Predicting the stability of early employment with its timing and childhood social and health-related predictors: a mixture Markov model approach

Satu Helske
Markus Keski-Säntti
Juha Kivelä
Aapo Juutinen
Antti Kääriälä
Mika Gissler
Marko Merikukka
Tea Lallukka

11.3.2021, update 16.3.2023

ISSN 2737-0534

The Inequalities, Interventions, and New Welfare State (INVEST) aims at increasing wellbeing of Finnish society during childhood, youth and early adulthood and preventing psychosocial risks compromising such development through innovative interventions. Based on cutting-edge research on the conditions and mechanisms involved at different periods of development, INVEST will evaluate and develop various universal and targeted interventions to improve the efficiency of the current welfare state institutions at critical points of the early life course. INVEST aims at providing a new model for the welfare states that is more equal, better targeted to problem groups, more anticipatory as well as economically and socially sustainable. INVEST is a Flagship project of the Academy of Finland.

Predicting the stability of early employment with its timing and childhood social and health-related predictors: a mixture Markov model approach

Satu Helske¹, Keski-Säntti Markus², Juha Kivelä², Aapo Juutinen², Antti Kääriälä², Mika Gissler^{2,3}, Marko Merikukka⁴, Tea Lallukka⁵

1 INVEST Research Flagship Center and Department of Social Research, University of Turku, Turku, Finland

2 Finnish Institute for Health and Welfare, Helsinki, Finland

3 Karolinska Institute, Department of Molecular Medicine and Surgery and Region Stockholm, Academic Primary Health Care Centre, Stockholm, Sweden

4 Itla Children's Foundation, Helsinki, Finland

5 Department of Public Health, University of Helsinki, Helsinki, Finland

Note: This manuscript has been published in *Longitudinal and Life Course Studies*. Please cite:

Helske, S., Keski-Säntti, M., Kivelä, J., Juutinen, A., Kääriälä, A., Gissler, M., Merikukka, M. & Lallukka, T. (2023). Predicting the stability of early employment with its timing and childhood social and health-related predictors: a mixture Markov model approach. *Longitudinal and Life Course Studies*, 14(1), 73-104. doi:10.1332/175795921X16609201864155

1. Introduction

Due to increased life expectancy and fallen birth rates in many countries, the proportion of people in the workforce has dramatically declined. Working life expectancy has also been increasing but remains relatively short (Nurminen, 2011). Many studies focus on older employees and risk factors of disability retirement. However, it is equally crucial to gain new information about entry into paid employment and the stability of early working life. For example, it is important to pinpoint risk groups of unstable working life as early as possible to target their determinants.

In young adults' lives, entry into paid employment is an important transition, which can be shaped by contextual factors, such as age at entry.. Accordingly, socioeconomic background and unhealthy lifestyle have been linked to an early entry into paid employment and to a lower occupational class in the first job (Halonen *et al.*, 2019a). Additionally, early entry into paid employment and physical heaviness of work in young adulthood shape the development of unhealthy behaviours and obesity in the long run (Shiri *et al.*, 2021). Hence, the timing of entry could be shaped by both parental and own socioeconomic position and health, further contributing to subsequent health and the stability of work participation later on. However, it remains unclear how the timing of entry into paid employment together with the accumulation of different life-course social and health-related determinants are linked to different labour market transitions such as unemployment during early careers and particularly the stability of early employment.

Labour market statistics show that the proportions of both under 25-year-olds and over 55-year-olds in the workforce have remained low compared to midlife groups (World Health Organization, 2012a, 2012b). Thus, there is a need to extend work careers by focusing on the determinants of work participation patterns among young employees particularly. One of the reasons for late labour market entries is entering tertiary education, which in Finland is often both delayed due to challenges in entering education and also extended due to delays in graduation (Statistics Finland, 2013). In the long run, higher education is, however, predictive of more stable labour market participation. Reasons for other types of unstable early work participation and spells of non-employment could further be attributable to early risk factors that have a long latency. Such factors could originate already from childhood, and be linked to both parental and own socioeconomic position and health. Particularly social disadvantage and early mental disorders of parents and the offspring have been linked to both work disability and long-term unemployment of young adults (Halonen *et al.*, 2019b; Lallukka *et al.*, 2019). As mental disorders increase already among young adults, even before entering paid employment, they need to be considered when focusing on work participation among young adults (Patel *et al.*, 2007; Croft, Blyth and Van Der Windt, 2011; Farioli *et al.*, 2014). Accordingly, Swedish register data shows that among mental disorders, e.g. neurodevelopmental disorders are typically diagnosed young and are strongly associated with subsequent work disability (Virtanen *et al.*, 2020) and unemployment trajectories (Lallukka *et al.*, 2020). A substantial part of early exits from paid employment is attributable to these disorders, thus the associations can be bidirectional with early ill health predicting unstable work participation, and non-employment is ill health later on (Kela, 2010; OECD, 2010; Coggon *et al.*, 2013).

Early detection of modifiable risk factors of unstable early work participation patterns is crucial. However, there is limited knowledge about the factors contributing to the stability of early careers, such as the opportunities of the disabled to enter the labour market and continue working (Achterberg *et al.*, 2009). Moreover, most studies focus on one exit type or the time to exit (e.g. work disability having dichotomous outcomes). Such approaches fail to consider different statuses and transitions between statuses that more comprehensively describe the true development of work participation. Young adults are unlikely a homogeneous group, even when they belong to the same birth cohort, and failure to consider different actual patterns of work participation patterns provide a rather limited picture of such patterns in young adulthood. More evidence is needed on what are the paths to entry and exit from work in young adulthood, what kind of early work participation patterns can be identified, how likely are transitions between different statuses, and what the determinants of different work participation patterns are. Importance of such life course approaches including childhood factors has been also stressed elsewhere (Dunn, 2010; Marmot *et al.*, 2012).

We used sequence analysis and mixture Markov models to analyse trajectories of monthly work participation statuses. The most important benefit of this approach is its holistic perspective: rather than focusing on specific events or transitions, this approach sheds light on the entire trajectory of early work participation. Sequence analysis has been increasingly used to study the sequencing of life events including labour market participation, family formation, and health (Barban, 2013; McKetta *et al.*, 2018; Roux, Grimaud and Leray, 2019; Gao *et al.*, 2020). Using mixture Markov models for such data allows us to identify typical patterns between and dynamics within individual trajectories as well as to study the associations between childhood predictors and different types of early labour market participation patterns (Helske, Helske and Eerola, 2018; Helske and Helske, 2019).

A key hypothesis of this study is that among a birth cohort, there are latent groups of work participation that have different social and health-related determinants that can shorten working life duration and lead to unstable early careers due to, e.g., health-related absences and unemployment.

Furthermore, we hypothesise that young adults who enter the labour market earlier, have more unstable work participation, which could be explained by their more disadvantaged social background and health-related factors. Such early entry is more likely for those with no or short education, and entering to more manual work with adverse exposures.

Using register data comprising an entire birth cohort and a flexible person-oriented mixture Markov modelling approach, the study therefore aims to (1) examine what kind of work participation patterns can be identified among young adults in their early careers and (2) how parental and own socioeconomic factors and health as well as age at entry into the labour market jointly predict such patterns. We focus on age at entry in the first long-term period of paid employment and study what kind of latent classes can be identified in subsequent work participation and how the accumulation of a number of childhood and adolescence social and health-related determinants are associated with different types of work participation patterns depending on the age at entry.

