Identification of High Likelihood of Dementia in Population-Based Surveys using Unsupervised Clustering: a Longitudinal Analysis

Authors

Amin Gharbi-Meliani, François Husson, Henri Vandendriessche, Eleonore Bayen, Kristine Yaffe, Anne-Catherine Bachoud-Lévi, Laurent Cleret de Langavant

Corresponding author:

Laurent Cleret de Langavant

Service de Neurologie

Hôpital Henri Mondor

Assistance Publique Hôpitaux de Paris

51 Avenue du Maréchal de Lattre de Tassigny, 94000 Créteil

3ème étage

laurent.cleret@gbhi.org

Affiliations :

(1) *Equipe neuropsychologie interventionnelle, Institut Mondor de Recherche Biomédicale, Département d'études cognitives, Ecole normale supérieure, Université PSL, Université Paris-Est Créteil, AP-HP Hôpital Henri Mondor-Albert Chenevier, Centre de référence Maladie de Huntington et Service de Neurologie, INSERM, 75005 Paris [ou 94000 Créteil]*, France
(Amin Gharbi-Meliani MPH, Anne-Catherine Bachoud-Lévi MD, Laurent Cleret de Langavant MD)

(2) *Institut Agro, Univ Rennes1, CNRS, IRMAR, 35000, Rennes, France*
(François Husson PhD)

(3) *Laboratoire de Neurosciences Cognitives et Computationnelles, Département d'études cognitives, Ecole normale supérieure, Université PSL*, INSERM, 75005 Paris, France
(Henri Vandendriessche MS)

(4) Global Brain Health Institute, University of California, San Francisco, CA, United States

(Eleonore Bayen MD, Kristine Yaffe MD, Laurent Cleret de Langavant MD)

(5) *Sorbonne Université, Hôpital Pitié-Salpêtrière–Assistance Publique Hôpitaux de Paris, Département de Rééducation Neurologique*, Paris, France

(Eleonore Bayen MD)

(6) Departments of Psychiatry, Neurology and Epidemiology and Biostatistics, University of California, San Francisco

(Kristine Yaffe MD)

SUMMARY

Background Dementia is defined by cognitive decline that affects functional status. Longitudinal ageing surveys often lack a clinical diagnosis of dementia though measure cognitive and function over time. We used unsupervised machine learning and longitudinal data to identify transition to probable dementia.

Methods Multiple Factor Analysis was applied to longitudinal function and cognitive data of 15,278 baseline participants (aged 50 years and more) from the Survey of Health, Ageing, and Retirement in Europe (SHARE) (waves 1, 2 and 4–7, between 2004 and 2017). Hierarchical Clustering on Principal Components discriminated three clusters at each wave. We estimated probable or "Likely Dementia" prevalence by sex and age, and assessed whether dementia risk factors increased the risk of being assigned probable dementia status using multistate models. Next, we compared the "Likely Dementia" cluster with self-reported dementia status and replicated our findings in the English Longitudinal Study of Ageing (ELSA) cohort (waves 1–9, between 2002 and 2019, 7,840 participants at baseline).

Findings Our algorithm identified a higher number of probable dementia cases compared with self-reported cases and showed good discriminative power across all waves (AUC ranged from 0.754 [0.722–0.787] to 0.830 [0.800–0.861]). "Likely Dementia" status was more prevalent in older people, displayed a 2:1 female/male ratio and was associated with nine factors that increased risk of transition to dementia: low education, hearing loss, hypertension, drinking, smoking, depression, social isolation, physical inactivity, diabetes, and obesity. Results were replicated in ELSA cohort with good accuracy.

Interpretation Machine learning clustering can be used to study dementia determinants and outcomes in longitudinal population ageing surveys in which dementia clinical diagnosis is lacking.

Funding French Institute for Public Health Research (IReSP), French National Institute for Health and Medical Research (Inserm), NeurATRIS Grant (ANR-11-INBS-0011), and Front-Cog University Research School (ANR-17-EUR-0017).

INTRODUCTION

Major neurocognitive disorder (MND), commonly known as dementia, is a clinical syndrome characterised by a decline in cognitive performance that compromises patient's independence[1]. Repeated clinical visits and assessments reveal the progression from a healthy state to dementia. International diagnostic criteria are available to identify dementia cases. Yet, more than half of the cases in high income countries (HIC)[2] and up to 90% in low and middle income countries (LMIC)[3] remain undetected. For such, new methods are needed to identify dementia cases and to study dementia determinants at the level of countries or continents.

Several population-based surveys modelled on the United-States Health and Retirement Study (HRS) are conducted in multiple countries to study the impact of the transition between late-life work and retirement in aging people[4]. The "HRS family" studies offer the opportunity to compare ageing outcomes internationally[5]. Yet, in these and in many other surveys, clinical dementia status is either not available or only self-reported by participants or their families, which underestimates the number of cases.

