Supplementary material of *Biclustering based longevity profiles characterization from a longitudinal database*

SM.1. Data Preprocessing Process

The preprocessing steps are:

(*Step 1*) Selection of records labeled as long-lived and non-long-lived, considering the answers provided from the first record of individuals who became long-lived and the last record of responses from non-long-lived individuals;

(*Step 2*) Filtering process to remove inconsistencies, such as a) individuals under the age of 50 or born outside the United Kingdom; b) responses considered inconsistent by the interviewer); c) individuals with deaths caused by accidents; d) records where the missing data rate is greater than 20%; e) information related to metadata (i.e., information that is not related to the respondent, only to the questionnaire itself).

(*Step 3*) Missing data estimation through the Last-Observation-Carried-Forward [38] method, where we replace a missing value with a previous wave of the same feature for the same record.

(*Step 4*) Conceptual feature selection of variables closely related to human aging, based on expert knowledge gathered in a previous literature review [3].

(*Step 5*) Merging of related attributes to reduce dimensionality, e.g., we merge questions about the participant engaging in an activity and the frequency of that activity into a single attribute that adequately represents the information of both questions.

(*Step 6*) Coding of the attribute values to keep the ordered nominal feature values in a [0..1] interval, where '0' always represents the least favorable out-

come for reaching longevity, and '1' represents the most favorable.
(*Step 7*) Fusion of feature sets of questions associated with the same aspect. Each feature set was merged into a single feature, having its value calculated by weighting the involved features. More about the preprocessing steps are in [10].

SM.2. Feature Merging

The process to calculate values for the merged variable is given in Equation SM.2.1 [10].

$$K = \sum_{i=1}^{I} V(X_i) * W(X_i) \tag{SM.2.1}$$

Where X_i is a feature belonging to the merged feature set K, which is composed of I features, $V(X_i)$ corresponds to the value of X_i for a given record, and $W(X_i)$ is its weight in the feature set. The resulting value for the merged feature set K is in the [0.0, N] interval, where $N = \sum_{i=1}^{I} W(X_i)$. In our values, '0' always indicates the least favorable for reaching longevity, and 'N' is the most favorable.

Table SM.2.1 illustrates with an example how to calculate the value assumed by a record in the merged feature set related to the alcohol and tobacco consumption habits reported by the respondents.

TABLE SM.2.1. Example calculation of merged feature set value - Tobacco and Alcohol consumption (set of variables A8).

ELSA feature	**Description**	**Answer Options**	**Value**	**Weight**
heska_smk	Tobacco	Never smoked	1.0	3
		Former smoker	0.5	
		Smoker	0.0	
scako	Alcohol	None in last 12 months	1.0	4
		Once every two months	0.8	
		Once or twice a month	0.6	
		Once or twice a week	0.4	
		Three to four times a week	0.2	
		Almost every day	0.0	
merged set A8 = $V(heska\_smk) * 3 + V(scako) * 4 = N \in [0.0, 7.0]$				

The weighting strategy allows you to calculate the value that represents the information for each ELSA questionnaire response in the set. In these merged feature sets, the higher the value answered by a respondent, the greater the chance that the respondent will achieve longevity. For instance, Table SM.2.2 shows the possible values assumed by the merged feature set A8. Regarding Tobacco Consumption (0: Smokers) and Alcohol Consumption (0: Almost every day), the lowest value of 0.0 contributes to a non-long-lived classification. On the other hand, high values for Tobacco Consumption (1: Never smoked) and Alcohol Consumption (1: Nothing in the last 12 months) results in a feature set value of 7.0, which contributes to a classification as long-lived.

TABLE SM.2.2. Possible values for the merged feature set A8 (Tobacco and Alcohol consumption)

heska_smk	scako	Total
0.0	**0.0**	**0.0**
0.0	0.2	0.8
0.0	0.4	1.6
0.0	0.6	2.4
0.0	0.8	3.2
0.0	**1.0**	**4.0**
0.5	0.0	1.5
0.5	0.2	2.3
0.5	0.4	3.1
0.5	0.6	**3.9**
0.5	0.8	**4.7**
0.5	**1.0**	**5.5**
1.0	0.0	3.0
1.0	0.2	**3.8**
1.0	0.4	**4.6**
1.0	0.6	**5.4**
1.0	0.8	**6.2**
1.0	**1.0**	**7.0**

SM.3. Tables from factor analysis in Section 3.2.1

As an example of variables highly correlated, independent of their representation by the 5, 9, and 11 factors dataset, Table SM.3.1 shows that the questionnaire variables related to depression, addressing negative feelings (set D2) and the feature related to difficulties reading the question cards ('fqhelp') maintained the correlation between themselves. Therefore they are considered to belong to the same set. Following this analysis, as shown in Table SM.3.1, we defined eight merged feature sets. This Table also presents a 9th block labeled as 'Undefined' containing features that did not remain correlated in the three possible transformations.