2. Methods

2.1 Data and variables

The Finnish Birth Cohort 1987 (from now on FBC) (Paananen and Gissler, 2012; Ristikari et al., 2016) includes data from several registers from all children born in 1987 as well as their parents (registered mother for all children and registered father for all but 821 children). The follow-up period for the original data lasts from the perinatal period until December 2015. In this study, we use data on employment, pensions and other benefits from the Finnish Centre for Pension (ETK), study grants, unemployment benefits, and parental leave data from the Social Insurance Institution of Finland (Kela), data on unemployment periods from the Finnish Ministry of Economic Affairs and Employment, data on social assistance from the Finnish Institute for Health and Welfare (THL), education level and date of death from Statistics Finland (TK), criminal convictions from Legal Register Centre (ORK) and comprehensive school achievement data from the Finnish National Board of Education (OPH). All the register datasets were merged using personal identification numbers assigned to each Finnish resident (Gissler and Haukka, 2004). Personal identification numbers are upheld by Digital and Population Data Services Agency (DVV). More precise information on the data and variables can be found on the FBC metadata web page (Finnish institute for health and welfare, 2021).

From the full dataset we created our study sample using information on early adulthood labour market participation to create the outcome and information on own and parental social and health-related events from age 0 to age 17 for creating predictor variables (covariates). More precisely, for the outcome we created sequences of monthly work participation statuses starting from the calendar month in which the individual started their first six-month long spell in employment or entrepreneurship without being in education at the same time (at age 17 the earliest). We chose six months as the threshold as it was long enough to exclude summer jobs and other short employment spells, but not too long to end up dropping too many cohort members from the analysis. We recorded the work participation status during the first five years after the first entry or for at least 12 months if they were censored sooner. We classified the work participation statuses into eight categories, presented in Table 1.

[Table 1 here.]

Altogether 59,476 children were born in Finland in 1987. We excluded 271 individuals that had received disability allowance at the middle or highest rate at the beginning of 2005 and 455 cohort members who had received a diagnosis for an intellectual disability. Furthermore, we dropped 464 individuals that had deceased before 2005 or during the five-year follow-up after their first entry into the labour market. Lastly, we dropped 1,980 cohort members who had lived abroad for more than 12 consecutive months during the follow-up, as changes in their work participation status could not be followed using national registers, ending up with 56,306 cohort members.

Among them, 52,344 individuals had started a six-month work spell (88% of the full cohort and 93% of those that had not been excluded earlier due to disabilities, death, or living abroad). Of them, 51,871 individuals were followed for at least 12 months after their labour market entry (87% of the cohort and 96% of the target population) and 42,770 individuals for full 60 months (72% of the cohort and 79% of the target population).

The timing of the labour market entry was categorized into three categories: 17–18 (28.5% of the sample with a work spell of at least 12 months; entering at age 17 was very rare), 19–22 (50.5%), and 23–26 (21.0%). We used this categorized variable to allow for possible nonlinear relationships between the timing of entry and the work participation pattern. The cut-off points were decided based on normative educational transitions and the distribution observed in the sample.

All childhood covariates were determined from records during the time period when the cohort members were 0–17-year-olds (by December 2004). These covariates can be divided into own and parental social and health-related variables. The cohort members' own variables included sex, grade point average (GPA), criminal conviction, teenage pregnancy, and psychiatric diagnosis. The sex of the cohort members was based on the Population Information System which is upheld by DVV. GPA is a measure of school success at the end of compulsory school around age 16. The grades range from four (lowest) to ten (highest) and the GPA variable was coded into six categories (from highest to lowest: 8.50–10.00, 8.00–8.49, 7.50–7.99, 7.00–7.49, 4.00–6.99, or missing). The GPA information comes Joint Application Register managed by OPH. This register has the GPA information of the individuals who have applied to secondary level education. Most adolescents apply immediately after comprehensive school. For this reason, we recorded GPA as missing if the student had not applied for further education at all or if their application was delayed (7.8% in total). We recorded criminal record as a dichotomous variable based on the Criminal records of the national central register managed by ORK and include only the cases decided in the court by the end of follow-up. Teenage pregnancy was defined as giving a birth or having an abortion before age of 20 according to Medical Birth Register of THL. Psychiatric diagnoses were sampled from the Care Register for Health Care of THL which includes only specialized healthcare visits. Our follow-up includes visits with a diagnosis based on ICD-9 (in 1987-1995) and ICD-10 (in 1996-2004), so we defined psychiatric diagnosis as ICD-9: 290–319 and ICD-10: F00–99. Inpatient care visits were included in the variable formation from 1987 onward whilst outpatient care visits from 1998 onward due to register limitations.

Parental social and health-related covariates included high parental education, nuclear family, social assistance, teenage mother, and parental psychiatric diagnosis. High parental education was defined as whether at least one parent had a tertiary degree in Register of Completed Education and Degrees of TK. Nuclear family is a variable based on the daily housing data from Population Information System of DVV. We defined it as a dichotomous variable that captures whether the cohort member and both parents occupied the same household at the end of the follow-up when cohort members were 17-year-olds (December 2004). Social assistance variable captured receiving social assistance as per the Register of Social Assistance of THL. It was coded as having received social assistance either if the

parents had received over 12 months of social assistance during the cohort member's childhood or if there was at least one calendar year during that time in which the family received social assistance for more than 7 months. Having a teenage mother was determined from the Medical Birth Register of THL and was defined as whether the mother was less than 20 years of age when giving birth to the cohort member. Lastly, parental psychiatric diagnoses were determined using the same ICD-code ranges as for cohort members (ICD-9: 290–319; ICD-10: F00–99) if either of the parents had received any of these diagnoses.

2.2 Statistical methods

Sequence analysis (SA) is an exploratory data-driven approach that has become central to the life course perspective for understanding various trajectories (Gauthier, Bühlmann and Blanchard, 2014). It is typically used for exploratory analysis and for assessing the relationship between trajectories and other variables. For the latter, a two-step approach is typical: first, sequence analysis and cluster analysis are used for finding meaningful groups in the data; second, the covariate–sequence relationship is assessed by using cluster memberships as a categorical variable. This approach, however, has been criticized for neglecting the within-cluster variation which could potentially lead to biased results if the variation is not small and random (Studer, 2013; Piccarreta and Studer, 2019).

Recently, probabilistic modelling of social sequence data has gained increasing interest. Unlike the data-mining type approach of SA, probabilistic approaches allow for the use of statistical inference to draw conclusions about the quality of the model and the covariate–sequence relationship. The multistate model (MSM) extends the focus of the basic event history model from one event or transition to multiple potentially recurring transitions (Meira-Machado *et al.*, 2009; Steele, 2011). However, unlike SA, the MSM requires making assumptions on the dependence of transitions on time. One commonly used assumption and the approach adopted in this paper is the (first-order) Markov assumption which states that the probability of a transition only depends on the current status, not on prior history. Markovian models can be regarded as one type of MSMs that allow for flexible analysis of life course data with multiple life statuses and a large number of different transitions (Helske, Helske and Eerola, 2018; Piccarreta and Studer, 2019; Han, Liefbroer and Elzinga, 2020).

