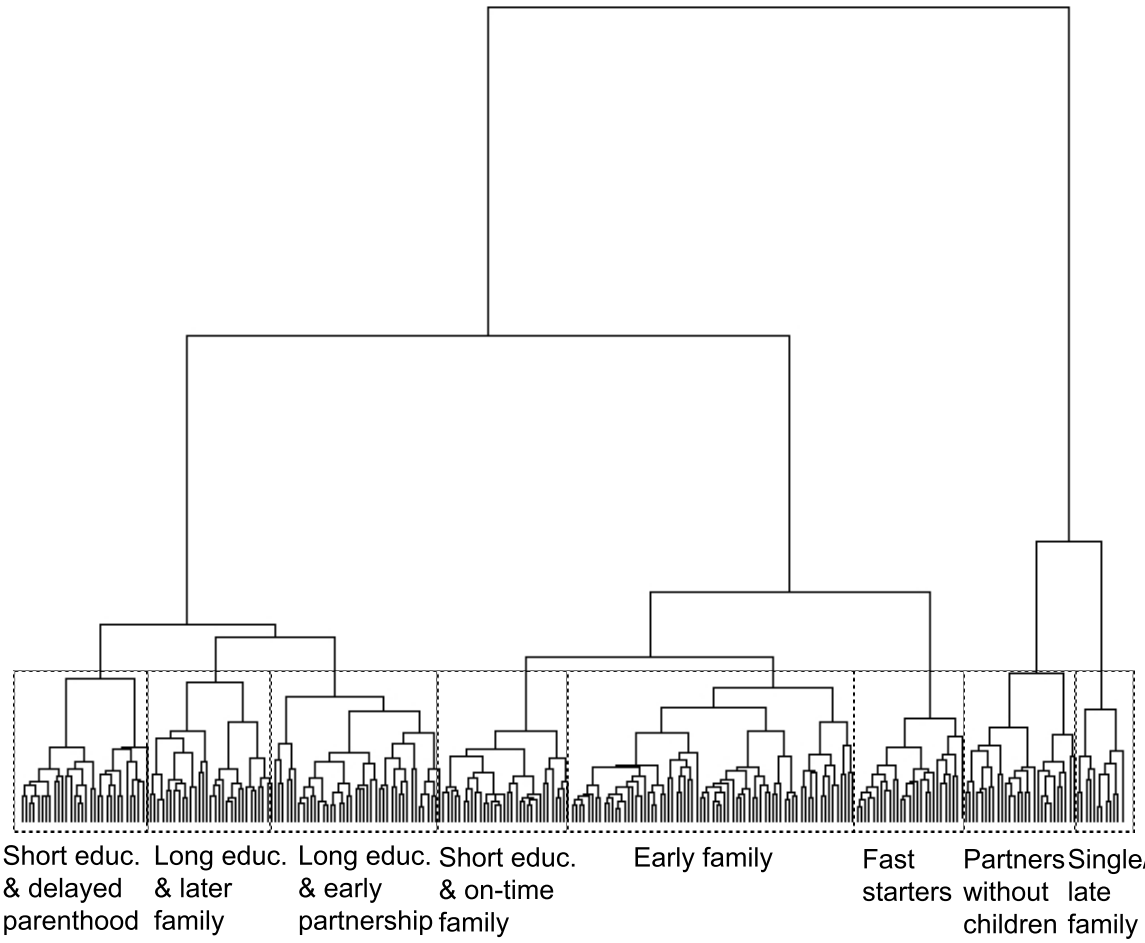

Figure 8: The dendrogram of the clustering based on Hamming distance with user-defined substitution costs.