In the absence of clinical diagnosis in population ageing surveys, unsupervised machine learning, generally used to discover clusters or patterns within datasets[6], can identify probable dementia cases. In a previous work, we applied an unsupervised clustering method to cross-sectional data from HRS and Survey of Health, Ageing and Retirement in Europe (SHARE) to identify high likelihood of dementia[7] based on variables related to demographics, comorbidities, functional status, mobility, cognition, and neuropsychiatric symptoms. However, applying this clustering method to cross-sectional data did not allow us to investigate longitudinal transition from normal to impaired functional status, or to assess risk factors associated with transition to dementia status.

Herein, we built a clustering analysis for identifying transition to high likelihood of dementia in population ageing surveys using repeated measurements of cognition and functional status with a modified unsupervised machine learning algorithm. Our objectives were to demonstrate that this method can identify probable dementia in population aging surveys where dementia is either poorly or non-diagnosed, and that this method is also efficient to study dementia risk factors. Three analyses were used to ascertain the internal validity of "Likely Dementia" status: (1) we compared "Likely Dementia" identification with self-reported dementia, (2) we studied the prevalence of "Likely Dementia" status according to sex and age, (3) we tested whether traditional dementia risk

factors were associated with a higher risk of transition to "Likely Dementia" cluster. To demonstrate replicability, we conducted our study using SHARE survey and replicated it in the English Longitudinal Study of Ageing (ELSA).

METHODS

Study design and participants

We used the harmonised dataset provided by the Gateway to Global Aging[5] of SHARE, a longitudinal panel study across multiple countries in Europe and Israel[8]. This population survey takes place every two years and follows a representative sample of individuals aged 50 years or older from each participating country. The harmonised version of SHARE consists of seven waves so far (the third being retrospective) conducted between 2004 and 2017. We included subjects from countries who have participated in SHARE since the first wave (ie, Austria, Belgium, Denmark, France, Germany, Greece, Israel, Italy, The Netherlands, Spain, Sweden, and Switzerland), aged 50 years or older with consecutive follow-ups.

Selected variables

We used variables related to cognition and function to remain close to the DSM-5 criteria of MND. The selected variables are listed in the supplementary tables (Supplementary Table 1 & 2). All variables with more than 30% missing values were discarded and the remaining data were imputed using the imputeMFA command of the missMDA R package[9].

Clustering

We ran Multiple Factor Analysis (MFA) followed by Hierarchical Clustering on Principal Components (HCPC) using FactoMineR R package[10] and longitudinal data from all waves at the same time. MFA is a principal component method that balances for differences in the number of active variables per domain by forming active groups (procedure details are shown in supplementary data). For the clustering, we retained only active groups that represented participants' function or cognition (see supplementary tables). Each participant, at each wave, was assigned to one of the three possible clusters (ie each participant could transition from one cluster to another, from one wave to another longitudinally). The number of clusters was set at three based on previous work for identification of high likelihood of dementia[7]. At the first wave, we singled out a cluster with a high probability of dementia (named "Likely dementia" cluster) based on the impaired cognition and function detected in its

participants. Any participant classified in "Likely dementia" cluster was permanently assigned to it (ie, making any incident case a prevalent one).

We took into account the attrition due to study dropout and death across waves. We applied Inverse Probability Weighting (IPW) using the ipw R package[11]. For each wave, a logistic regression model was built based on the participants' age, sex, and country of origin characteristics collected at the previous wave. Weights were derived by inverting the product of the predicted probabilities computed by the model. We integrated those weights into both imputation and clustering methods.

Self-reported diagnosis of dementia

We evaluated the discrimination power of our clustering method counting on its identification of "Likely dementia" status compared with the self-reported dementia status, data collected from the second wave of SHARE, using Sensitivity, Specificity and Area Under the Curve (AUC) metrics.

Effect of age, sex, and risk factors for dementia

We computed the prevalence of "Likely dementia" status of each wave by sex and by age. Participants were divided into six age groups (under 65 years, 65–69 years, 70–74 years, 75–79 years, 80–85 years, and more than 85 years).

We examined the role of several modifiable risk factors[12] in transitioning to "Likely dementia" cluster: low education, hearing loss, hypertension, excessive alcohol drinking, current smoking, depression, social isolation, physical inactivity, diabetes, obesity, and air pollution. Past history of traumatic brain injury was not available in the database and could not be tested. All risk factors were measured at baseline and were imputed if necessary.