The Markov model estimates the probabilities for starting in each labour market state and transitioning between the states (initial and transition probabilities). We used an extension of the basic model, the mixture Markov model (MMM), which expects that the population consists of latent classes (clusters) with varying longitudinal employment patterns and allows for different specifications of initial and transition probabilities between the classes (see Appendix A for more information on the model specification). The MMM is similar to the *latent class model* (Han, Liefbroer, & Elzinga, 2017) in that it accounts for *unobserved heterogeneity* between the sequences but is more general as it does not assume conditional independence between time points – quite the opposite, as the idea is to estimate transition probabilities from time point to time point.

Covariates can be included in the model to explain class membership probabilities and/or initial and transition probabilities. However, allowing the same variable to influence various types of probabilities may hinder the interpretation and evaluation of its effect. Therefore, it is common practice to relate each covariate to one type of probability only (Han, Liefbroer and Elzinga, 2016). Since our interest was to use childhood covariates to predict labour market participation patterns, we constructed a model where the covariates explain class membership probabilities, allowing us to

evaluate the relationship between the timing of the labour market entry and further work participation and to predict labour market participation outcomes using childhood information.

In practice, the estimation of a complex Markovian model with covariates can be time consuming, especially if and typically when we do not know the number of latent classes beforehand (Helske and Helske, 2019). To facilitate the analysis, we used an approach described in Helske, Helske and Eerola (2016) who used SA as a preliminary step before estimation of a mixture hidden Markov model.

We chose ten clusters based on the SA and estimated the MMM with covariates predicting latent class memberships. More detailed information on the estimation process is given in Appendix A.

2.3 Software

All analyses were conducted with R (R Core Team, 2020). We used the TraMineR package (Gabadinho *et al.*, 2011) for creating the sequences and calculating sequence dissimilarities, seqHMM (Helske and Helske, 2019) for estimation and visualization of the MMM, and ggplot2 (Wickham, 2016) for other visualizations.

3. Results

3.1 Descriptive statistics

Virhe. Viitteen lähdettä ei löytynyt. shows the duration of the follow-up for men and women in the sample after their first entry into the labour market as well as the duration of their first at least six-month long work episode. We see that for most of the cohort members, the first work spell was fairly short: less than two years for the majority of the cohort, while only 20% are continuously employed for four years or more. Even if accounting for all work spells, the durations were fairly short, as illustrated in Figure B-1 in Appendix B.

[Figure 1 here]

Virhe. Viitteen lähdettä ei löytynyt. shows the mean time spent in each work participation state. The work status was the most prevalent: about 70–78% of all months of the follow-up were spent in work (slightly more for men than women). Women spent more time in parallel work and studies as well as on parental leave while men spent more time in the unknown state (living with parents, abroad, or otherwise unknown, typically military or non-military service) and in the state of unemployment or receiving social assistance.

[Figure 2 here]

Virhe. Viitteen lähdettä ei löytynyt. shows the most frequent work participation patterns in the full study sample (before any classification), when not taking timing and duration of the spells into account. Continuous work was the most frequent pattern, but only 14% of the individuals were working continuously during the first five years after their entry into the labour market. All the other patterns were much rarer and typically described temporary exits from work, most typically due to unknown status, studying parallel to work, or being unemployed. The only “permanent” exit among the most frequent patterns was exiting due to parental leave. The transitory nature of exits from work can also to some extent be seen in Figure B-2 in Appendix B – even numbers of transitions within

sequences are much more common than odd numbers (slightly less common among women due to the “permanent” exits to parental leave).

[Table 2 here]

Virhe. Viitteen lähde ei löytynyt. shows descriptive statistics of the predictors. We find that 28% of the individuals that had entered the labour market during the follow-up were aged 18 or younger and a half were aged 19–22. Men and women differed in terms of their GPA (higher for women), in having had a conviction (more often men), and their own psychiatric diagnosis (more often women).

[Table 3 here]

3.2 Latent classes of labour market participation

[Figure 3 here]

Virhe. Viitteen lähde ei löytynyt. describes the latent classes after assigning each individual to the latent class for which they had the highest membership probability, showing monthly prevalence of each status. At the start, almost all individuals were in the work state (apart from few exceptions that had a work contract but were on parental or health-related leave). Monthly transition probabilities are visualized in Figure A-4 in Appendix A..

Class A: Continuous work (most probable class for 19% of the individuals). Individuals hardly leave paid work – the probability for exiting work is 1% – but in the rare occasions that they do, they swiftly return back to work.

Class B: Work and early gap (17%). The individuals mainly work but often have an early gap for which we have no data. Likely for many this gap consists of military or non-military service which is mandatory by law for all men in Finland (and a voluntary option for women, quite rarely used). Another typical reason is being supported by parents. Other exits from work (mainly related to health and unemployment) are also a bit more probable than in the continuous work class but tend to be short in duration.

Class C: Work and short parental leaves (4%). In this male-specific class, men mainly work but they also have short exits, typically for parental leave. Early spells of unknown status are fairly common.

Class D: Parental leave and health-related exits (12%). This female-specific class is mainly characterised by exits from work to longer parental leave and short exits due to health-related non-attendance.

Class E: Unemployment (9%). Individuals exit work for unemployment (some are also only receiving social assistance without other main benefits or work-related income). Many transitions between unemployment and employment and also other short exits – mainly to parental leave or health-related or unknown status – are fairly common.

Class F: Studies and work (6%). Flexible transitions between work and studies and both in parallel. Entering other states is rare.

Class G: Studies and parallel work (13%). Individuals mainly study and work in parallel. They tend to have a later entry into studies and and/or their studies are short in duration. Entering other states than work is rare.

Class H: Gap, studies and summer jobs (7%). Recurrent transitions between studies and work, likely indicating full-time studies and summer jobs. In addition, individuals have spells of unknown status, typically soon after first entering the labour market but also later during the follow-up. Many of these gaps likely indicate participation in (non-)military service or being supported by parents. Individuals rarely enter other states, but if they do, they tend to exit swiftly.

Class I: Gap, studies and parallel work (8%). Mainly work and studies in parallel. Also some unknown (military-service type) gaps of unknown status, mainly soon after first entering the labour market. Entering other states is rare and short in duration.

Class J: Fragmented studies and health-related exits (5%). Students who typically lack the usual pattern of alternating between or combining studies and work. Instead, they have spells of parental leave, unemployment or health-related exits between study spells. This is also the most likely class for individuals with long health-related non-attendance spells, sometimes even without studies.

3.3 Predictors to early work participation patterns

[Figure 4 here]

Virhe. Viitteen lähde ei löytynyt. shows estimates and 95% confidence intervals for the timing of the labour market entry and cohort members' own and parental social and health-related predictors. The estimates are interpreted similarly to the log odds ratios from multinomial logistic regression; the *Continuous work* class was used as the reference. Log odds ratios are difficult to interpret as such, but due to the large variation in the estimates, the figure would be illegible with odds ratios. We do, however, show odds ratios in supplementary material and in the interpretations of each variable. Exact estimates for variable coefficients are given in supplementary material along with the rest of the estimated model parameters (Table C-1 for log-odds and Table C-2 for exponents).