TABLE SM.3.1. Variable Loadings for the three possible factor analysis sets

Feature	5 Factors		9 Factors		11 Factors		Feature Set
hebck	0.85	F3	0.82	F7	0.82	F8	
hefet	0.81	F3	0.82	F7	0.82	F8	
hehip	0.85	F3	0.85	F7	0.85	F8	
hekne	0.84	F3	0.87	F7	0.90	F8	D1
heji	0.32	F3	0.36	F7	0.33	F8	
helim	0.56	F3	0.55	F7	0.52	F8	
mmschs	0.41	F3	0.38	F7	0.39	F8	
psceda	0.86	F4	0.89	F4	0.87	F5	
pscedb	0.75	F4	0.74	F4	0.70	F5	
pscedc	0.53	F4	0.57	F4	0.62	F5	
pscede	0.72	F4	0.70	F4	0.68	F5	D2
pscedg	0.81	F4	0.82	F4	0.79	F5	
pscedh	0.71	F4	0.69	F4	0.68	F5	
fqhelp	-0.68	F4	-0.58	F4	-0.78	F5	
scfam	-	F4	0.99	F2	0.98	F2	D3
scfrd	0.43	F4	0.70	F2	0.78	F2	
pscedd	0.56	F4	0.66	F4	0.69	F9	D4
pscedf	0.58	F4	0.64	F4	0.57	F9	
hepbs	0.94	F2	0.96	F1	0.96	F1	D5
hespk	0.69	F2	0.74	F1	0.67	F1	
wplljy_rage	0.98	F1	0.98	F3	0.97	F3	
wpdes	0.67	F1	0.67	F3	0.70	F3	D6
heaga_h	0.55	F1	0.54	F3	0.56	F3	
hemdb	-	F5	-	F5	0.52	F11	D7
wpphi	-	F5	-	F5	0.57	F11	
hefrac	0.52	F5	0.56	F5	0.63	F6	
mmhss	0.34	F5	0.41	F5	0.42	F6	
hevsi	0.67	F5	0.66	F5	0.71	F6	
heins	0.45	F5	0.45	F5	0.41	F6	D8
wpedc	-0.31	F5	-0.43	F5	-0.39	F6	
iadebt_owe	0.75	F5	0.62	F5	0.67	F6	
wpbus	-0.51	F5	-0.49	F5	-0.54	F6	
dhsex	-	F5	-	F4	-0.47	F4	
heyra	0.32	F3	0.36	F6	0.50	F7	
hefla	-	F3	-	F4	-	F8	
hewks	0.97	F2	0.98	F5	0.97	F1	
hecat	-	F5	-	F9	0.36	F11	
hecanb	0.31	F5	-0.41	F6	0.31	F6	
heama	-	F4	-	F9	-	F9	

Table SM.3.1 – *continue*

Table SM.3.1 – continuation

Feature	5 Factors		9 Factors		11 Factors		Feature Set
hemda	-	F2	-	F1	-	F3	D9
helng	-	F3	-	F3	-	F11	Undefined
hehra	-	F2	-	F8	0.33	F7	
mmaid	0.37	F3	0.33	F7	0.34	F6	
chouthh	-	F2	-	F9	0.34	F4	
scptr	0.35	F4	0.30	F2	0.92	F4	
hoveh_spcar	-	F4	-	F5	0.44	F11	
hoevm_hopay	0.54	F5	0.38	F8	0.46	F6	
iacisa_npb_sava	-	F5	-0.32	F6	-0.34	F7	

Factors with loading values lower than 0.30 were not considered significant [36, 37]. We indicate these cases by a dash (-) in the Table. Note that are no features correlated in the 10th factor (for 11 factors analysis). For this reason, this factor does not appear in Table SM.3.1. Data dictionary are available in ELSA. Some variables are coded in this work so that lower values are related to non-long-lived classification and higher values to long-lived.