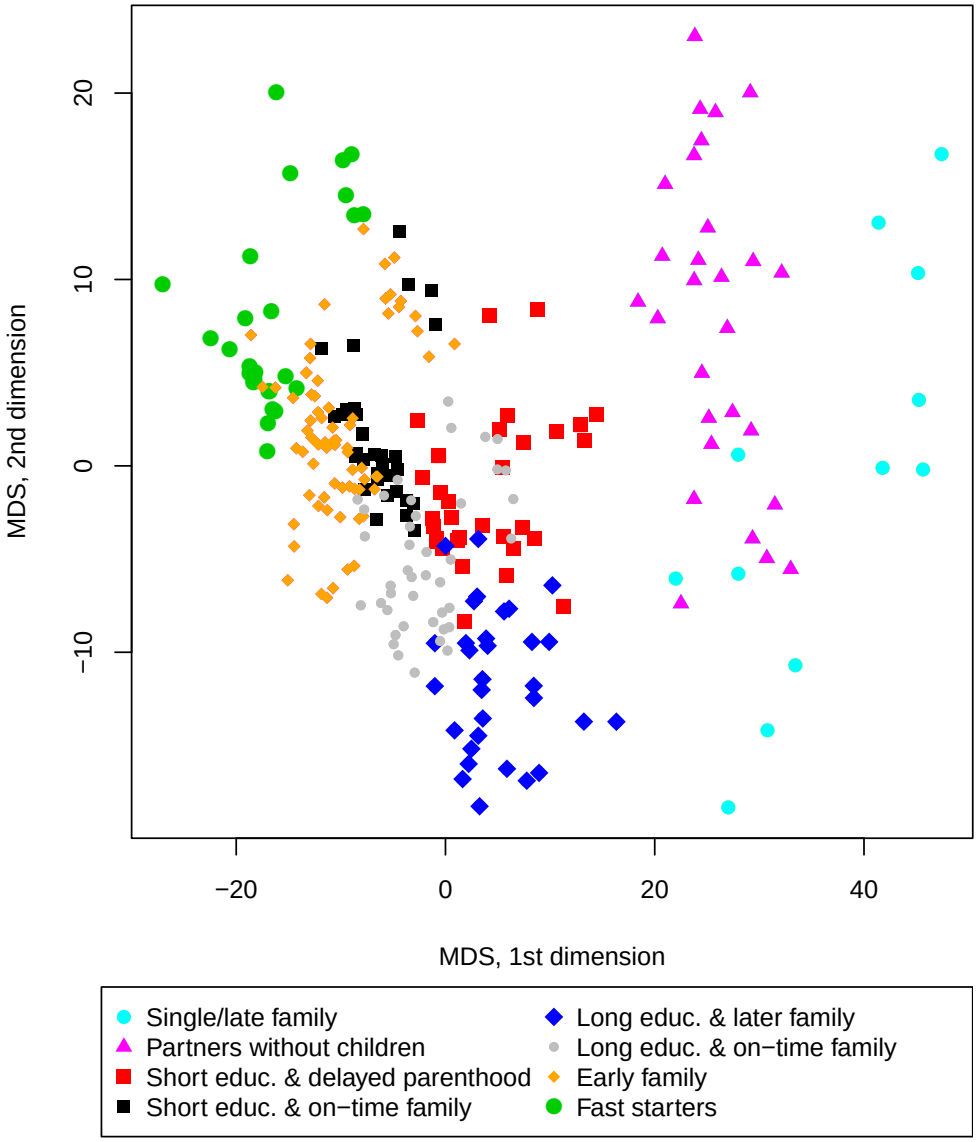

Figure 9: Scatter plot of the cluster-specific MDS scores based on the first two dimensions of multidimensional scaling. Dissimilarities were computed by using theory-based Hamming distance for the three-dimensional sequences.

Hamming, resulted in clusters where the family-related life domains dominated and the career domain was hardly represented. The dendrogram based on the Hamming distance in figure 8 supported the eight cluster solution.

Time-preserving Hamming distance instead of OM seems more reasonable for sequences of uneven length. In OM, using indels in our data would mean aligning, for example, a state at age 15 with a state at age 23 in another sequence. Within the same metric, user-defined substitution costs gave better results than costs based on transition probabilities in this three-domain setting. However, preliminary studies with only one life domain suggested the opposite so no general guidelines can be given.

We present the results with the Hamming distance, at the same time illustrating different ways of investigating the clustering results by sequence plots, by comparing sequence variation in clusters, by reducing dimensionality in the multivariate categorical analysis with multidimensional scaling and finally, by using the cluster indicators in the regression of depressive symptom scores on cluster membership.

Table 6: Logistic regression of depression score on cluster membership.

