Case Studies of Combining Sequence Analysis and Modelling

Mervi Eerola

1 Introduction

The recent two decades have shown that sequence analysis is a valuable tool for life course analysis. While the significance of the past and future is of fundamental importance in event-history analysis, sequence analysis, which in its basic form is ignorant to this distinction, seems to highlight the diversity of life patterns in a way that cannot be achieved with traditional statistical modelling. Several improvements and contributions either to the dissimilarity measures or to the cost matrix have been suggested to the original version. However, the question still remains whether the colourful figures of individual index plots or state distribution plots merely pose interesting hypotheses and further questions rather than provide analytic answers to the causes of life course differences. Therefore, current consensus in this research area seems to emphasize combining the benefits of both approaches. In this methodological paper, I shall present three case studies of life course analysis in which the clustering of the sequences has been combined, or contrasted, with modelling. Two of the studies are already published and one is an ongoing project. The results and complete versions can be found in the References. In the Discussion, the experiences of using both methods are compared in more detail and their role in life course analysis is evaluated.

M. Eerola (✉)
Centre of Statistics, University of Turku, Turku, Finland
e-mail: mervi.eerola@utu.fi

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_3

2 Case Study 1: Prediction of Excess Depressive Symptoms and Life Events

This study investigated how the timing and pattern of certain life events, here partnership formation and steady employment, affect the prediction of parenthood, especially remaining childless, and whether this is associated with excess depressive symptoms in middle age. The participants of the Finnish Jyväskylä Longitudinal Study of Personality and Social Development (JYLS), born in 1959, were from 12 randomly selected second-grade classes in Jyväskylä, Central Finland. They were followed from age 8 to 50. The original sample consisted of 173 girls and 196 boys. A life history calendar (LHC) was used to 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 50 during interviews in which 275 participants gave reports based on memory and visual aids provided by the LHC-sheet. Since both partnership formation and career events can have variable patterns in time, and be interpreted as 'states' also, we were interested in investigating what information probabilistic multistate models on one hand, and sequence analysis, on the other hand, can provide about the study question.

2.1 Multistate Models

We considered the life course events in an observation interval $\mathscr{T}$ as a *marked point process* (T, X) specifying the sequence of events by a pair of random variables, the occurrence time T and a mark X identifying the event (e.g. Arjas 1989). Other presentations of multistate models can be found, for example, in Andersen and Keiding (2002).

The discrete time event-specific hazard of event x is the conditional probability

$$p_x(t) = P(\Delta N_x(t) = 1 \mid \mathscr{F}_{t-1}^N)$$

of a jump of type x in time interval $t \in \mathscr{T}$ in the counting process $N_x(t) = \sum_{t \geq 1} 1(T \leq t, X = x)$, given its internal history $\mathscr{F}_{t-1}^N$ generated by the points and marks until time $t-1$. We denote the extended history by $\mathscr{H}_t = \mathscr{F}_t^N \vee \mathbf{Z}_0$, where $\mathbf{Z}_0$ are covariates fixed at time $t = 0$ already. The crude hazard that any event happens in the interval t is the sum over event-specific hazards $p(t) = \sum_x p_x(t)$.

While the hazard of an event gives a very short-term prediction of the life course, *prediction probabilities* associated with the marked point process give a long-term prediction of some random event for the *whole observed* interval in a life course (e.g. Eerola 1994; Putter et al. 2007; Eerola and Helske 2016). They are functions of event-specific hazards but provide more comparable results with sequence analysis than simple hazard analysis.