Age at labour market entry

The later the entry into the labour market, the more likely the *Continuous work* class (class A) was and the less likely the study classes (F–I) and the *Work and early gap* class (class B) were. A later entry (at age 23–26) also predicted somewhat lower odds for *Work and short parental leaves* (class C), *Parental leave and health-related exits* (class D), and *Fragmented studies and health-related exits* (class J). The only positive estimate was found in the *Unemployment* class (class E), indicating an increase of 30% in the odds for those with a later entry rather than an early entry (all else equal).

The relationship between age at labour market entry and early work participation pattern in early adulthood seemed fairly linear.

Own social and health-related predictors

Sex. The *Parental leave and health-related exits* (class D) was basically an all-female class (probabilities close to 0 for all men) and the *Work and short parental leaves* class (class C) an all-male class (probabilities close to 0 for all women). Women were also much more likely to be

members of the *Studies and work* (3.5-fold odds) and *Studies and parallel work* (2-fold odds). Male sex predicted higher odds for all other classes, most notably in *Work and early gap* (4-fold odds) and *Gap, studies, and summer jobs* (3-fold odds).

GPA. Higher GPAs predicted higher odds for membership for all classes characterised by studies and work (classes F–I), whereas lower GPAs predicted higher probabilities for all other classes. The exception was the *Work and short parental leaves* class where the differences were statistically insignificant (and close to 0) for all GPAs between 7 and 10.

Conviction. Having been convicted predicted higher odds for membership in the *Unemployment* class (2-fold) and in the *Work and short parental leaves* class (50% increase). Not having been convicted predicted higher odds of membership in the *Work and early gap* (by 35%) and three classes of combining studies and work (G–I; by 58–360%).

Teenage pregnancy. Own teenage pregnancy predicted an increase in the odds for membership by 57% in the *Parental leave and health-related exits* and by 51% in the *Unemployment* class, and reduced the odds for membership in three study classes (F, H, and I).

Psychiatric diagnosis. Own childhood psychiatric diagnosis (before age 18) predicted higher odds for membership in the same classes that also had higher odds for individuals with a low GPA, most notably for *Fragmented studies and health-related exits* (3.5-fold odds), *Unemployment* (2.5-fold odds), and *Parental leave and health-related exits* (73% increase in odds).

Parental social and health-related predictors

Parental education. Parental education was a significant predictor of early work participation patterns, all else equal and even after controlling for own educational attainment. More specifically, parents' high education predicted higher odds of membership in all study classes (by 70–248% for classes F–I and 15% for class J). Parent's average or low education, on the other hand, predicted an increase in the odds of membership in the *Unemployment* class by 62%, in the *Parental leave and health-related exits* by 35%, and in the *Work and parental leaves* by 21%.

Nuclear family. Having grown up in a nuclear family predicted higher odds of membership in classes *Studies and work*; *Gap, studies, and work*; and *Gap studies, and parallel work* by 36–51% and in classes *Work and early gap* and *Work and short parental leaves* by 16–22%. Having grown up in a non-nuclear family predicted higher odds for classes *Studies and parallel work*, *Fragmented studies and health-related exits*, and *Unemployment* by 14–17%.

Social assistance. Receiving social assistance in childhood predicts higher odds for membership in the *Unemployment* class by 88% and in classes *Work and short parental leaves*, *Fragmented studies and health-related exits*, and *Parental leave and health-related exits* by 36–57%. Not having received social assistance predicted higher odds for membership in all classes of combining studies and work (F–I).

Teenage mother. Being born to a teenage mother predicted higher odds for membership in the *Parental leave and health-related exits* class by 75%, in the *Work and short parental leaves* by 68%, in the *Unemployment* class by 38%, and in the *Fragmented studies and health-related exits* by 32%. Being born to a mother who was at least 20 years old predicted higher odds for membership in the *Gap, studies and parallel work* class by 77%, in the *Gap, studies, and summer jobs* class by 68%, and in the *Work and early gap* class by 32%.

Psychiatric diagnosis. When accounting for all other predictors, having had a parent with a psychiatric diagnosis did not predict higher odds for membership in any of the latent classes (all positive coefficients were small and not statistically significant). However, *not* having a parent with a psychiatric diagnosis predicted slightly higher odds for membership in the *Work and early gap* class (by 20%) and in the *Work and short parental leaves* class (by 15%).

3.4 Accumulation of (dis)advantage

In this section, we look more closely into the accumulation of childhood advantage and disadvantage, by illustrating with a number of examples how different kinds of backgrounds (combinations of predictors) and different ages at the labour market entry predicted different work participation patterns. We present five typical cases defined by selected risk or protective factors. The cases and their respective risk/protective factors are: (1) *Reference case* with having had a nuclear family, intermediate GPA, low/average parental education, and no risk factors; (2) *Accumulated advantage* with nuclear family, high GPA, high parental education, and no risk factors; (3) *“Typical” accumulated disadvantage* with non-nuclear family, low GPA, social assistance, and parental psychiatric diagnosis; (4) *Female-typical accumulated disadvantage* with non-nuclear family, low GPA, social assistance, and teenage pregnancy; and (5) *Male-typical accumulated disadvantage* with non-nuclear family, low GPA, social assistance, conviction. Table D-1 in Appendix D shows the exact definitions for each example case.

All but the sex-typical cases are shown separately for women and men. The *typical accumulated disadvantage* case was chosen according to the data: it was the most frequent combination of predictors among cohort members that had at least four risk factors. The sex-typical cases were chosen accordingly, but only among profiles that had differing frequencies between the two sexes.

[Figure 5 here]

Virhe. Viitteen lähdettä ei löytynyt. shows the predicted probabilities for each latent class for the example cases. For all cases, the continuous employment pattern was the more likely the later the individual entered the labour market. There are considerable differences in the levels though: individuals entering the labour market after age 23 with a “reference” background or an advantaged background had much higher probabilities compared to the disadvantaged cases, and the probabilities were the lowest for individuals with a sex-typical disadvantaged background. Among the late entrees, the continuous work class was the most likely pattern for women and men with an advantaged background while for women in the other groups the *Long parental leave and health-related exits* class was the most likely (43–53%) and for the disadvantaged men the most likely class was the *Unemployment* class (58–62%). For late male entrees with a reference background, there was more variation: the probabilities were about equally likely (14–18%) for classes *Work and early gap*, *Work and short parental leaves*, and *Unemployment*.

For the younger entrees with a reference background and especially with an advantaged background, entering studies was a more likely pattern than continuous employment. For the youngest entrees with a reference background, women’s probabilities were divided between *Work and early gap*, *Parental leave and health-related exits*, and the four study classes of combining studies and work (of which the *Studies and parallel work* was the most likely with a 24% probability). For the youngest male entrees with a reference background, the *Work and early gap* class was the most likely (40%).

For the individuals with a disadvantaged background, the unemployment class was likely, the more likely the later the entry into the labour market was. Women had high probabilities for the *Parental-leave and health-related exits* class. Among the men and women from disadvantaged backgrounds that did enter studies, the *Fragmented studies and health-related exits* class was the most likely for almost all groups, with the exception of the youngest female entrees, among whom the *Studies and parallel work* class was slightly more likely.

4. Discussion

4.1 Main findings

This study sought to identify distinct clusters of work participation among young employees, transitions between employment statuses, and their association with age at entry into the labour market as well as their life course social and health-related determinants. Using data from an entire birth cohort, we could follow individuals from their birth until young adulthood, entry into paid employment and different work participation patterns and transitions between employment statuses after the entry.