	β	s.e.	p	OR
Short education & delayed parenthood	−0.21	0.37	0.58	0.81
Short education & on-time family	−0.51	0.37	0.16	0.60
Long education & late family	−0.59	0.39	0.14	0.56
Partners without children	−0.24	0.40	0.55	0.79
Early family	0.18	0.25	0.46	1.20
Single/late family	1.61	0.77	0.04	5.00
Long education & early partnership	−0.05	0.33	0.87	0.95
Fast starters	0.47	0.40	0.24	1.60

Multidimensional scaling (MDS). MDS provides a concise visual representation of cluster results, first by showing how well the clusters actually separate but also by providing a visual aid when the original sequences are ordered according to MDS scores. The first few scaling dimensions capture the most prominent variation in the sequences. The rotation of the solution is arbitrary, but principal component axes can be used for achieving a meaningful rotation. The resulting dimensions often sort the sequences according to an attribute, such as the timing of some transition.

In figure 9, the sequences were plotted as points on the plane spanned by the first two MDS dimensions with cluster identification. The eigenvalues of the MDS solutions with different dimensions supported two MDS dimensions. Correlation between the original Hamming distances and the distances computed from two-dimensional MDS scores was 0.93. The timing of initial partnership and parenthood seemed to separate the clusters best (1st principal component dimension); length of education follows (2nd dimension). Clusters of individuals with no children were clearly separated from the others, which were more or less connected but not completely overlapping.

Sequence plots. *Index plots* show the individual life courses, merely re-organizing the original data according to the similarity defined by clustering (figure 10). Ordering according to some MDS dimension assists in interpretation. *State distribution plots* show the prevalence of states at each time point. We combined the different life domains to give an overview of the dynamics of the state distribution (figure 11).

Sequence variability. Shannon's entropy[39,40] is often used as a measure of disorder of a system. In life sequence analysis, entropy is used to characterize variation in the states within one sequence, or more interestingly, within and between clusters of sequences. When entropy is 0, all cases (of a cluster) are in the same state. When entropy is 1, there are equally large amount of cases in each state. Important transition times are easily seen as peaks in the cluster-specific plots (figure 12).

Regression analysis. External explanatory variables can be taken into account either in the clustering phase (covariance analysis instead of ANOVA), or as independent variables in a multinomial analysis of cluster membership indicators. We used the membership indicators as explanatory factors in a logistic regression predicting higher than median depression scores as in EHA. The "single/late family" cluster was the only one that shows statistically significant differences (with higher odds of having excess depressive symptoms). This result supports the finding of Salmela-Aro et al.[41] that postponing or lack of some stages in the transitory process to adulthood anticipate lower life satisfaction in adulthood. Note that although we used individual-specific cluster membership indicators in the regression models, the cluster characteristics may not be representative to all members of the cluster. Clustering was based on the matrix of pairwise distances, not on the individual sequences any more. It was therefore expected that only the most different clusters (here singles) would have