We dichotomised all ordinal risk factors variables. Education level was categorised as high (upper secondary and vocational training or tertiary education) or low (less than upper secondary). For hearing loss, we used self-reported hearing capacity as a proxy considering it either being normal (excellent, very good, and good) or bad (fair or poor). Moderate and vigorous physical activity were merged into physically active (frequency: more than once per week, once per week, one to three times a month) or inactive (hardly ever or never). The remaining risk factors were treated as dichotomous as they were in the database: hypertension (ever had high blood pressure *vs* never had high blood pressure), drinking (21 units or more of alcohol per week *vs* less than 21 units of alcohol

per week), smoking (current smoker *vs* non-current smoker), depression (Centre for Epidemiologic Studies Depression [CES-D] scale score greater than or equal to five *vs* CES-D scale score less than five), social isolation (participating in social activities weekly *vs* non-participating in social activities weekly), diabetes (ever had diabetes *vs* never had diabetes), obesity (Body Mass Index [BMI] ≥ 30 kg/m2 *vs* BMI < 30 kg/m2), air pollution (living in urban area *vs* living in rural area).

Multistate models

In each wave, a participant could be classified in one of the three clusters (Cluster 1, Cluster 2 or Cluster 3; see above). Data being interval-censored, we applied multistate models using MSM package[13] to study the impact of dementia risk factors on the risk of transition to "Likely Dementia" cluster.

We used age as the time-scale. It was calculated as the difference between birth date and interview date in years. In the multistate models, age was divided by 10 to facilitate the computational process without altering the Hazard Ratios (HR) results. Sex was considered as binary (male or female). All transitions were adjusted for sex, and transition towards "Likely Dementia" cluster was further adjusted for age. All covariates were set at baseline. For each risk factor, we computed its corresponding HR.

We checked the robustness of the multistate models in two steps. First, we considered death as a competing risk and added it as an absorbing state in the models. This was investigated in SHARE where vital status was reported consistently. Second, we excluded early prevalent and incident dementia cases by excluding participants categorised with a likelihood of dementia at first and second waves and ran multistate analyses again.

Replication cohort

In order to confirm our results, we chose the harmonised version of ELSA[14] as a replication cohort. It is a representative longitudinal panel study of people aged 50 years and over in England. It comprises nine waves ranging from 2002 to 2019.

Ethical approval and guidelines

All participants provided informed consent and both studies obtained ethical approvals from local research committees. We followed both STROBE (STrengthening the Reporting of OBservational studies in Epidemiology)

and MELODEM (The MEthods in LOngitudinal research on DEMentia) guidelines[15,16] for the reporting of this study.

Role of the funding source

Sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

RESULTS

Identification of probable dementia

Of the initial sample of SHARE (n=30,419), we restricted our analyses to participants aged 50 years and over at baseline (n=29,102), who had consecutive follow-ups (n=15,278) (Figure 1). After running the clustering, the distribution between the clusters was uneven. At baseline, the first cluster (n=11,369) and the second (n=1,294) encompassed the majority of the sample, leaving a small part for the third cluster (n=535) (Table 1). Participants of the first and second clusters had similar baseline characteristics evoking healthy ageing. Participants of the third cluster were older (mean age 76.5 years [SD 11]), often female (n=368 [68.6%]), had lower education level (n=426 [79.6%] attained less than upper secondary education), more mobility impairment (mean mobility impairment score 4.9 [SD 1.5]), more functional impairment (mean Activities of Daily Living [ADL] score 3.1 [SD 1.7] and mean Instrumental Activities of Daily Living [IADL] score 4.2 [SD 1.8]), and more impaired cognition (mean immediate word recall test 2.6 [SD 1.9] and mean verbal fluency 10.4 [SD 6]) than participants of the first and second clusters at baseline. These characteristics corroborated that the third cluster was the one reflecting a high likelihood of dementia, thus we named it “Likely Dementia” cluster. Conversely, the first and second clusters were composed of participants considered dementia-free.

Discrimination power

We compared our algorithm identification with the self-reported dementia diagnosis in the SHARE dataset, which was available from wave 2 (Table 2). Our clustering algorithm allowed the identification of a higher number of “Likely Dementia” cases compared with self-reported dementia cases. The AUC metric ranged from 0.754 (0.722–0.787) to 0.830 (0.800–0.861), suggesting good discrimination power. Sensitivity peaked at wave 4 reaching

0.714 (0.659–0.770) then slowly decreased after. Specificity remained high (> 0.9) in all waves. Results by country are given in supplementary data (Supplementary Table 1).

Effect of age and sex

Older age and being female were both associated with an increased risk of entering “Likely Dementia” cluster. The prevalence of “Likely Dementia” was higher in women with approximatively a 2:1 female to male ratio across all waves (Fig 2.A). The number of “Likely Dementia” cases increased with age (Fig 2.B). For instance, at wave 2, the prevalence of “Likely Dementia” cases gradually rose with age: 1.8% in those under 65 years, 3.1% in 65–69 years, 5.9% in 70–74 years, 10.2% in 75–79 years, 18.9% in 80–85 years, and 37.4% in more than 85 years old participants.