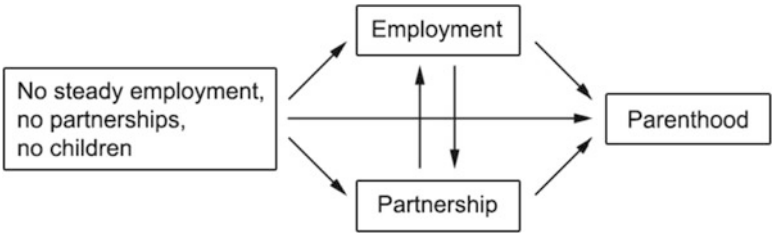

Fig. 1 A schematic model of multistate model for the JYLS data. (Source: Eerola and Helske 2016, reprinted by permission of SAGE publications)

In a multistate model the state-space can increase rapidly, so we only considered the first occurrences of partnership formation (P), child births (C), and steady employment (W) for each of which we specified event-specific hazards (that is, for $X = W$, P or C). For example, the hazard of entering steady employment when the other events have not yet occurred, is in the general form

$$p_W(t) = P(T_W = t \mid T_W \geq t, T_P \geq t, T_C \geq t)$$

where T_W is the time (age) of first steady employment, and the other event time variables are defined accordingly. As a statistical model for the discrete time hazard of event x, a piecewise constant logistic model with time-dependent indicator variables for earlier events

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

was used. The covariate vector $Z(t)$ comprises indicator variables for the events in Fig. 1, as well as piecewise constant indicators for time (age). For example, the covariate $Z_W(t) = 1$ if steady employment was reached at age t and 0 before that.

The prediction probabilities of remaining childless are sums of the probabilities of all paths of remaining childless within the prediction interval when all possible timings of partnership formation and steady employment are considered. The most complicated path results when nothing has yet happened at the prediction time t, the other paths being special cases of it. In particular, when initial partnership (P) and entering working life (W) have already occurred by the prediction time t, the prediction is simply the survival probability (for time points $0 < v \lesseqgtr w < t < u$)

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

Fixing the prediction time t, the prediction interval from $t + 1$ to u (the last observation time), or the history, results in different visual representations of the predictions. For example, fixing prediction interval and history, and identifying

t with the occurrence time of a life event, compares *factual* and *counterfactual* predictions of remaining childless, depending on whether the event x in fact occurred at t, or not. Finally, the prediction of excess depressive symptoms if the person remained childless until age 42, given the history of partnership formation and entry to stable employment, is the joint conditional probability (for $15 < t \leq 42$)

$$P(T_C > 42,\, D(42) > d^* \mid \mathscr{H}_t) = P(D(42) > d^* \mid T_C > 42)\, P(T_C > 42 \mid \mathscr{H}_t)$$

in which $D(42)$ is the score of depressive symptoms at age 42 and d^* is the median score in the study population.

2.2 Sequence Analysis

As a comparison, we performed multidimensional sequence analysis. Pairwise comparison of the original sequences using the Hamming distance resulted in eight clusters. They differed mostly in terms of timing of partnership and parenthood, and to a lesser extent in terms of the length of education. To associate these results with depression in middle age, we used the individual cluster membership indicator as a covariate in a logistic regression predicting higher than median depression score d^*, as before. This covariate was used as a 'proxy' variable for parenthood, partnership and employment history. For a generic individual, the model was

$$\text{logit}(P(D(42) > d^* \mid c)) = \alpha + \beta Z(c)$$

with $Z(c) = 1$, if the individual was a member of cluster c. Only the most deviant cluster ("singles or late family") had significantly higher odds of excess depressive symptoms than the other clusters. This supports our results with the prediction probabilities but is less informative in terms of separating individual effects or timing effect in general. More results of the study can be found in Eerola and Helske (2016).