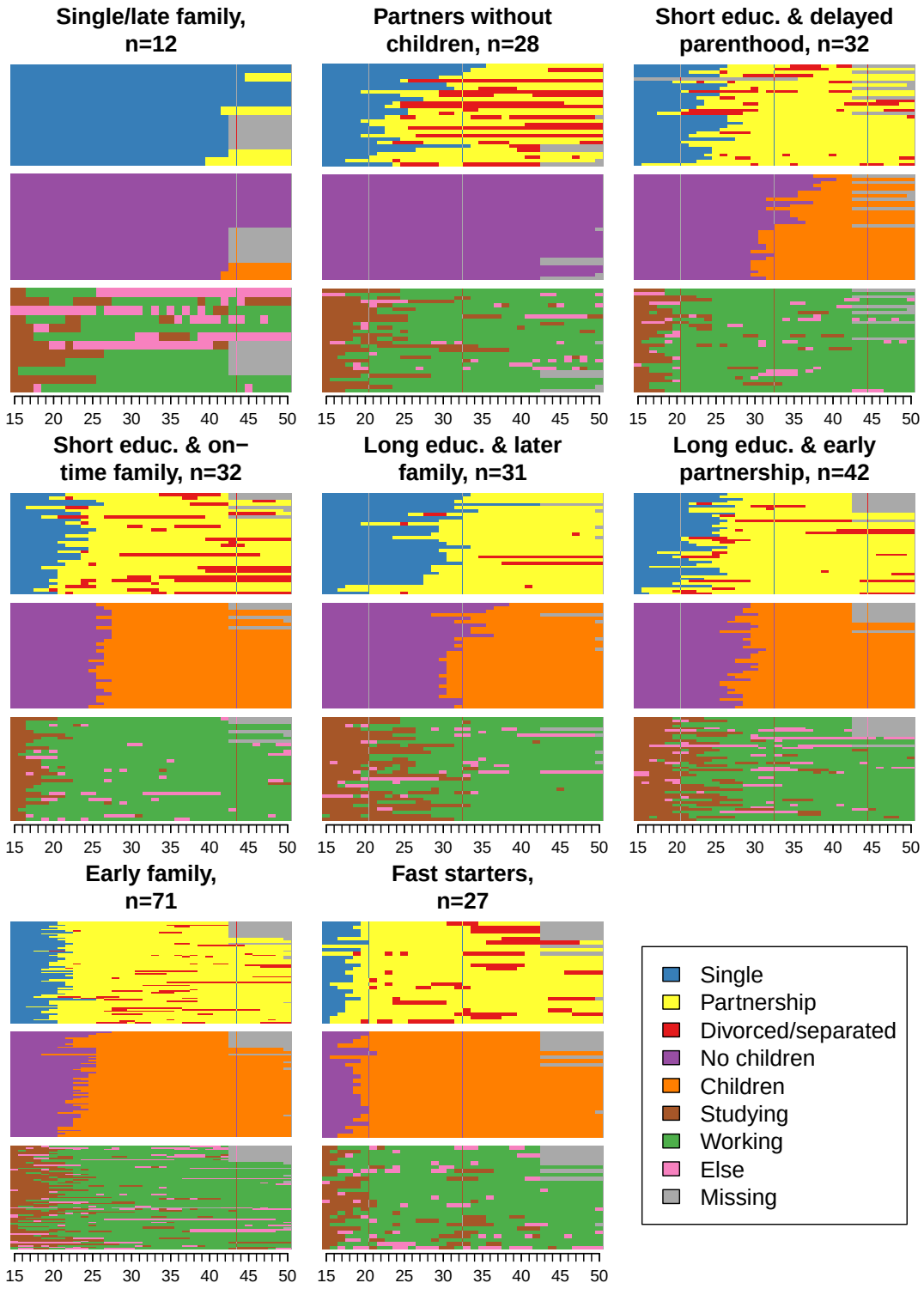

Figure 10: Index plots of partnership (top), parenthood (middle), and career (bottom) in the eight clusters based on Hamming distance. The sequences are ordered according to the first dimension of multidimensional scaling that represents the timing of partnership and children.