Multistate models

To assess the associations of dementia risk factors with the risk of transitioning to ‘’Likely Dementia’’ cluster (Table 3), we computed a multistate model (Figure 3.A). Nine of the eleven dementia risk factors chosen a priori were associated with an increased risk of transition from cluster 1 to “Likely Dementia” cluster: low education level (Hazard Ratio [HR] 1. 92% CI [1.58−2.33]), poor hearing (1.74 [1.45−2.09]), hypertension (1.35 [1.14−1.16]), smoking (1.45 [1.13−1.87]), depression (2.51 [1.06−3.07]), social isolation (1.66 [1.39−1.98]), physical inactivity (3.66 [2.97−4.51]), diabetes (2.4 [1.94−2.96]), and obesity (1.7 [1.39−2.07]). Some of these associations were also significant for transition from cluster 2 to “Likely Dementia” cluster: depression (2.39 [1.62−3.53]), social isolation (2.31 [1.51−3.53]), physical inactivity (3.21 [2.12−4.87]), and obesity (1.58 [1.08−2.32]).

In the first sensitivity analysis which took into account death (Figure 3.B), we excluded 105 participants due to inconsistencies between interview and death dates. All of the above-described associations between dementia risk factors and transition to “Likely dementia” cluster remained significant albeit with lower HR, except for hypertension. Of more, smoking became significantly associated with the risk of transition from cluster 2 to “Likely Dementia” cluster (2.23 [1.57−3.16]). In the second sensitivity analysis, where prevalent and incident cases at wave 1 (2004−05) and wave 2 (2006−07) (n=1,000) were further removed, HRs of transition from cluster 1 to ‘’Likely Dementia’’ cluster did not change, but excessive alcohol drinking became a significant risk factor (1.34 [1.17−1.53]). As for transitions from cluster 2 to ‘’Likely Dementia’’ cluster, only smoking (2.68 [1.79−4.03]) and depression (1.65 [1.07−2.54]) remained significant.

Replication in ELSA

Of the initial sample of ELSA (n=12,099), we restricted our analyses to participants over 50 years at baseline (n=11,522) and further restricted to participants who had consecutive follow-ups (n=7,840) (Supplementary Figure 2). Overall, results obtained with ELSA participants were similar to those found in the SHARE cohort (Supplementary Table 4).

At baseline, participants of the third cluster (n=659) were more likely older (mean age 69.8 [SD 11]), more likely female (n=401 [60.8%]), of lower education level (n=423 [64.2%] attained less than upper secondary education), had more mobility impairments (mean mobility impairment score 4.9 [SD 1.5]), more functional impairment (mean ADL score 2.7 [SD 1.5] and mean IADL score 2.5 [SD 1.4]) and worse cognition (mean immediate word recall test 4.6 [SD 1.9] and mean verbal fluency 16.5 [SD 6]) than the other clusters.

Our clustering algorithm identified a higher number of “Likely Dementia” cases compared with self-reported dementia cases. Except for wave 1 (2002−03) in which the number of self-reported dementia cases was small (n=24), the algorithm identification AUC metric values were similar to those found with SHARE (Supplementary Table 5). Sensitivity and specificity were balanced.

Women were more likely to be in the “Likely Dementia” group, and prevalence of “Likely Dementia” status rose with age (Supplementary Figure 3).

Ten dementia risk factors were tested (not air pollution due to missing urbanicity data). Their associations with transition to “Likely Dementia” cluster remained similar to those found with the SHARE dataset (Supplementary Table 6) except for excessive alcohol drinking which was protective for the transition from cluster 1 to “Likely Dementia” cluster (HR 0.6 [0.43−0.83]). Four risk factors were associated with an increased risk of transition from cluster 2 to “Likely Dementia” cluster: hypertension (1.64 [1.13−2.38]), depression (2 [1.26−3.17]), physical inactivity (2.69 [1.73−4.18]), and diabetes (2.23 [1.26−3.95]). We did not take death into account in the multistate models as vital status data were not available for each wave in sensitivity analysis.

Removing prevalent and incident cases at wave 1 (2002−03) and wave 2 (2004−05) in sensitivity analysis led to similar results with few exceptions. Excessive alcohol drinking was no longer significant for the transition from cluster 1 to ‘’Likely Dementia’’ cluster (0.79 [0.58−1.08]). Only physical inactivity remained significant for the risk of transition from cluster 2 to ‘’Likely Dementia’’ cluster (2.02 [1.1−3.69]).