We describe the dichotomous features belonging to each merged feature set, shown in Table SM.3.1, below:

1. D1 (Musculoskeletal problems) - Features related to pain in the back (*hebck*), feet (*hefet*), hip (*hehip*) or knee (*hekne*), joint replacement (*heji*), limitations due to health problems (*helim*) and difficulties with mobility (*mmschs*);

2. D2 (Negative feelings) - Features related to depression and the occurrence of negative feelings during the previous week: felt depressed (*psceda*), felt that everything required effort (*pscedb*), felt that sleep was restless (*pscedc*), felt lonely (*pscede*), felt sad (*pscedg*), felt they could not go on (*pscedh*) and problems reading the question cards (*fqhelp*).

3. D3 (Social Relationships) - Features related to the presence of an immediate family member (*scfam*) and friends (*scfrd*).

4. D4 (Positive feelings) - Features related to feeling happy (*pscedd*) and enjoying life (*pscedf*) in the previous week.

5. D5 (Stroke-related problems) - Features that describe limitations caused by health problems associated with stroke with sequels (*hepbs*) or difficulty speaking/swallowing (*hespk*).

6. D6 (Work activities) - Features related to retirement (*wplljy_rage*), work (*wpdes*) and frequency of physical activities related to diagnosis of serious health problems (*heaga_h*).

7. D7 (Health plan and comorbidities) - Features related to diabetes medication (*hemdb*) and private health insurance (*wpphi*).

8. D8 (Mobility and economic situation) - Features related to hip fracture (*hefrac*), reduced mobility caused by health problems (*mmhss*), difficulty with vision caused by stroke (*hevsi*), insulin self-injection for diabetes (*heins*), whether the respondent had formal education in the previous year (*wpedc*) and whether the respondent have their own business (*wpbus*) or debt (*iadebt_owe*).

9. D9 (Undefined) - Features not correlated to any other in all three datasets. They describe the gender (*dhsex*); whether the respondent had angina or chest pain in the last two years (*heyra*); fall (*hefla*) or weakness in the limbs (*hewks*); cataract surgery (*hecat*); cancer (*hecanb*); medication for asthma (*heama*), hypertension (*hemda*) or

lung problems (*helng*); hearing difficulties (*hehra*); walking aid (*mmaid*); children living outside the household (*chouthh*) and whether they have a partner (*scptr*); vehicle access (*hoveh_spcar*); home ownership (*hoevm_hopay*) and whether they have savings (*iacisa_npb_sava*).

It can be seen in Table SM.3.1 that some factors considered less relevant (with loadings < 0.30, indicated with -) become relevant when represented by a dataset with more of them. For instance, block D7 features are not considered significant by the representations of 5 and 9-factors, but the features become relevant in the 11-factors. It suggests that the number of factors affects the dimensions refinement quality, giving a higher or lower significance level for correlated features.

SM.3.1. Assessing correlations across factors through published research

To analyze and substantiate the influence of features in sets D1 to D8 of Table SM.3.1, we searched for scientific contributions that can justify the observed correlations because these features are correlated with each other regardless of the number of factors representing the dataset.

- Regarding the features in set D1 (musculoskeletal problems), four of them are strongly correlated (loading$>$0.70): the presence of back pain (*hebck*), foot pain (*hefet*), hip pain (*hehip*) and pain in the knees (*hekne*). According to [39], a study carried out with 258 elderly reported musculoskeletal pain in 82.1% of the interviewees. Musculoskeletal problems derive from the loss of muscle mass and strength that tend to affect movement and aerobic capacity, leading to disability in older people. The lowest correlation, but not less important, exists between

features related to joint replacement surgery (*heji*), physical limitations resulting from health problems (*helim*), and mobility difficulties (*mmschs*). The gradual increase in pain can lead to surgical procedures for joint replacement [40]. These features correlate with musculoskeletal problems.

- A common characteristic seen in people having depression symptoms is difficulty reading which is associated with low self-esteem, feelings of incompetence that lead to inability to concentrate, and self-isolation [41]. So regarding features in set D2 (Negative psychological problems), we can suggest that depression would be a possible cause for the reading difficulties of the question card observed by the interviewer.

- Albeit the features in set D5 (physical limitations derived from the occurrence of stroke (*hepbs* and *hespk*)) have a significant correlation on three analyses (5, 9, and 11 factors), we also note a high correlation on the feature *hewks* (set D9), corresponding to the presence of weakness in the limbs, and it has high correlation to features in the set D5 on the analyses of 5 and 11 factors. We support the high correlation between these features by studies presented in [42] and [43]. In the first one, 80% of people affected by stroke report weakness in the limbs or inability to perform movements on one side of the body. [43] related that 65% of people with a stroke are affected by dysphagia, characterized by swallowing difficulties.