3 Case Study 2: Antecedents and Consequences of Transitional Pathways to Adulthood

This study from developmental psychology linked two types of longitudinal data: the sequences of young people's transitions to adulthood and longitudinal data of psychological resources. The study investigated the extent to which university students' depressive symptoms, and the strategies they deploy in achievement and social situations (Nurmi et al. 1995) at the beginning of university studies

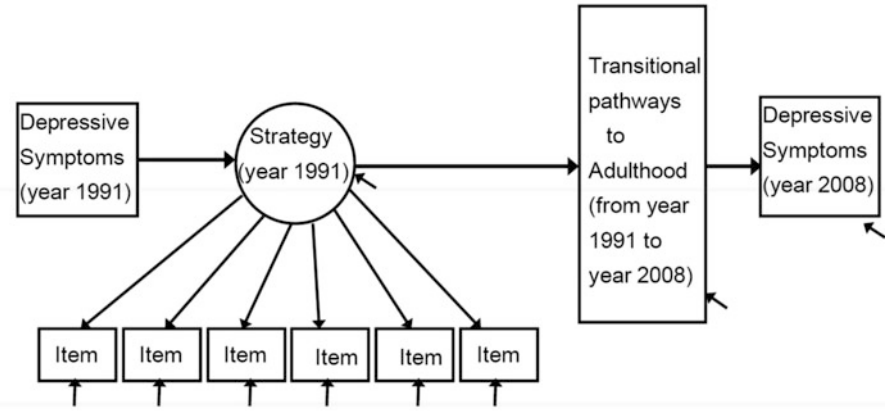

Fig. 2 A schematic model of the structural equation model for the HELS data. (Source: Salmela-Aro et al. 2014, reprinted by permission of Springer)

would predict their transitional pathways, and the extent to which the pathways contribute to depressive symptoms later in adulthood. The study was part of the Helsinki Longitudinal Student Study (HELS study). The participants were 182 undergraduates who started their studies at the University of Helsinki in 1991 and were born in or around 1970. A life history calendar was completed in 2008, retrospectively reporting on key life events during the years 1991–2008 (residence, partnership, parenthood and career).

In a previous study (Salmela-Aro et al. 2011), six clusters (transitional pathways) were identified by sequence analysis. Figure 2 shows a schematic model of the study. A strategy-specific *structural equation model* combined the submodels for achievement and social strategies with depressive symptoms at the start of follow-up, with the model for transitional pathways, and finally with the model for depressive symptoms at the end of follow-up.

3.1 Model for Strategies Accounting for Depressive Symptoms

We defined hierarchical factor models for social and achievement strategies (social optimism, social withdrawal, achievement optimism or task-irrelevant behaviour) as

$$y_{ijs} = \lambda_i \eta_{js} + \epsilon_{ijs}, s = 1, \ldots, 4, \; j = 1, \ldots, 182$$
$$\eta_{sj} = \eta_{0s} + \gamma_s x_j^{pre} + \zeta_{sj}$$

where y_{ijs} is the ith item measuring a particular strategy s and η_{sj} is the factor corresponding to that strategy for individual j, λ_i is the factor loading of item i, and ϵ_{ijs} is the unique factor of item i for strategy s. The hierarchical or

multilevel structure of the factor model (Muthén 1994; Goldstein 2011) assumes that depressive symptoms at the start of the follow-up may affect the strategies, and this is accomplished in the model by allowing each strategy factor to depend on the individual's depression score. The parameter η_{0s} is the mean level of the factor s, γ_s is the regression coefficient of depressive symptoms score x_j^{pre} at the start of studies in 1991, and ζ_{sj} refers to the individual-specific deviation from the mean factor level η_{0s}. In this model, only the items of the strategy measures and the depression scores are observable. The factors η_{sj} and the error variables ϵ_{ijs} for items and ζ_{sj} for factors are assumed zero-mean normal random variables.

3.2 Model for Transitional Pathways Accounting for Strategies

The predictive value of social and achievement strategies for the probability of following a particular pathway was studied by binary or multinomial logistic regression models with the membership indicator of a particular transitional pathway as the dependent variable. Achievement and social strategies were included as separate predictors. Since the most distinguishing factor between the six pathways was that some life events did not occur at all, or that their timing was exceptionally late, we estimated a joint model for the pathways that we called postponed (singles with slow career who never lived in a partnership during the follow-up, and slow starters, whose transitions were postponed in general). They were combined to a *postponed group* (n = 56). The remaining four pathways (fast starters, fast partnership and late parenthood, career and family, and career and unsteady partnerships) were combined to a *non-postponed group* (n = 126).