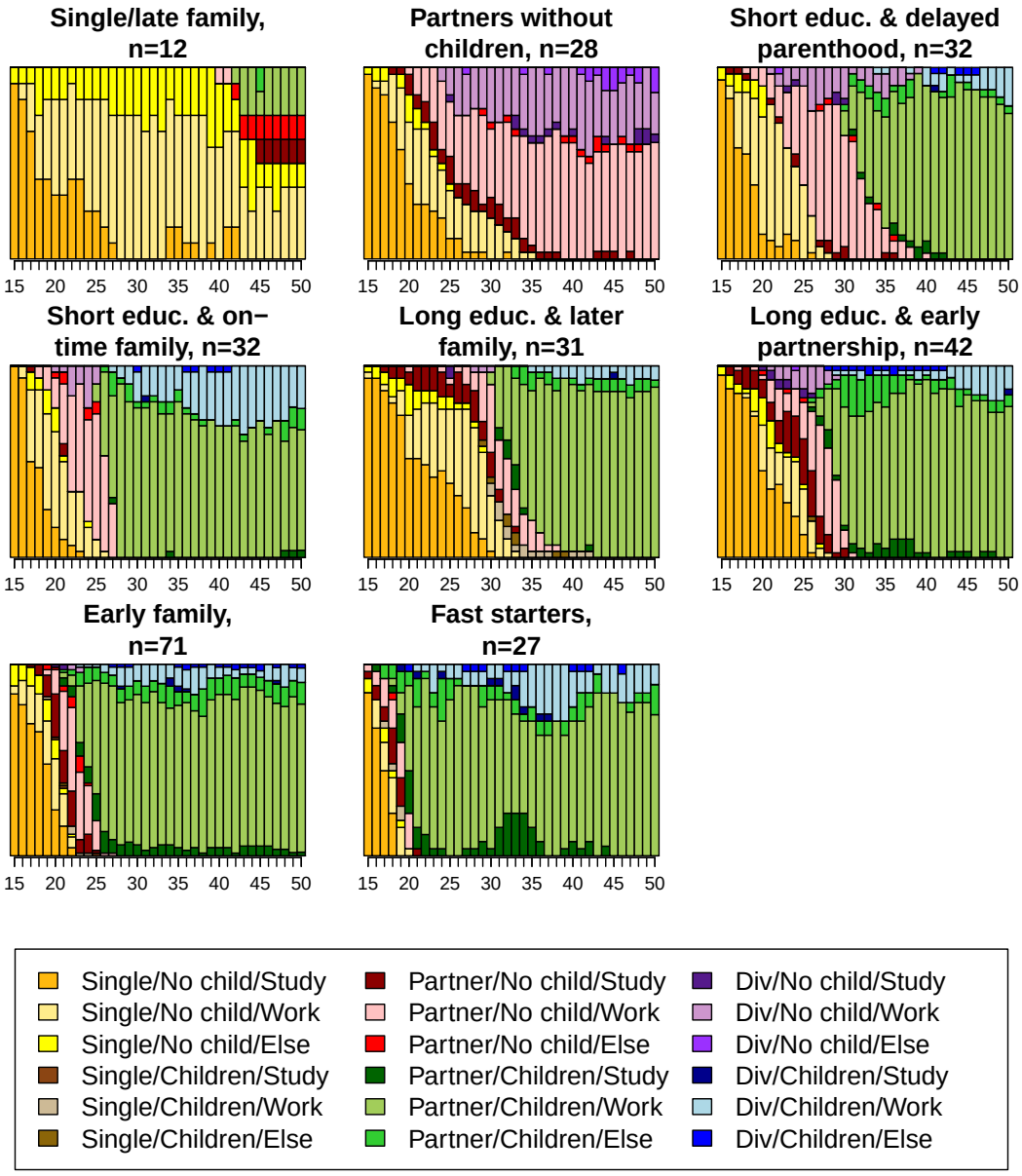

Figure 11: State distribution plots of combined partnership, parenthood, and career states in the eight clusters based on Hamming distance. Note, that "divorced" can mean either a broken marriage or cohabitation. Positions with missing states in any life domain are excluded from the plots.