DISCUSSION

Unsupervised clustering applied to two longitudinal population-based surveys of ageing (SHARE and ELSA) identified participants with high likelihood of dementia using longitudinal data related to functional and cognitive measures. In both surveys, this method had a good discrimination performance when compared with self-reported diagnosis of dementia. “Likely Dementia” status was more common in older participants and in women with a 2:1 sex ratio. Low education, hearing loss, hypertension, smoking, depression, social isolation, physical inactivity, diabetes, and obesity were associated with a higher risk of subsequent transition to “Likely dementia” cluster. Results for excessive alcohol drinking and air pollution were inconclusive. Applying clustering to longitudinal cohorts for the identification of high likelihood of dementia paves the way for researchers to conduct future secondary analyses on population ageing surveys worldwide.

Although supervised machine learning algorithms have already been used in population surveys to identify persons with dementia[17], they have their limitations, eg, they require a subsample of data to be labelled “diagnosis of dementia”, and their external validity remains variable. Conversely, unsupervised machine learning may overcome such limitations as suggested in a previous cross-sectional study[7]. Here, we used an improved clustering method combining longitudinal data and a limited number of variables related to participants’ cognition and daily functions. Our clustering algorithm identified a greater number of people with a high likelihood of dementia in both SHARE and ELSA compared with self-reported dementia cases. Identifying a higher number of probable dementia cases in population ageing surveys might give a better statistical power to future studies of dementia determinants and outcomes. Moreover, this clustering method relies on cognitive and functional status data, largely available in HRS family studies and in several population ageing surveys, which makes it very suitable to apply to other ageing surveys including those in LMIC. Noteworthy, our study took into account many biases inherent to longitudinal studies, in particular attrition[18] due to loss to follow-up or death. Internal validity was assessed using different approaches: comparison with self-reported diagnosis of dementia, impact of age and sex on dementia prevalence, and impact of known dementia risk factor on the risk of being classified as a “Likely Dementia” case. Results were obtained using data of 12 countries participating in SHARE, and then replicated in ELSA.

Nonetheless, our results should be carefully examined. Our algorithm detects a “Likely Dementia” status which cannot, by any stretch, be taken as a diagnosis of the disease without clinical validation. Future studies that

compare our identification method with the recently developed cognitive assessment in HRS family cohorts using the Harmonized Cognitive Assessment Protocol (HCAP)[19] are warranted. Our method cannot distinguish the aetiology of dementia, whether Alzheimer's disease (AD) or other origins. Contrary to the results of our prior cross-sectional study, Cluster 1 and Cluster 2 participants were similar in terms of daily function, cognition, and mobility, yet they differed in their risk of transition to Cluster 3. However, we cannot rule out the possibility that the non-significant HRs observed for the transition from Cluster 2 to "Likely Dementia" cluster resulted from a lack of statistical power. Although this three-cluster partition remains consistent with our earlier work[7], future investigation will test the interest of further simplification by merging the first two clusters together. The lack of biological or imaging biomarkers in this study could be seen as a limitation. Yet, biomarkers are often costly, human expert-dependent and rarely available in large population ageing studies. As for genetics, Apolipoprotein E (*APOE*)[20] and polygenic scores[21] are associated with a higher risk of AD, but the role of genetic factors in explaining future risk of dementia remains modest[21,22]. Although most dementia risk factors were associated with a higher risk of being assigned a "Likely Dementia" status, results for excessive alcohol drinking were ambiguous. We observed a deleterious drinking effect in SHARE, whereas it was protective in ELSA. Excessive drinking has been entangled for the brain damage it causes[23], yet its exact relationship with dementia risk is debated since alcohol thresholds and time of exposure differ between studies[24,25]. The association between air pollution and dementia was inconclusive in SHARE and could not be explored in ELSA. Urbanicity (ie, geographical variation between urban and rural areas) was used as a proxy for air pollution as proposed recently[12]. Yet, people living in rural areas have shown higher rates of dementia compared with their urban counterparts [26,27]. Switching to quantifiable pollution markers (fine particulate matter or ozone) that have been linked to an increased risk of dementia[28] is more than desirable.

Unsupervised clustering is an efficient method to detect people with probable dementia in population ageing surveys using their cognitive and functional characteristics in a longitudinal setting. This approach opens new perspectives for the analyses of population data sets already available worldwide in HIC and LMIC to better compare and understand dementia determinants and outcomes.

Contributors

AGM and LCL contributed to the conceptualisation and design of the study. AGM analysed the data. FH granted statistical methods expertise. HV helped with the computational analyses. AGM and LCL wrote the first draft. All

authors reviewed and approved the final version of the manuscript. AGM decided to submit this manuscript for publication.

Declaration of interests

EB is an advisory board member of SafelyYou. EB has vested stock options in SafelyYou stock. KF is a participant on a data safety monitoring board Eli Lilly and NIH-sponsored trials. KF is a board member of Alector. ACBL has received academic grants from the French National Research Agency (ANR-17-EURE-0017) and NeurATRIS (ANR-11-INBS-0011). ACBL has received fees for consultancy work from Roche. All other authors declare no competing interests.