The model for the log-odds of belonging to the postponed vs. non-postponed pathway which accounts for social and achievement strategies was of the form

$$\theta_{sj}^{post} = \text{logit}(P(y_j^{post} = 1 \mid \eta_{sj})) = \alpha_s + \beta_s \eta_{sj}$$

Here y_j^{post} is the membership indicator of postponed pathway, β_s the regression coefficient of factor η_{sj} of strategy s for individual j.

3.3 Model for Depressive Symptoms When Accounting for Pathways

The model for the expected level of depressive symptoms in 2008, which accounts for pathway and indirectly also the initial level of depression and the strategies, is for each strategy

$$\mu_{sj} = \mu_0 + \delta_{1s} \exp(\theta_{sj}^{post}).$$

The parameter μ_0 is the mean level of depressive symptoms in 2008, $exp(\theta_{sj}^{post})$ the individual- and strategy-specific odds of following the postponed pathway, and δ_{1s} are the direct effects of pathway, containing also the *indirect* effects of initial-level depressive symptoms and strategies. This model was contrasted with the model of *direct* effect of the initial level of depressive symptoms

$$\mu_j = \mu_0 + \delta_2 x_j^{pre}$$

on the level of depression in 2008, where the parameter δ_2 is the direct effect of the initial level of depression symptoms in 1991, without considering the effect of strategies or pathways. This 18-year follow-up showed that depressive symptoms at the beginning of studies were associated with pessimistic and avoidant strategies in both achievement and social situations, which further predicted postponed pathways later on. The transitional pathways also contributed subsequent changes in depressive symptoms. More results of this study can be found in Salmela-Aro et al. (2014).

4 Case Study 3: Pathways to Social Exclusion

The so called NEET problem (Not in education, not in employment or training) has in many countries initiated special government policy acts to prevent young people, especially young men, from ending up in social exclusion. By social exclusion is usually meant a combination of problems such as unemployment, unfinished education, low incomes, alcohol problems, crime, bad health and unstable family conditions. These problems are linked and mutually reinforcing, and can create a vicious cycle in a person's life course. Cross-sectional studies are not helpful when trying to understand the dynamics of this process.

In this ongoing study, we are in particular interested in originating events or factors of risk accumulation and potential turning points in a young person's trajectory. We use the Finnish National Birth Cohort 1987 (around 60,000 individuals) data from the years 2005–2012 when the members were 18 to 25 years old. The cohort can be combined with all official registers, from which we restrict to data on unemployment, education and use of social benefits, episodes in mental health care and reimbursement of medicine expenses for mental illness, inpatient days due to intoxicant abuse and notifications in crime register. As usual with register data, it is important to analyse carefully which outcomes are results of the social benefit system itself to prevent from meaningless modelling.

4.1 Sequence Analysis

Sequence analysis is here used to find the most vulnerable individuals for the follow-up. Two clusters out of 12 (together around 10% or 6000 individuals) having the most fragmentary trajectories in terms of the main activity classification ("Employed", "Unemployed", "Studying", "Other") are chosen for further analysis. Several approaches can be suggested to analyse underlying lifetime periods characterised by the accumulation of risk factors.

4.2 Risk Pattern Analysis

Let $y_a = (y_{1a}, \ldots, y_{6a})'$ be individual's observed *risk pattern* at age a where $y_1, \ldots, y_6$ are indicators of the measured risk factors (outside of work force, lowest educational attainment, living on social benefits, mental health care or medication, intoxicant abuse and criminal record, respectively). This amounts to $M = 2^6$ possible binary risk patterns in each follow-up year.

We assume that η_a is a latent state with values $s \in S$ representing underlying situational characteristics of a young person at age a. As usual in hierarchical modelling, we assume that the observed indicators $\{y_{ia}\}$ are conditionally independent given η_a at each a. This is a *latent transition model* (e.g. Collins and Lanza 2010) of observed risk patterns given the dynamics of the underlying latent states.