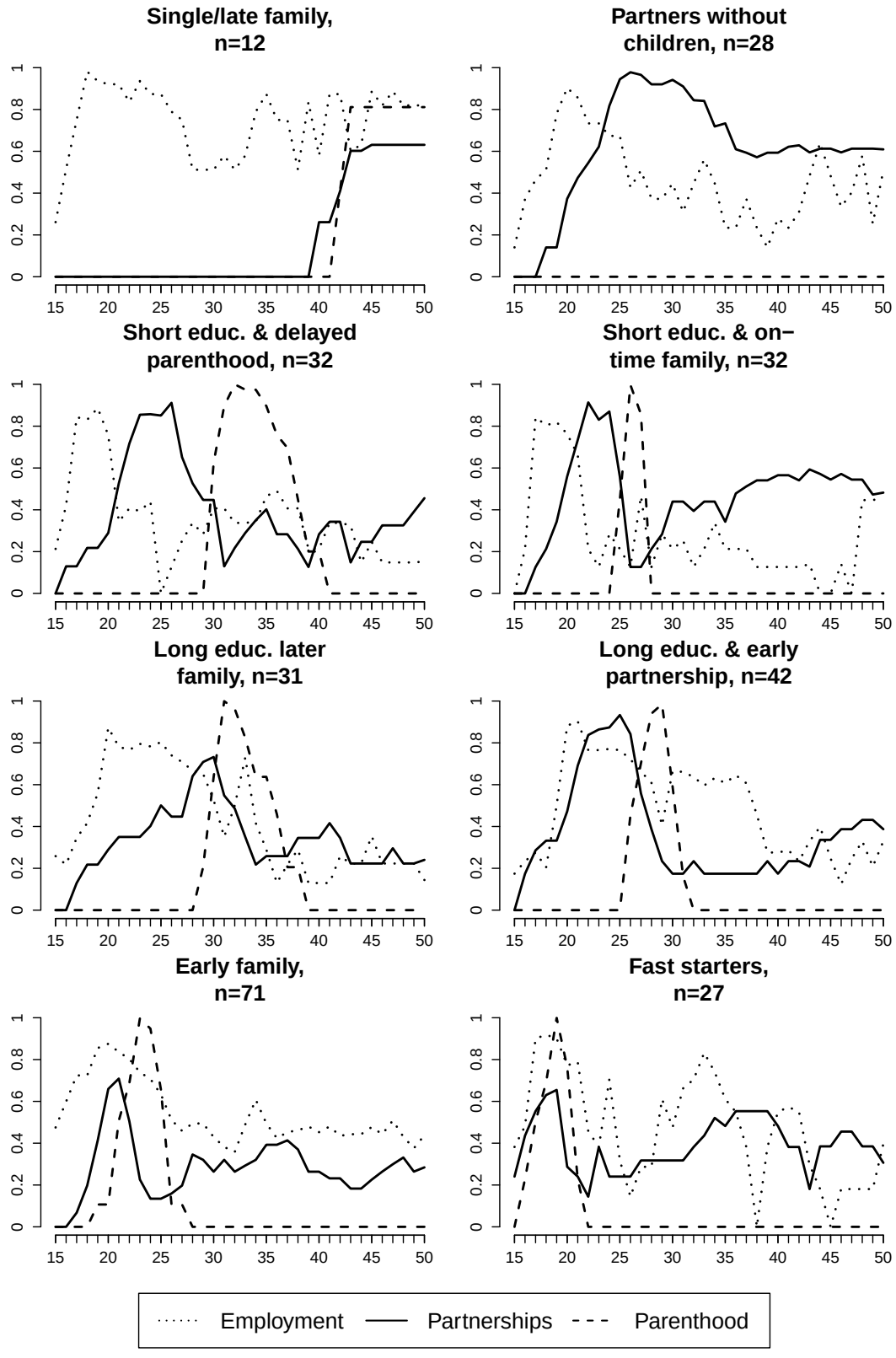

Figure 12: Transversal entropies of partnerships, parenthood, and career sequences in clusters based on Hamming distance.

a significant role in the regression analysis. We conclude that, unlike making individual-level predictions of parenthood given the history of partnership and employment, the aim of sequence analysis was to find subpopulations or clusters of individuals whose life courses were similar in terms of the timing of initial partnership, parenthood and employment.

Computations. Sequence analyses were carried out with the *TraMineR* library in *R*.[42] Logistic hazard models and programs calculating the prediction probabilities and their bootstrap intervals in section 3.2 were implemented with R.

4 Discussion

We compared two approaches of analysing data collected with a life history calendar: the model-based probabilistic method of event history analysis and the model-free data mining method of sequence analysis. Traditionally, EHA models the risk of a transition from one state to another, but here we instead estimated the cumulative prediction probabilities of life events in a multistate model to have a more comparable setup with sequence analysis. Instead of transitions, the analysis was extended to the entire observed trajectory, which was the unit of analysis in SA as well. In sequence analysis, we compared several dissimilarity metrics and contrasted data-driven and user-defined substitution costs. To illustrate the two methods, we studied young adults' transition to adulthood as a sequence of landmark events in several life domains. These landmark events defined our multistate event-history model and the parallel life domains in multidimensional SA. Finally, we analysed the relationship between life trajectories and excess depressive symptoms at age 42 by first estimating their joint predictions in the multistate model and then by using the individual-specific cluster indices of multidimensional SA in a further explanatory analysis for depressive symptoms.

When the same life course problem was analysed with both methods, we found that the two approaches complement each other. SA is a descriptive tool synthesizing large amount of information to obtain a broad picture of multidimensional data. As other dimension-reducing methods, SA helps developing an intuitive understanding of complex relationships but the resulting clusters should not be given a confirmatory status. Finding descriptions for the clusters mirrors the rather subjective way of naming factors in factor analysis. In our case, sequence analysis could reveal typical and atypical patterns of young adults transition process to adulthood which supported the earlier findings that no normative pathway to adulthood exists any longer. Individuals' enhanced opportunities to make choices in their own lives increases diversity in the life course. These individual choices are affected by various governmental and other external decisions, the results of which are difficult to conceive at the population level. In particular, sequence analysis has offered new means for large scale comparative analysis of life patterns across nations and between age cohorts.

Multistate event history analysis, on the other hand, is a predictive method which requires structured hypotheses and a well-defined system of hazard models. This is opposite to the data mining approach of SA in which no assumptions about the data generating mechanisms are made, or needed, for that matter. We believe, however, that the analysis of increasingly complex life course data, combining perhaps both biological and behavioural data, will require methods that at the initial stage can reveal underlying structures and help generating causal hypotheses for further analysis. Causal inquiries can only be addressed with proper "book keeping" of risk sets for transitions. Thus, correct individual-level conditioning of the history is possible only in EHA. Multiple time scales, inherent in many life course problems, and their separate effects can only be quantified by modelling. Furthermore, time-varying covariates indicating individual status changes or contextual changes in time are only possible in model-based analysis.