Data sharing

All the data we used are publicly available. We used the harmonised versions of SHARE and ELSA provided by the Gateway to Global Aging Data (https://g2aging.org/). ELSA raw data can be downloaded via the UK Data Service (https://ukdataservice.ac.uk/find-data/). SHARE raw data can be accessed on their website (http://www.share-project.org/data-access.html).

Acknowledgements

The Gateway to Global Aging Data project is developed by Centre for Economic and Social Research (CESR) at University of Southern California (https://cesr.usc.edu/). The project is funded by National Institute on Aging, National Institutes of Health (R01 AG030153, RC2 AG036619, R03 AG043052, R24 AG048024)

SHARE was funded by the European Commission, through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA 211909, SHARE-LEAP: GA 227822, SHARE M4: GA 261982, DASISH: GA 283646), and Horizon 2020 (SHARE-DEV3: GA 676536, SHARECOHESION: GA 870628, SERISS: GA 654221, SSHOC: GA 823782), and by DG Employment, Social Affairs and Inclusion (VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332 and VS 2020/0313). Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the US National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, and RAG052527A) and from various national funding sources is gratefully acknowledged.

ELSA was funded by the National Institute on Aging (R01AG017644) and by a consortium of UK government departments: Department for Health and Social Care; Department for Transport; Department for Work and Pensions, which is coordinated by the National Institute for Health Research (NIHR, 198-1074). Funding has also been provided by the Economic and Social Research Council (ESRC).

First and foremost, we would like to express our gratitude to Julie Josse for all the methodological recommendations she gave for this project.

We would like to thank both Bioinformatics services of *Institut Mondor de Recherche Biomédicale* (IMRB) and *Ecole Normale Supérieure* (ENS) for their help in employing computer clusters for our analysis.

References

1 American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5, 5th edition. Washington, DC, 2013.

2 Lang L, Clifford A, Wei L, *et al.* Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis. *BMJ Open* 2017; **7**: e011146.

3 Alzheimer's Disease International, McGill University. World Alzheimer's Report 2021. 2021 https://www.alzint.org/resource/world-alzheimer-report-2021/.

4 Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JWR, Weir DR. Cohort Profile: the Health and Retirement Study (HRS). *Int J Epidemiol* 2014; **43**: 576–85.

5 Lee J, Phillips D, Wilkens J, Gateway to Global Aging Data Team. Gateway to Global Aging Data: Resources for Cross-National Comparisons of Family, Social Environment, and Healthy Aging. *J Gerontol B Psychol Sci Soc Sci* 2021; **76**: S5–16.

6 Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. *Annu Rev Public Health* 2020; **41**: 21–36.

7 Langavant LC de, Bayen E, Yaffe K. Unsupervised Machine Learning to Identify High Likelihood of Dementia in Population-Based Surveys: Development and Validation Study. *J Med Internet Res* 2018; **20**: e10493.

8 Börsch-Supan A, Brandt M, Hunkler C, *et al.* Data Resource Profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *Int J Epidemiol* 2013; **42**: 992–1001.

9 Josse J, Husson F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *J Stat Softw* 2016; **70**: 1–31.

10 Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. *J Stat Softw* 2008; **25**: 1–18.

11 Wal WM van der, Geskus RB. ipw: An R Package for Inverse Probability Weighting. *J Stat Softw* 2011; **43**: 1–23.

12 Livingston G, Huntley J, Sommerlad A, *et al.* Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet* 2020; **396**: 413–46.

13 Jackson C. Multi-State Models for Panel Data: The msm Package for R. *J Stat Softw* 2011; **38**: 1–28.

14 Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: the English longitudinal study of ageing. *Int J Epidemiol* 2013; **42**: 1640–8.

15 von Elm E, Altman DG, Egger M, *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet Lond Engl* 2007; **370**: 1453–7.

16 Weuve J, Proust-Lima C, Power MC, *et al.* Guidelines for reporting methodological challenges and evaluating potential bias in dementia research. *Alzheimers Dement J Alzheimers Assoc* 2015; **11**: 1098–109.

17 Hurd MD, Martorell P, Langa KM. Monetary costs of dementia in the United States. *N Engl J Med* 2013; **369**: 489–90.

18 Schmidt SCE, Woll A. Longitudinal drop-out and weighting against its bias. *BMC Med Res Methodol* 2017; **17**: 164.

19 Langa KM, Ryan LH, McCammon RJ, *et al.* The Health and Retirement Study Harmonized Cognitive Assessment Protocol Project: Study Design and Methods. *Neuroepidemiology* 2020; **54**: 64–74.

20 Gharbi-Meliani A, Dugravot A, Sabia S, *et al.* The association of APOE ε4 with cognitive function over the adult life course and incidence of dementia: 20 years follow-up of the Whitehall II study. *Alzheimers Res Ther* 2021; **13**: 5.