Denote the conditional probability of risk factor i at age a by $P(Y_{ia} = 1 \mid \eta_a = s) = p_a(i \mid s)$ and the transition probability to latent state s at age a by $P(\eta_a = s \mid \eta_{a-1} = r) = q_a(s \mid r), s, r \in S$, given that the previous latent state at age $a - 1$ was r. For a generic individual, the (marginal) probability of risk patterns in the follow-up is then

$$
\begin{aligned}
P(Y = y) &= \sum_{s_a} \prod_{a=18}^{25} \{P(\eta_a = s \mid \eta_{a-1} = r) P(Y_a = m \mid \eta_a = s)\} \\
&= \sum_{s_a} \pi_s \prod_{a=19}^{25} q_a(s \mid r) \prod_{a=18}^{25} [\prod_{i=1}^{6} p_a(i \mid s)^{y_{ia}} (1 - p_a(i \mid s))^{1-y_{ia}}] \\
&= \sum_{s_a} \pi_s \prod_{a=19}^{25} q_a(s \mid r) \prod_{a=18}^{25} [\prod_{m=1}^{M} p_a(m \mid s)^{1(y_a=m)}]
\end{aligned}
$$

when summing over all possible latent states at ages $a = 18, \ldots, 25$. π_s is the initial probability of latent state s at age $a = 18$ and $1(y_a = m) = 1$ if the observed risk pattern at age a is m.

If we, in turn, assume that $\eta_a = \eta$, where η is an inherent tendency or 'trait', predisposing to marginalisation, which can be partially observed in terms of the accumulating risk factors, we would consider it as a fixed continuous latent variable. If the probability of the observed risk factors change by age, this is a dynamic *latent trait model* (e.g. Lord and Novick 1968).

A *hidden Markov model* (e.g. Rabiner 1989) has a similar probability structure but the observed states y would then be the main activity groups "Employed", "Unemployed", "Studying", "Other". To include the risk factors, we can either enlarge the state space by combining the statuses of the risk factors with the main activity groups resulting in states such as "Other/LowEdu/MHealth/Drugs/Crime" which would resemble multidimensional sequence analysis. A more natural way is to define the transition rates or transition probabilities between the four main activity groups with time-dependent covariates as in Case 1.

4.3 Predictions of Positive Trajectories

Since there already exists several studies on the prevalence of risk factors for NEET, yet another approach is to estimate predictions of *no* marginalisation, that is, predictions of integration into the labour market or in educational trajectories by age 26 when *avoiding* a particular risk factor along the developmental pathway while experiencing others. As in Case study 1, such "What if" -analyses compare two probabilities: that of an individual, initially at high risk, but who avoids a particular risk factor (mental health problems, criminal records, living on social benefits, lowest educational level) *at least until age a*, with the probability of not avoiding it, given other risk factors. Since we are interested in the effects of risk factors on the *positive outcome* (integration into labour market or education), it is the difference of these probabilities that allows us to evaluate the effect of timing on the positive trajectory.

5 Discussion

This paper has illustrated three case studies which combine sequence analysis and probabilistic modelling in life course analysis. In the first, prediction probabilities for individual's entire observed life trajectory were estimated to find out how the timing of certain life events affects the prediction of an outcome. Since all of the life events could repeat in time, sequence analysis provided a much more detailed picture of the life patterns while multistate models restricted to the first events only. Nevertheless, for the analytic and 'causal-like' questions posed in the study, sequence analysis turned out to be less helpful.