- Feature set D6 is formed by the features related to being retired (*wplljy_rage*), currently working (*wpdes*) and medical diagnosis for serious health

problems (*heaga_h*) . In a study conducted in 2014, considering 40,000 families in the UK, 70% of people over 60 years old, 18% over 70 years old and 7% over 80 years old were still active in the workforce [44].

- None feature in set D7 has a significant value in the analyses of datasets reduced to 5 and 9 factors. However, the attributes related to the use of diabetes medication (*hemdb*) and possession of a private health plan (*wpphi*) had a relatively moderate correlation in the representation by 11 factors. We found no studies that explain or corroborate this correlation.

- Finally, feature set D8 has a diversity of information as it covers dimensions related to physical health, recent formal education, and economic situation. In this set, the features related to hip fracture (*hefrac*), to vision problems caused by stroke (*hevsi*), and the current existence of financial debts (*iadebt_owe*) are the variables with the highest loading weight (above 0.6) considering the representation by 11 factors, so these factors partially describe the respondent's physical health and economic situation. However, there are no studies that directly relate to the features of this set.

- Regarding feature sets D3 (Social relationship), D4 (Positive feelings), D7 (Health plan and comorbidities), and D8 (Mobility and economic situation), we were unable to find an article to substantiate the correlations found. Therefore, we suggest carrying out studies to understand the causes and impacts of these variables on elderly lives.

SM.3.2. Explanation about the biggest groups of Section 3.2.2

To illustrate this procedure, see Figure SM.3.1. Suppose that X, Y and Z are factors. The white spaces are factor scores settled as zero in BiMax dichotomization and the gray ones as 1. One can observe that the subsets of records *b,c,d* and *e* are biclustered in X; *c* and *d* in Y and X simultaneously; and *c* and *e* in Z and X simultaneously. One can also observe that *c* is biclustered on factors X, Y and Z. So one can say that X is the biggest bicluster containing four records, while Y and Z contain only two records. Therefore, when we chose to analyze only biclusters with one factor, we are interested in knowing which factors can characterize records as long-lived or not long-lived despite these records having inter-class similarities when analyzing all subgroups of biclustering.

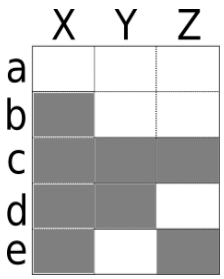

Figure SM.3.1. Example of assignment of records in the biclusters

SM.4. Tables of the Section 4

Training datasets results:

TABLE SM.4.1. Random Forest classification of the training dataset

Dataset	Class	FP(rate)	Precision	Recall	F-Measure
	L	42.5	81.9	96.3	88.5
d1	**NL**	3.8	88.5	57.5	69.7
	W.A.	29.6	84.1	83.3	82.2
	L	30.9	86.1	96.9	**91.2**
d2	**NL**	3.1	91.8	69.1	**78.9**
	W.A.	21.5	88.0	87.6	**87.0**
	L	27.5	87.0	91.9	89.4
d3	**NL**	8.1	81.7	72.5	76.8
	W.A.	21.0	85.2	85.4	85.2
	L	35.0	84.2	93.1	88.4
d4	**NL**	6.9	82.5	65.0	72.7
	W.A.	25.6	83.6	83.8	83.2
	L	27.5	87.1	92.5	89.7
d5	**NL**	7.5	82.9	72.5	77.3
	W.A.	20.8	85.7	85.8	85.6

L: Long-lived. NL: Non-long-lived. WA: Weighted Average. We highlight in bold the best results by class or by the weighted average of the algorithm.

TABLE SM.4.2. J48 classification of the training dataset

Dataset	Class	FP(rate)	Precision	Recall	F-Measure
	L	28.8	86.6	93.1	**89.8**
d1	**NL**	6.9	83.8	71.3	77.0
	W.A.	21.5	85.7	85.8	85.5
	L	27.2	86.4	87.5	87.0
d2	**NL**	12.5	74.7	72.8	73.7
	W.A.	22.2	82.5	82.6	82.5
	L	31.3	85.1	89.4	87.2
d3	**NL**	10.6	76.4	68.8	72.4
	W.A.	24.4	82.2	82.5	82.3
	L	28.8	86.0	88.1	87.0
d4	**NL**	11.9	75.0	71.3	73.1
	W.A.	23.1	82.3	82.5	82.4
	L	23.8	88.4	90.6	89.5
d5	**NL**	9.4	80.3	76.3	**78.2**
	W.A.	19.0	85.7	85.8	**85.7**

L: Long-lived. NL: Non-long-lived. WA: Weighted Average. We highlight in bold the best results by class or by the weighted average of the algorithm.