21 Bellenguez C, Küçükali F, Jansen IE, *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* 2022; **54**: 412–36.

22 Leonenko G, Baker E, Stevenson-Hoare J, *et al.* Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat Commun* 2021; **12**: 4506.

23 Evangelou E, Suzuki H, Bai W, *et al.* Alcohol consumption in the general population is associated with structural changes in multiple organ systems. *eLife* 2021; **10**: e65325.

24 Sabia S, Fayosse A, Dumurgier J, *et al.* Alcohol consumption and risk of dementia: 23 year follow-up of Whitehall II cohort study. *BMJ* 2018; **362**: k2927.

25 Mewton L, Visontay R, Hoy N, *et al.* The relationship between alcohol use and dementia in adults aged more than 60 years: a combined analysis of prospective, individual-participant data from 15 international studies. *Addict Abingdon Engl* 2022; published online Aug 22. DOI:10.1111/add.16035.

26 Liu C-C, Liu C-H, Sun Y, Lee H-J, Tang L-Y, Chiu M-J. Rural-urban disparities in the prevalence of mild cognitive impairment and dementia in Taiwan: A door-to-door nationwide study. *J Epidemiol* 2021; published online April 10. DOI:10.2188/jea.JE20200602.

27 Rahman M, White EM, Mills C, Thomas KS, Jutkowitz E. Rural-urban differences in diagnostic incidence and prevalence of Alzheimer's disease and related dementias. *Alzheimers Dement* 2021; **17**: 1213–30.

28 Weuve J, Bennett EE, Ranker L, *et al.* Exposure to Air Pollution in Relation to Risk of Dementia and Related Outcomes: An Updated Systematic Review of the Epidemiological Literature. *Environ Health Perspect* 2021; **129**: 96001.