Labour market entry and exit patterns were measured using register data. We could consider all relevant routes of exit in young adulthood: in addition to unwanted, involuntary health-related exit and unemployment, we could also consider more voluntary types of exit such as studying and parental leave.

First of all, we observed that entry into the labour market as measured by six months in continuous employment was not a permanent entry for many, not only due to unemployment and health-related exits but also due to studies and parental leave. With our definition we were able to omit summer jobs and other short-term employment as well as working in parallel with studies, but not complete gap years while waiting for (another) attempt to enter higher education. Such gap years are fairly common in the Finnish context: entrance into tertiary education is very competitive and for this cohort typically required passing entrance exams. In 2011 when the cohort members turned 24, the median age at entrance was 20 for universities and 21 for polytechnics (Statistics Finland, 2013), indicating that many young adults had to apply multiple times before being admitted, during which time many of them worked. Another typical reason for exiting unemployment in early adulthood is military and non-military service (compulsory for men, voluntary for women), of which we had indirect evidence via lack of other register data at the typical time of service.

Using sequence analysis and the mixture Markov model we were able to identify different patterns in early labour market participation, to study dynamics within different types of patterns, and study the association of typical patterns and early life course determinants. A 10-class solution was selected to most meaningfully describe the different work participation patterns.

Regarding the timing of labour market entry, we were interested in whether it was predictive of future labour market attachment and whether the relationship was linear or nonlinear. During the first five years we observed a fairly linear relationship between the age at entry and early work participation pattern: individuals entering the labour market later were more likely to be in continuous employment thereafter (all else equal). When interpreting these results, we have to keep in mind that due to the nature of our data we were only able to study individuals entering the labour market by January 2015 when most of the cohort members were aged 27 (87% of the cohort), and only up to age 24 for a full

five-year follow-up (74% of the cohort). Given that the median age for getting a lower tertiary degree was as high as 26 in Finland in 2011 (Statistics Finland, 2013), some of the cohort members had not yet entered the labour market at least partly due to ongoing studies (19% of the cohort members who were excluded due to not entering the labour market, 843 cases, were students in 2015). On the other hand, with this analytical set up we are also missing the individuals who had yet to experience their labour market entry at age 27 for other reasons such as long-term ill health. It is possible that the linear association would not hold for individuals entering the labour market at later ages, but this remains to be confirmed with further studies.

Regarding both own and parental social- and health-related childhood factors, latent classes show clear patterns of accumulated advantage and disadvantage as hypothesised. As expected, we observed more advantaged background to be related to exits due to studies (or a combination of studies and work) or – when following a late entry – stable employment, while disadvantaged background factors were more often related to unstable work and long-term exits from the labour market.

While the continuous work pattern was about equally likely among men and women, we observed clear gendered pathways in the other groups. Women often had exits from work for long periods of parental leave – especially women from more disadvantaged backgrounds. Men, on the other hand, often had short gaps shortly after the labour market entry (likely due to military or non-military service) and seemed to have higher risks for frequent or long-term unemployment and social assistance. We also observed a male-typical parental leave pattern consisting of short-term exits from work. When combining studies and work, women more often had flexible transitions between work and studies and a combination of both, while the pattern of full-time studying and full-time summer jobs was more common among men.

Higher GPAs were related to exiting work for studies, while lower GPAs increased the risk of exits from work for other reasons. Parent's higher education further increased the probability of entering studies while low background was related to higher risk of long-term exits from work for other reasons, even after accounting for own educational attainment in adolescence. In fact, we observed that children with average school success but highly educated parent(s) had similar predicted probabilities for labour market participation patterns as highly-achieving children from less educated homes. This is well in line with the risk aversion theory of social mobility research that has shown that children from higher social origins are more likely than children from lower social origins to be compensated for the risks related to lower educational accomplishments and failure (Breen and Goldthorpe, 1997; Heiskala, Erola and Kilpi-Jakonen, 2020).

Having been convicted was a strong risk factor for men for long-term or frequent exits from employment. Most notably, exits due to unemployment and living on social assistance were very likely, especially after a late entry into the labour market and in a case of accumulation of risk factors. This is in line with previous findings from the same data (Ristikari *et al.*, 2016).

Teenage pregnancy and having a teenage mother both predicted long parental leaves and short health-related exits and frequent and long-term unemployment. The former was a particularly strong predictor for the earlier entrees. This is also in line with previous studies (Merikukka *et al.*, 2018; Lallukka *et al.*, 2019).

Own psychiatric diagnosis increased the risk of long-term exits from work, including studies without work spells in between. Parent's psychiatric diagnosis did not (independently) predict higher risks for non-participation patterns compared to continuous employment, all else equal. This finding is

somewhat contrary to an earlier study from the same cohort that did find an independent effect of parent's mental disorders on offspring's psychiatric work disability and social disadvantage in adolescence that was not completely mediated by offspring's mental disorders (Halonen *et al.*, 2019b). It is possible that this difference in findings stems from the different formulations of the outcomes – as we focus on the big picture (general and multifaceted labour market participation pattern) we may miss some of the finer details.

Regarding other family background factors, having grown up in a nuclear family was a fairly weak predictor alone. It did, however, somewhat increase the probability of exits due to studies or exits for other reasons that are short in duration, while non-nuclear family increased the probabilities for unemployment and fragmented studies but also for later or shorter studies in parallel with work. The association of family structure and educational attainment is well in line with previous inequality research: numerous studies have shown that parental separation is related to lower educational attainment (McLanahan and Sandefur, 1994; Amato, 2000, 2010; McLanahan and Percheski, 2008; Bernardi and Radl, 2014).

Receiving social assistance in childhood predicted higher probability for exits from work for other reasons than studies, while not receiving social assistance increased the probability of exiting work for studies. The results are in line with previous research suggesting that receiving social assistance in childhood predicts higher probability for labour market exclusion (Bäckman and Nilsson, 2007).

While some risk factors were clearly independently predictive of higher risks, this was not the case with all of them but rather the most elevated risks for less favourable labour market participation patterns were associated with the accumulation of different risk factors.

4.2 Methodological considerations

Unique datasets available for this study enabled detailed examination of work participation patterns during early careers among young adults and their social and health-related determinants. The use of complex, large register-based merged datasets and sophisticated person-oriented analysis methods further provided unique opportunities to produce novel evidence with implications at various levels. These innovative, novel approaches are needed and have been encouraged to be applied in the field of epidemiology (Ness, 2013).

Generally, the coverage and quality of the register data used in this study are exceptionally good. The datasets are large with statistical power to detect smaller differences and pinpoint risk groups and examine also rarer determinants of work participation. The inclusion of an entire birth cohort, to study social and health-related factors of both parents and their offspring in determining work participation is a unique strength of the study. At the same time, it is important to keep in mind that the information on services is based on usage, not need.