Sequence analysis has been criticized for violating the basic principles of prospective analysis because the "past" and the "future" are treated symmetrically in vertical alignment. In this sense,

it is not suitable for any causal analysis. Subjectivity of substitution cost specification and non-uniqueness of clustering results have also raised scepticism about its usefulness. In recent years, several improvements have been suggested to the specification of substitution costs, to handle censoring, and to preserve timing and the order of states in sequence analysis.[43] They all modify the substitution cost matrix in some way because this is the only way of tuning the values of the distance matrix. According to our examinations, also Elzinga's non-alignment methods[26–28] seem promising, but no multidimensional method exists yet. As a data mining method, SA is best suited for large register-based data sets. With small data sets and large state space, all trajectories tend to be unique. If the substitution cost matrix is based on estimated transition probabilities, small data sets run out of observations. This was shown by Helske et al.,[44] who used Hidden Markov models to cluster life sequences probabilistically.

Statistical analysis of life sequences still has many unresolved questions, compared to the well developed theory of event history analysis. Sequence analysis is less conventional, but its use is expected to increase in the future, especially now that there is an easy-to-use software available in R. Event history analysis will certainly remain the main tool for analytical life course studies. We believe that although the prediction probabilities are not a standard tool in EHA, they are valuable for synthesizing information in a multistate model. Although the probabilistic statements and programming require careful specification, the probabilities can be estimated in a straightforward manner from state-specific hazards. Confidence intervals can be calculated, for example, by bootstrapping (as we did) or, in a fully parametric case, analytically (cf. Eerola[8]).

As in epidemiology, prevalence indicates what is typical or atypical at a particular time, whereas incidence is related to change, the underlying concept in all causal inquiry. Life course analysis is obviously dynamic, but the complex pattern of interacting factors also requires "zooming" into details. Therefore, one could summarize the complementary advantages of the methods: while sequence analysis provides detailed information about "how things are", event history analysis answers the "why".

Acknowledgements. We thank The Jyväskylä Longitudinal Study of Personality and Social Development, led by Lea Pulkkinen, for letting us use the data in our study. We thank especially Katja Kokko and Eija Räikkönen for their comments.

References

[1] Belli RF, Stafford FP, Alwin DF. Calendar and time diary: methods in life course research. Sage Publications, Inc; 2008.

[2] Caspi A, Moffitt TE, Thornton A, Freedman D, et al. The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. International Journal of Methods in Psychiatric Research. 1996;6(2):101–114.

[3] Andersen PK, Borgan Ø, Gill RD, Keiding N. Statistical models based on counting processes. Springer Verlag; 1993.

[4] Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. Journal of Epidemiology and Community Health. 2003;57(10):778–783.

[5] Andersen PK, Keiding N. Multi-state models for event history analysis. Statistical Methods in Medical Research. 2002;11(2):91.

[6] Arjas E. Survival Models and Martingale Dynamics. Scandinavian Journal of Statistics. 1989;p. 177–225.

[7] Arjas E, Eerola M. On predictive causality in longitudinal studies. Journal of statistical planning and inference. 1993;34(3):361–386.

[8] Eerola M. Probabilistic causality in longitudinal studies. vol. 92 of Lecture Notes in Statistics. Springer-Verlag; 1994.

[9] Klein JP, Keiding N, Copelan EA. Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients. Statistics in Medicine. 1993;12(24):2315–2332.

[10] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine. 2007;26(11):2389–2430.

[11] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.

[12] Abbott A. Sequence Analysis: New Methods for Old Ideas. Annual Review of Sociology. 1995;21(1):93–113.

[13] Pollock G. Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2007;170(1):167–183.

[14] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10; 1966. p. 707–710.

[15] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology. 1970;48(3):443–453.

[16] Halpin B. Optimal Matching Analysis and Life-Course Data: The Importance of Duration. Sociological Methods & Research. 2010;38(3):365–388.

[17] Stovel K, Savage M, Bearman P. Ascription into achievement: Models of career systems at Lloyds Bank, 1890–1970. The American Journal of Sociology. 1996;102(2):358–399.