J48 training results from datasets *d1* and *d5* are similar. But *d5* has a better outcome for random forest training. For both datasets, the false-positive rate associated with F-Score is smaller in dataset *d5*. But *d2* presents better results of the F-score associated with the average of false positives from random forest training. Figure SM.4.1 summarizes these results from Tables SM.4.1 and SM.4.2.

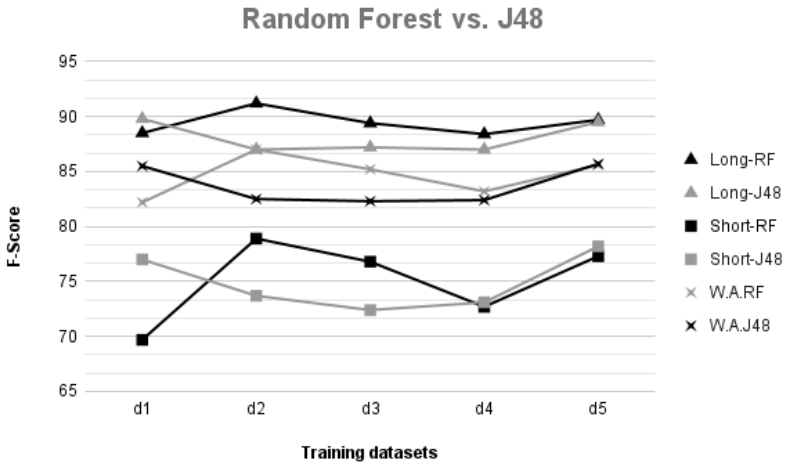

Figure SM.4.1. F-Score classification results by training dataset

TABLE SM.4.3. The merged feature sets composed of categorical variables added to the dataset

Sets	Description	Sets	Description
A1	Frequency of Physical Activity	B8	Positive Feelings
A7	Sensory Efficiency	B9	Life Satisfaction and Outlook
A8	Tobacco and Alcohol Consumption	C3	Relationship with their Children
A9	Memory Test Results	C4	Relationship with Family
B5	Anxiety and Stress	C5	Relationship with Friends
B6	Self-reported Wellbeing	C6	Relationship with Partner
B7	Negative Feelings		

TABLE SM.4.4. The original (un-merged) categorical variables added to the dataset.

Categorical variables	**Description**
hepaa	Whether feels pain, and how severe
hefunc	Difficulty in walking 400m without aid
hhtot	How many people in the household
disib	How many living siblings
dimar	Marital state
dignmy	How many grandchildren
wpvw	Frequency of voluntary work
hotenu	Whether owns the household, pays rent or mortgage
exrslf	Expectation of financial hardship in the future

TABLE SM.4.5. Aspects addressed by the dichotomous features.

Age-related aspects		
Gender	Recent formal education	Social Structure
Career	House and vehicles	Current economic status
Serious health problems	Medical history	Pain severity
Depression symptoms	Cognitive issues	Medication use
Physical limitations	Mobility test results	

TABLE SM.4.6. Volunteer work frequency variable

Feature	**Answer Options**	**Value**
wpvw	never	1.0
	less than once a year	0.8
	about once or twice a year	0.6
	every few months	0.4
	about once a month	0.2
	twice a month or more	0.0

TABLE SM.4.7. Best parameters for Random Forest algorithm by dataset

Dataset	**Parameters**
d1	-P 100 -attribute-importance -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
d2	-P 100 -I 110 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 -depth 3
d3	-P 100 -attribute-importance -I 110 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 20 -depth 2
d4	-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 -depth 3
d5	-P 90 -print -attribute-importance -I 280 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 -depth 3 -B

TABLE SM.4.8. Best parameters for J48 algorithm by dataset

Dataset	Parameters
d1	-R -N 8 -Q 1 -M 1
d2	-R -N 4 -Q 1 -M 1
d3	-C 0.05 -M 1
d4	-C 0.05 -M 1
d5	-R -N 8 -Q 1 -M 1