Mixture Markov models allowed us to examine longitudinal outcomes and multiple predictors simultaneously as well as to calculate predicted probabilities for following different types of patterns based on protective and risk factors experienced during childhood. In comparison to more traditional event history models, we were able to account for different types of exits from employment as well as the following sequence of events with our approach. Furthermore, by using sequence analysis we could have been able to study more complex patterns, but this approach also comes with its own

issues, most importantly the fact that the two-step process of first finding clusters with SA and cluster analysis and then taking cluster memberships as fixed at the modelling stage is often problematic and can even lead to wrong conclusions about the nature and existence of the studied relationships (e.g., Warren et al. 2015, Piccarreta and Studer 2019, and Helske et al. 2021). With the chosen Markovian approach, we were able to study not just specific exits but a longer trajectory, often including a number of transitions, and also to take the uncertainty of the classification into account better than with other methods. For example, we were able to differentiate between, e.g., short-term and long-term exits due to parental leave (classes C and D) and different types of patterns of exits due to studies (e.g., differentiating the less favourable pattern of fragmented studies from more favourable combinations of work and studies). Furthermore, we were able to study the probabilities of following different types of labour market participation patterns by taking the uncertainty of memberships into account.

In a mixture model, the classifications of the outcome and the effects of the predictors are modelled jointly; the purpose is to find balance between an optimal partitioning of the trajectories (latent classes) and the mixing probabilities (including the effects of the predictors). This approach is different to one where the classification of the outcome is done separately from modelling the effects of the predictors, and is arguably a more correct approach for analysing heterogeneous populations when we assume that the predictors have an effect (as we do). Due to the different methodological approach, however, direct comparisons with earlier more narrowly focused studies can only be tentative. It is also important to keep in mind that the analyses here are based on prediction, not on causal estimation. Selection based on attributes other than those that were directly studied was not accounted for.

5. Conclusions

There are varied work participation patterns and transitions between employment statuses among young employees. More in-depth understanding of these patterns and their determinants is important to plan targeted interventions and to promote more stable work participation among young adults.

Information about work participation patterns and their determinants can be used in efforts to promote work participation, particularly considering young people's entry into their first job, their subsequent transitions in the labour market alongside the opportunities for re-employment, and the determinants for different patterns. If people can have a sustained work participation and more stable work careers, this also has a notable societal impact in workplaces and beyond.

On the one hand, early (and possibly also very late) entry into paid employment could reflect unstable work participating during early working life span. This is related to low socioeconomic status, low education, and own health problems. On the other hand, early entry is fairly common, and in addition to more unstable labour market participation it is also related to gap years before studies (which in turn predict better labour market attachment in the future), so an early entry is likely not a risk factor as such. It is important to target policies and actions in the high risk and most disadvantaged groups in particular, using the information of this study. Intervention studies, e.g. supporting the most disadvantaged families, teenaged mothers, children of teenaged mothers, and individuals with early psychiatric problems, could confirm, if such support can be linked to later more stable working life as well.

Data availability

The data that support the findings of this study are not publicly available due to data protections laws and regulations. Thus, strict restrictions apply to the availability of these data, which were used solely under the permission from the national, administrative register data holders. Data are available from the register data holders upon reasonable request pending their permission.

References

- Achterberg, T. J. *et al.* (2009) 'Factors that promote or hinder young disabled people in work participation: A systematic review', *Journal of Occupational Rehabilitation*. Springer, 19(2), pp. 129–141. doi: 10.1007/s10926-009-9169-0.
- Amato, P. R. (2000) 'The consequences of divorce for adults and children', *Journal of Marriage and Family*. doi: 10.1111/j.1741-3737.2000.01269.x.
- Amato, P. R. (2010) 'Research on divorce: Continuing trends and new developments', *Journal of Marriage and Family*. doi: 10.1111/j.1741-3737.2010.00723.x.
- Bäckman, O. and Nilsson, A. (2007) 'Childhood Poverty and Labour Market Exclusion. Findings from a Swedish Birth Cohort', *Arbetsrapport 2007*.
- Barban, N. (2013) 'Trajectoires familiales et santé: Une approche sous l'angle de parcours de vie', *European Journal of Population*, 29(4), pp. 357–385. doi: 10.1007/s10680-013-9296-3.
- Baum, L. E. and Petrie, T. (1966) 'Statistical inference for probabilistic functions of finite stateMarkov chains', *The Annals of Mathematical Statistics*. JSTOR, 67(6), pp. 1554–1563. doi: 10.1214/aoms/1177699147.
- Bernardi, F. and Radl, J. (2014) 'The long-term consequences of parental divorce for children's educational attainment', *Demographic Research*. doi: 10.4054/DemRes.2014.30.61.
- Breen, R. and Goldthorpe, J. H. (1997) 'EXPLAINING EDUCATIONAL DIFFERENTIALS', *Rationality and Society*. SAGE Publications Ltd, 9(3), pp. 275–305. doi: 10.1177/104346397009003002.
- Coggon, D. *et al.* (2013) 'International variation in absence from work attributed to musculoskeletal illness: findings from the CUPID study', *Occupational and Environmental Medicine*, 70(8), pp. 575 LP – 584. doi: 10.1136/oemed-2012-101316.
- Croft, P., Blyth, F. M. and Van Der Windt, D. (2011) *Chronic Pain Epidemiology: From Aetiology to Public Health*. Oxford University Press. doi: 10.1093/acprof:oso/9780199235766.001.0001.
- Dunn, K. M. (2010) 'Extending conceptual frameworks: Life course epidemiology for the study of back pain', *BMC Musculoskeletal Disorders*. BioMed Central, 11(1), pp. 1–11. doi: 10.1186/1471-2474-11-23.
- Farioli, A. *et al.* (2014) 'Musculoskeletal pain in Europe: The role of personal, occupational, and social risk factors', *Scandinavian Journal of Work, Environment and Health*. Europe PMC Funders, 40(1), pp. 36–46. doi: 10.5271/sjweh.3381.
- Finnish Institute for Health and Welfare (2021) *Meta Visualisation of the Finnish Birth Cohort Studies*. Available at: <https://fbc-studies.github.io/meta-visu/?lang=en&ds=finnish-birth-cohorts> (Accessed: 10 February 2021).