In the second case study, multivariate psychological measures before and after the pathway analysis were combined into a larger structural equation model. Embedding the clustering results from sequence analysis in it allowed for including the multidimensional information about the trajectories in a way that could hardly be achieved with a few covariates. However, this information is often weak because individuals may in fact have characteristics of several overlapping clusters. Since clustering is based on the matrix of pairwise distances, and not on the individual sequences any more, explanatory models based on membership indicators can be rather unspecific. The cluster characteristics are not then representative to all its members, and sensitivity analysis with, for example, MDS plots can be useful. Lundevaller et al. (2018) used the combined SA states directly as covariates in Cox regression models but this approach would require a larger dataset than they had. Rossignon et al. (2018) propose to add the individual trajectory as a time-dependent covariate which resembles our logistic risk models with time-dependent indicators for the events of the multistate model in Case 1.

The third case study uses sequence analysis to find the most plausible cases for the follow-up from a large register data while leaving the rest of the cohort as a reference, if needed. Initial clustering with SA allows again multidimensional and time-dependent criteria to extract out a subgroup for further analysis. In Helske et al. (2018) clustering with SA was used to get initial values for the latent states of a hidden Markov model.

Preserving diversity in life-histories in the preliminary stage usually means that we need dimension reduction in later stage. In this paper, we have used latent variable (hierarchical) modelling in various ways for this purpose. Latent variables may have different interpretations: individual-specific tendency to respond (latent response models), deviation from group-specific mean behaviour (mixed models), frailty (excess risk for an event in survival models), or an underlying hypothetical construct or trait which can be observed as a pattern of multiple items (latent trait models, IRT models etc.). In latent transition models or hidden Markov models the latent structure is dynamic and, especially for multichannel problems, the interpretation of the latent states becomes sometimes quite difficult.

Sequence analysis is undoubtedly most effective in grasping the ‘big picture’ of the state dynamics in population-level studies. It provides an easily understandable visual representation (proportions of states by time) of multidimensional longitudinal data with minimal simplification of the original data. The figures lead to questions as to why these observed differences between population groups exist. This often requires individual-level information, which unfortunately, apart from the membership indicator, is lost in clustering. More specific causal inquiries, such as “How would the trajectory of an individual of certain type most likely be, had he/she faced (or avoided) a particular life event, which he/she didn’t, given that

everything else had been the same?" are only possible in probabilistic modelling. In a more general sense, however, combining SA with statistical modelling allows quantitative comparative analysis of observed differences in terms of explanatory covariates, and evaluation of their significance.

Acknowledgements The HELS study led by Katariina Salmela-Aro and Jari-Erik Nurmi and the JYLS study initiated by Lea Pulkkinen and led by Katja Kokko are acknowledged for the permission to use the data in the Case studies. The referees are acknowledged for their valuable comments.

References

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research, 11*(2), 91.

Arjas, E. (1989). Survival models and martingale dynamics. *Scandinavian Journal of Statistics*, 177–225.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Probability and statistics). New York: Wiley.

Eerola, M. (1994). *Probabilistic causality in longitudinal studies* (Lecture notes in statistics, Vol. 92). New York: Springer.

Eerola, M., & Helske, S. (2016). Statistical analysis of life history calendar data. *Statistical Methods in Medical Research, 25*(2), 571–597.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Hoboken: Wiley.

Helske, S., Helske, J., & Eerola, M. (2018). Analysing complex life sequence data with hidden Markov modelling. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lundevaller, E., Vikström, L., & Haage, H. (2018). Modelling mortality using life trajectories of disabled and non-disabled individuals in 19th-century Sweden. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 376–398.

Nurmi, J.-E., Salmela-Aro, K., & Haavisto, T. (1995). The strategy and attribution questionnaire: Psychometric properties. *European Journal of Psychological Assessment, 11*, 108–121.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in medicine, 26*(11), 2389–2430.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Rossignon, F., Studer, M., Gauthier, J.-A., & Goff, J.-M. L. (2018). Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

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

Salmela-Aro, K., Kiuru, N., Nurmi, J.-E., & Eerola, M. (2014). Antecedents and consequences of transitional pathways to adulthood among university students: 18-year longitudinal study. *Journal of Adult Development, 21*(1), 48–58.