[18] Rohwer G, Pötter U. TDA User's Manual; 2004.

[19] Wiggins R, Erzberger C, Hyde M, Higgs P, Blane D. Optimal matching analysis using ideal types to describe the lifecourse: an illustration of how histories of work, partnerships and housing relate to quality of life in early old age. International Journal of Social Research Methodology. 2007;10(4):259–278.

[20] Hollister M. Is Optimal Matching Suboptimal? Sociological Methods & Research. 2009;38(2):235–264.

[21] Lesnard L. Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. Sociological Methods & Research. 2010;38(3):389–419.

[22] Hamming RW. Error detecting and error correcting codes. Bell System Technical Journal. 1950;29(2):147–160.

[23] Wu LL. Some Comments on "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect". Sociological Methods Research. 2000;29(1):41–64.

[24] Stovel K. Local sequential patterns: The structure of lynching in the Deep South, 1882–1930. Social Forces. 2001;79(3):843–880.

[25] Marteau PF. Time warp edit distance with stiffness adjustment for time series matching. IEEE transactions on pattern analysis and machine intelligence. 2008;31(2):306–318.

[26] Elzinga CH; Citeseer. Sequence similarity: a nonaligning technique. Sociological Methods and Research. 2003;32(1):3–29.

[27] Elzinga CH. Sequence analysis: Metric representations of categorical time series. Manuscript. 2006;.

[28] Elzinga CH, Liefbroer AC. De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. European Journal of Population/Revue européenne de Démographie. 2007;23(3):225–250.

[29] Han SK, Moen P. Clocking out: Temporal patterning of retirement. American Journal of Sociology. 1999;105(1):191–236.

[30] Gauthier JA, Widmer ED, Bucher P, Notredame C. Multichannel sequence analysis applied to social science data. Sociological Methodology. 2010;40(1):1–38.

[31] Ward JH Jr. Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 1963;p. 236–244.

[32] Studer M, Ritschard G, Gabadinho A, Müller N. Discrepancy analysis of complex objects using dissimilarities. Advances in Knowledge Discovery and Management. 2010;292(4):3–19.

[33] Pulkkinen L, Lyyra AL, Kokko K. Life success of males on nonoffender, adolescence-limited, persistent, and adult-onset antisocial pathways: follow-up from age 8 to 42. Aggressive Behavior. 2009;35(2):117–135.

[34] Pulkkinen L. The Jyväskylä Longitudinal Study of Personality and Social Development. In: Pulkkinen L, Kaprio J, Rose RJ, editors. Socioemotional development and health from adolescence to adulthood. Cambridge University Press, New York; 2006. p. 29–55.

[35] Pulkkinen L, Kokko K. Tiivistelmä [Summary]. In: Pulkkinen L, Kokko K, editors. Keski-ikä elämänvaiheena [Middle-age as a stage of life]. Jyväskylän yliopisto, Jyväskylä; 2010. p. 5–13.

[36] Kokko K, Pulkkinen L, Mesiäinen P. Timing of parenthood in relation to other life transitions and adult social functioning. International Journal of Behavioral Development. 2009;33(4):356–365.

[37] Depue R. General Behavior Inventory; 1987. Ithaca, NY: Department of Psychology, Cornell University.

[38] Kokko K, Pulkkinen L. Unemployment and psychological distress: Mediator effects. Journal of Adult Development. 1998;5(4):205–217.

[39] Shannon C. A mathematical theory of communication. The Bell System Technical Journal. 1948;27(7):379–423.

[40] Billari FC. The Analysis of Early Life Courses: Complex Descriptions of the Transition to adulthood. Journal of Population Research. 2001;18(2):119–142.

[41] Salmela-Aro K, Kiuru N, Nurmi JE, Eerola M. Mapping pathways to adulthood among Finnish university students: Sequences, patterns, variations in family-and work-related roles. Advances in Life Course Research. 2011;16(1):25–41.

[42] Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software. 2011;40(4):1–37.

[43] Aisenbrey S, Fasang AE. New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. Sociological Methods & Research. 2010;38(3):420–462.

[44] Helske J, Eerola M, Tabus I. Minimum description length based hidden Markov model clustering for life sequence analysis. In: Proceedings of the Third Workshop on Information Theoretic Methods in Science and Engineering, August 16–18, 2010, Tampere, Finland; 2010. .