- Gabadinho, A. *et al.* (2011) 'Analyzing and Visualizing State Sequences in R with TraMineR', *Journal of Statistical Software*. American Statistical Association, 40(4), pp. 1–37. doi: 10.18637/jss.v040.i04.
- Gao, N. *et al.* (2020) 'Paid work, household work, or leisure? Time allocation pathways among women following a cancer diagnosis', *Social Science and Medicine*. Elsevier Ltd, 246, p. 112776. doi: 10.1016/j.socscimed.2019.112776.
- Gauthier, J.-A., Bühlmann, F. and Blanchard, P. (2014) 'Introduction: Sequence Analysis in 2014', in *Advances in Sequence Analysis: Theory, Method, Applications*. New York, NY: Springer-Verlag, pp. 1–17.
- Gissler, M. and Haukka, J. (2004) 'Finnish health and social welfare registers in epidemiological research', *Norsk Epidemiologi*, 14, pp. 113–120.
- Halonen, J. I., *et al.* (2019a). Socioeconomic and health-related childhood and adolescence predictors of entry into paid employment. *European Journal of Public Health*, 29(3), pp. 555–561
- Halonen, J. I. *et al.* (2019b) 'Pathways from parental mental disorders to offspring's work disability due to depressive or anxiety disorders in early adulthood—The 1987 Finnish Birth Cohort', *Depression and Anxiety*. Blackwell Publishing Inc., 36(4), pp. 305–312. doi: 10.1002/da.22847.
- Han, S. Y., Liefbroer, A. C. and Elzinga, C. H. (2020) 'Mechanisms of family formation: an application of Hidden Markov Models to a life course process', *Advances in Life Course Research*. Elsevier BV, 43, p. 100265. doi: 10.1016/j.alcr.2019.03.001.
- Han, Y., Liefbroer, A. C. and Elzinga, C. H. (2016) 'Understanding social-class differences in the transition to adulthood using Markov chain models', in Ritschard, G. and Studer, M. (eds) *Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016*, pp. 155–177.
- Heiskala, L., Erola, J. and Kilpi-Jakonen, E. (2020) 'Compensatory and Multiplicative Advantages: Social Origin, School Performance, and Stratified Higher Education Enrolment in Finland', *European Sociological Review*. Oxford University Press (OUP). doi: 10.1093/esr/jcaa046.
- Helske, S., Helske, J., & Chihaya, G. K. (2021). From sequences to variables – Rethinking the relationship between sequences and outcomes. SocArXiv preprint. doi:10.31235/osf.io/srxag
- Helske, S. and Helske, J. (2019) 'Mixture hidden Markov models for sequence data: The seqhmm package in R', *Journal of Statistical Software*, 88(1). doi: 10.18637/jss.v088.i03.
- Helske, S., Helske, J. and Eerola, M. (2016) 'Analysing Complex Life Sequence Data with Hidden Markov Modelling', in *LaCOSA II: Proceedings of the International Conference on Sequence Analysis and Related Methods*.
- Helske, S., Helske, J. and Eerola, M. (2018) 'Combining Sequence Analysis and Hidden Markov Models in the Analysis of Complex Life Sequence Data', in. Springer, Cham, pp. 185–200. doi: 10.1007/978-3-319-95420-2_11.
- Kela, E. & (2010) 'The Statistical Yearbook of Pensioners in Finland', *Official Statistics of Finland. Social Protection*.
- Lallukka, T. *et al.* (2019) 'Determinants of long-term unemployment in early adulthood: A Finnish birth cohort study'. doi: 10.1016/j.ssmph.2019.100410.
- Lallukka, T. *et al.* (2020) 'Unemployment trajectories and the early risk of disability pension among young people with and without autism spectrum disorder: A nationwide study in Sweden', *International Journal of Environmental Research and Public Health*, 17(7). doi:

10.3390/ijerph17072486.

Marmot, M. *et al.* (2012) 'WHO European review of social determinants of health and the health divide', *The Lancet*, 380(9846), pp. 1011–1029. doi: 10.1016/S0140-6736(12)61228-8.

McKetta, S. *et al.* (2018) 'Social sequencing to determine patterns in health and work-family trajectories for U.S. women, 1968–2013', *SSM - Population Health*, 6, pp. 301–308. doi: 10.1016/j.ssmph.2018.10.003.

McLanahan, S. and Percheski, C. (2008) 'Family Structure and the Reproduction of Inequalities', *Annual Review of Sociology*. doi: 10.1146/annurev.soc.34.040507.134549.

McLanahan, S. and Sandefur, G. (1994) 'Growing up with a single parent: What helps, what hurts', *Cambridge, MA: Harvard University Press*.

Meira-Machado, L. F. *et al.* (2009) 'Multi-state models for the analysis of time-to-event data', *Statistical Methods in Medical Research*. SAGE PublicationsSage UK: London, England, pp. 195–222. doi: 10.1177/0962280208092301.

Merikukka, M. *et al.* (2018) 'Childhood determinants for early psychiatric disability pension: A 10-year follow-up study of the 1987 Finnish Birth Cohort', *International Journal of Social Psychiatry*. SAGE Publications Ltd, 64(8), pp. 715–725. doi: 10.1177/0020764018806936.

Ness, R. B. (2013) 'Counterpoint: The Future of Innovative Epidemiology', *American Journal of Epidemiology*. Oxford University Press, 177(4), pp. 281–282. doi: 10.1093/aje/kws416.

Nurminen, M. (2011) *Working-Life Expectancy in Finland: Development in 2000–2009 and Forecast for 2010–2015*. Eläketurvakeskus. Available at: <https://www.julkari.fi/handle/10024/129058> (Accessed: 22 February 2021).

OECD (2010) *Sickness, Disability and Work: Breaking the Barriers. A synthesis of findings across OECD countries*.

Paananen, R. and Gissler, M. (2012) 'Cohort Profile: The 1987 Finnish Birth Cohort', *International Journal of Epidemiology*. Oxford Academic, 41(4), pp. 941–945. doi: 10.1093/ije/dyr035.

Patel, V. *et al.* (2007) 'Series Mental health of young people: a global public-health challenge', *www.thelancet.com*, 369. doi: 10.1016/S0140.

Piccarreta, R. and Studer, M. (2019) 'Holistic analysis of the life course: Methodological challenges and new perspectives', *Advances in Life Course Research*, 41, p. 100251. doi: 10.1016/j.alcr.2018.10.004.

R Core Team (2020) 'R: A language and environment for statistical computing'. Vienna, Austria: R Foundation for Statistical Computing,. Available at: <https://www.r-project.org/>.

Rabiner, L. (1989) 'A tutorial on hidden Markov models and selected applications inspeech recognition', *Proceedings of the IEEE*. IEEE, 77(2), pp. 257–286. doi: 10.1109/5.18626.

Ristikari, T. *et al.* (2016) *Suomi nuorten kasvu ympäristönä – 25 vuoden seuranta vuonna 1987 Suomessa syntyneistä nuorista aikuisista [Finland as a growth environment for children – 25-year follow-up of those born in Finland 1987]*.

Roux, J., Grimaud, O. and Leray, E. (2019) 'Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis', *Statistical Methods in Medical Research*, 28(6), pp. 1651–1663. doi: 10.1177/0962280218772068.

Shiri, R. *et al.* (2021) 'Work participation and physicality of work in young adulthood and the

development of unhealthy lifestyle habits and obesity later in life: a prospective cohort study', *Occup Environ Med*, 78, p. 11. doi: 10.1136/oemed-2020-106526.

Statistics Finland (2013) *Oppilaitostilastot 2012*. Available at: http://www.stat.fi/tup/julkaisut/tiedostot/julkaisuluettelo/ykou_opla_201200_2013_10094_net.pdf.

Steele, F. (2011) 'Multilevel discrete-time event history analysis with application to the analysis of recurrent employment transitions', *Australian & New Zealand Journal of Statistics*. Wiley Online Library, 53(1), pp. 1–20. doi: 10.1111/j.1467-842X.2011.00604.x.

Studer, M. (2013) 'WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R', (24). doi: 10.12682/lives.2296-1658.2013.24.

Studer, M. and 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)*. Wiley Online Library, 179(2), pp. 481–511. doi: 10.1111/rssa.12125.

Virtanen, M. *et al.* (2020) 'Neurodevelopmental disorders among young adults and the risk of sickness absence and disability pension: a nationwide register linkage study', *Scandinavian Journal of Work, Environment & Health*. Nordic Association of Occupational Safety and Health, 46(4), pp. 410–416. doi: 10.5271/sjweh.3888.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

World Health Organization (2012a) *Ageing and health in focus in 2012, The World Health Organization Regional Office for Europe*. Available at: <https://www.euro.who.int/en/health-topics/Life-stages/healthy-ageing/news/news/2012/01/ageing-and-health-in-focus-in-2012> (Accessed: 22 February 2021).

World Health Organization (2012b) *Executive summary of the European health report 2012: Moving Europe towards health and well-being*.

Supplementary information

Figure S-1: Sequence cluster results from 2 to 15 clusters using optimal matching.

Figure S-2: Sequence cluster results from 2 to 15 clusters using Euclidean distance.

Figure S-3: Length of individual follow-up.

Figure S-4: Predicted probabilities for the (almost) reference case, changing one variable at a time.

Table S-1: Estimated coefficients for covariates from the mixture Markov model.

Table S-2: Estimated transition probabilities from the mixture Markov model.

Table S-3: Estimated initial probabilities from the mixture Markov model.

Appendix

Appendix A: Model estimation

Sequence analysis

Sequence analysis was used as an intermediary step to help with setting up informative starting values for the parameters of the more complex mixture Markov model. The results of SA were not used directly in the final analyses. As sequence dissimilarity measures we considered measures that are known to be sensitive to the duration of spells spent in one state, namely optimal matching (OM) and Euclidean distance (Studer and Ritschard, 2016). We then used Ward’s agglomerative clustering algorithm for grouping sequences into relatively homogeneous clusters. For the choice of the dissimilarity measure, we used a sample of 15,000 sequences to reduce computation times, and then chose the measure with more meaningful clusters. Finally, we used the OM measure for clustering the full sample of 60-month sequences and chose the number of clusters by subjective assessment of meaningfulness as well as using a range of cluster quality indices suggested by Studer (2013)(Studer, 2013).

[Figure A-1 here]

We chose the 10-cluster solution as producing the most meaningful classification. Up to ten clusters the substantive meaning of each cluster was useful and each cluster was clearly distinctive, while after ten clusters the substantive differences were small. As shown by Figure A-1, the cluster quality indices did not give a clear indication of one solution being preferable to another (except at five clusters, shown as a peak – this solution, however, was not distinctive enough given the complexity of the data). As the SA part was only a preliminary step, we do not present the full results here, but only give short descriptions and proportions in Table A-1. For an illustration of the number of clusters ranging from two to 15 for OM and Euclidean dissimilarities, see the supplementary material.

[Table A-1 here]

Mixture Markov model

A simple Markov (chain) model for sequence data can be described with using the following notations and probabilities:

- y_{it} : observation of individual i at time t
- s : state (or status, e.g., employed, studying, unemployed, ...)
- $\pi(s)$: Probability, that a sequence starts in state s (initial probability)
- $a(s, r)$: Probability to transition from state s to state r (transition probability)

The mixture Markov model (MMM) expects that the population consists of latent classes (clusters) with varying employment patterns and allows for different specifications of initial and transition probabilities between these classes. Each *submodel* M_k consists of initial probabilities $\pi_k(s)$ and

transition probabilities $a_k(s, r)$, with w_k defining the class membership probability for latent class k , which here depends on the covariates x_i . The (log) likelihood of the model is calculated as

$$\begin{aligned}\log L &= \sum_{i=1}^N \log P(y_i | x_i, M_k) = \sum_{i=1}^N \log \sum_{k=1}^K P(M_k | x_i) P(y_i | M_k) \\ &= \sum_{i=1}^N \log \sum_{k=1}^K P(M_k | x_i) P(y_{i1} | M_k) \prod_{t=2}^T P(y_{it} | y_{i(t-1)}, M_k) \\ &= \sum_{i=1}^N \log \sum_{k=1}^K w_k \pi_k(y_{i1}) \prod_{t=2}^T a_k(y_{i(t-1)}, y_{it})\end{aligned}$$

where N is the number of individuals, K is the number of classes (submodels), T is the number of time points, and x_i includes the covariates for individual i .

The maximum likelihood estimates of the probabilities are typically calculated with the expectation–maximization (EM) algorithm (Baum and Petrie, 1966; Rabiner, 1989). The EM algorithm is iterative, which means that we make a guess on the values of the parameters (or often use random starting values) and the algorithm then changes these values until it has found a solution that (locally) maximizes the value of the likelihood function. To reduce the risk of being trapped in a poor local maximum (and ending up with a suboptimal model), estimation should be started multiple times from different starting values. See Figure A-2 for a simple illustration of the problem. The full estimation process can be extremely slow, so using meaningful starting values are often needed for finding an optimal solution in a reasonable time.

[Figure A-2 here]

To facilitate the analysis, we used an approach described in Helske, Helske and Eerola (2016) who used SA as a preliminary step before estimation of a mixture hidden Markov model. The outline of our analytical strategy is thus as follows:

1. *Apply sequence analysis and cluster analysis* to determine the number of latent classes. At this step, we only accounted for individuals whose work participation was recorded for full 60 months, as accounting for missing information in SA can be problematic. We then used the optimal matching algorithm for calculating sequence dissimilarities and Ward’s clustering algorithm for identifying groups with similar labour market participation trajectories.
2. *Estimate Markov models independently for each cluster* of work participation sequences from step 1. At this step, we used the clusters of 60-month work participation sequences and estimated a simple Markov model independently for each cluster.
3. *Estimate a MMM with covariates* using all sequences with 12–60 month follow-up and the number of clusters decided at step 1 as the number of latent classes and the estimated initial probabilities π_s and transition probabilities a_{sr} from step 2 as meaningful starting values for the submodels of the MMM. We tested between randomized starting values (using the default randomization option for multiple EM restarts in the seqHMM package) and chose the final best-fitting model according to the likelihood function.

Fixing the number of submodels based on the results of step 1 saved us from having to estimate several MMMs with different numbers of latent classes, and the probabilities estimated at step 2

allowed us to greatly reduce the estimation time by using starting values that are plausible and likely relative close to the model with the highest likelihood.

In order to avoid being trapped in a local maximum, we estimated the model 16 times (of which 14 estimations were successful) and as per convention chose the model with the highest likelihood as the final model (Figure A-3).

[Figure A-3 here]

[Table A-2 here]

Table A-2 presents the mean probabilities of each latent class by the most probable latent class. The diagonal (bold) shows the mean probability of the respective class; e.g., the average membership probability for class A (Continuous work) was 75% for individuals with the highest class probability for class A and in general these ranged from 75% for class 1 to 87% for class C (Work and short parental leaves). High probabilities on the diagonal and low probabilities elsewhere indicate highly distinctive classes with little membership uncertainty. As can be seen, we had some non-negligible uncertainty for some clusters, but nothing too alarming. The highest off-diagonal probability was found for class A with an 11% probability for being assigned to class B (Work and early gap) instead, so even in the worst case the probability on the diagonal was almost 7-fold to the highest off-diagonal probability.

[Figure A-4 here]

Figure A-4 shows the transition probabilities between labour market states in each latent class A–J. The numerical values for the estimated parameters of the mixture Markov model are given as supplementary material.

Appendix B: Descriptive statistics

[Figure B-1 here]

[Figure B-2 here]

Appendix C: Estimated model parameters

[Table C-1 here]

[Table C-2 here]

Appendix D: Variable definitions for exemplary cases for predicted probabilities

[Figure D-1 here]

[Table D-1 here]