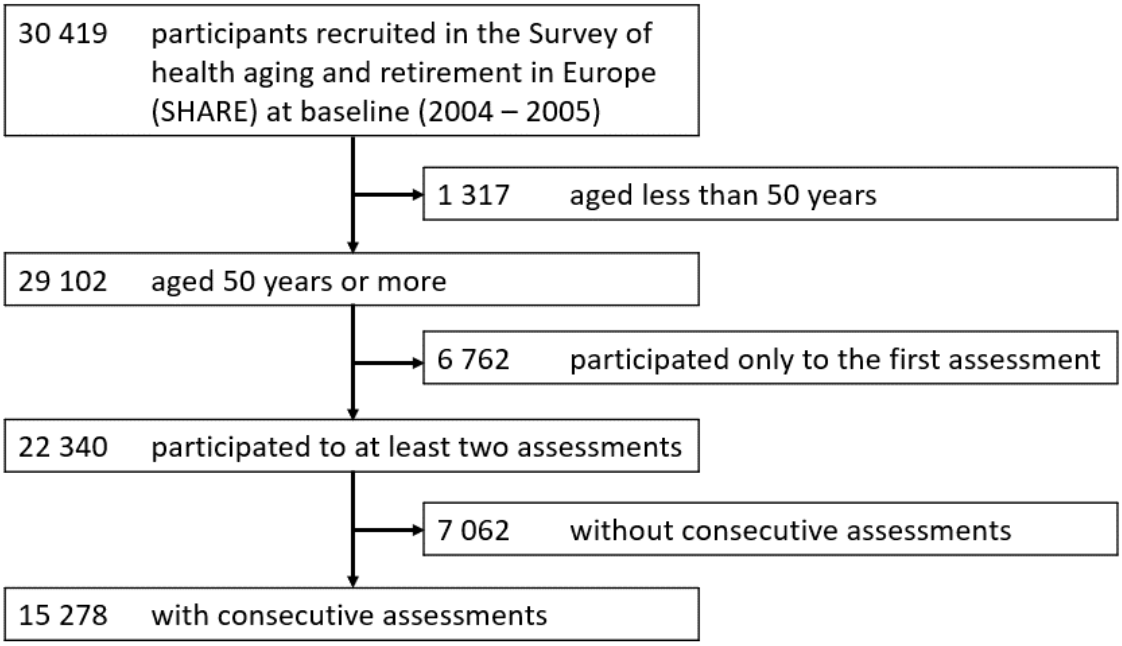


Figure 1: SHARE flow chart

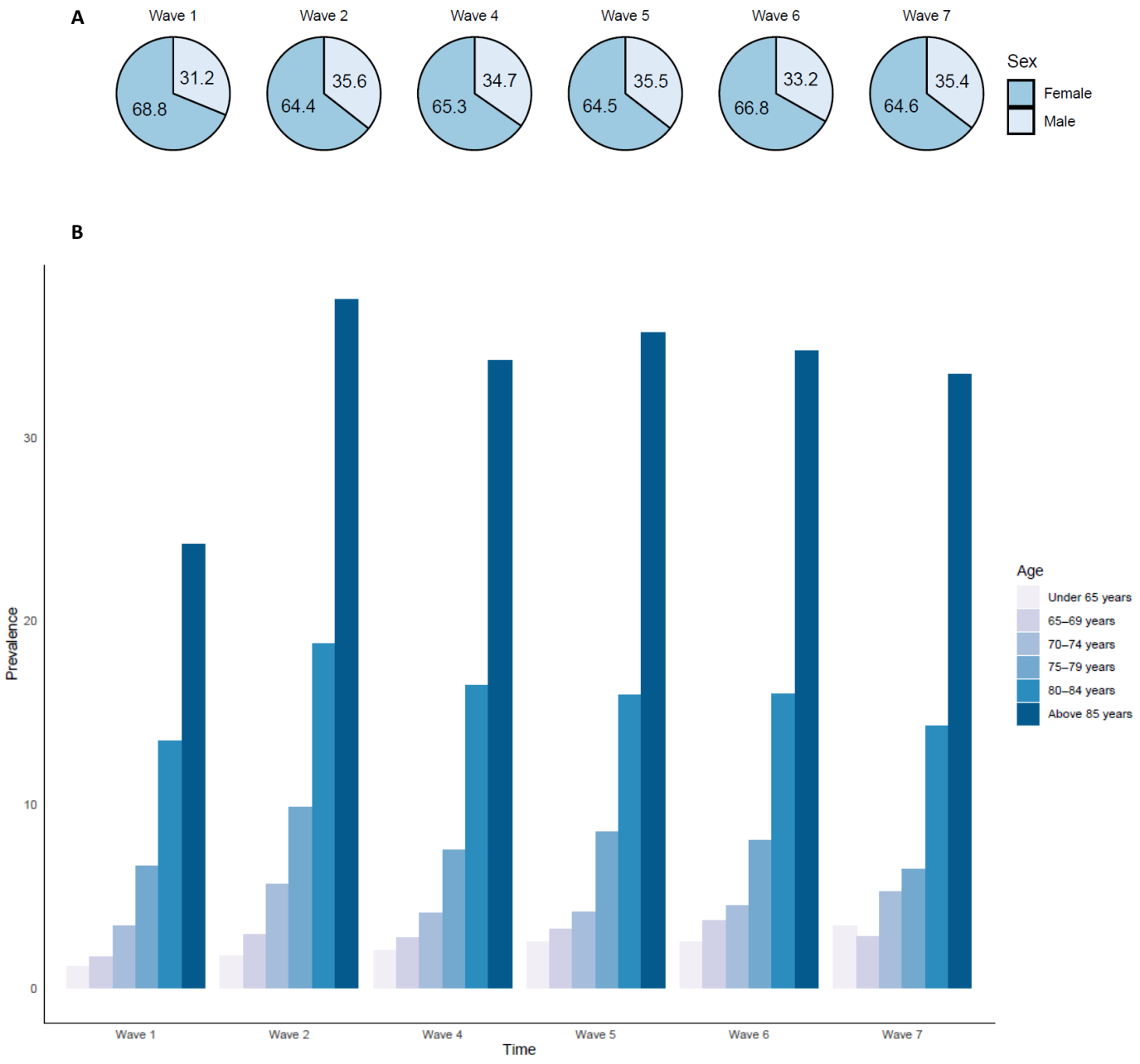
	SHARE		
	Cluster 1 (n=11,369)	Cluster 2 (n=3,374)	Cluster 3 (n=535)
Age, Years	64.6 (9.6)	65.2 (9.7)	76.5 (11)
Sex			
Female	6,251 (55%)	1,704 (50.5%)	368 (68.8%)
Male	5,118 (45%)	1,670 (49.5%)	167 (31.2%)
Education			
Less than upper secondary education	5,632 (49.5%)	1,834 (54.4%)	426 (79.6%)
Upper secondary and vocational training	3,446 (30.3%)	928 (27.5%)	69 (12.9%)
Tertiary education	2,291 (20.2%)	612 (18.1%)	40 (7.5%)
Mobilitity impairment score [0–7]	1 (1.4)	1 (1.4)	4.9 (1.5)
Autonomy			
ADL score* [0–6]	0.1 (0.3)	0.1 (0.4)	3.1 (1.7)
IADL score* [0–7]	0.2 (0.5)	0.2 (0.6)	4.2 (1.8)
Cognition			
Immediate Word Recall [0–10]*	5 (1.7)	4.7 (1.9)	2.6 (1.9)
Verbal Fluency [0–67]*	19.7 (7.1)	18.7 (7.3)	10.4 (6)

Table 1: Baseline characteristics of the SHARE study participants according to the three clusters identified by the algorithm.

(*) Values were imputed using MissMDA package.

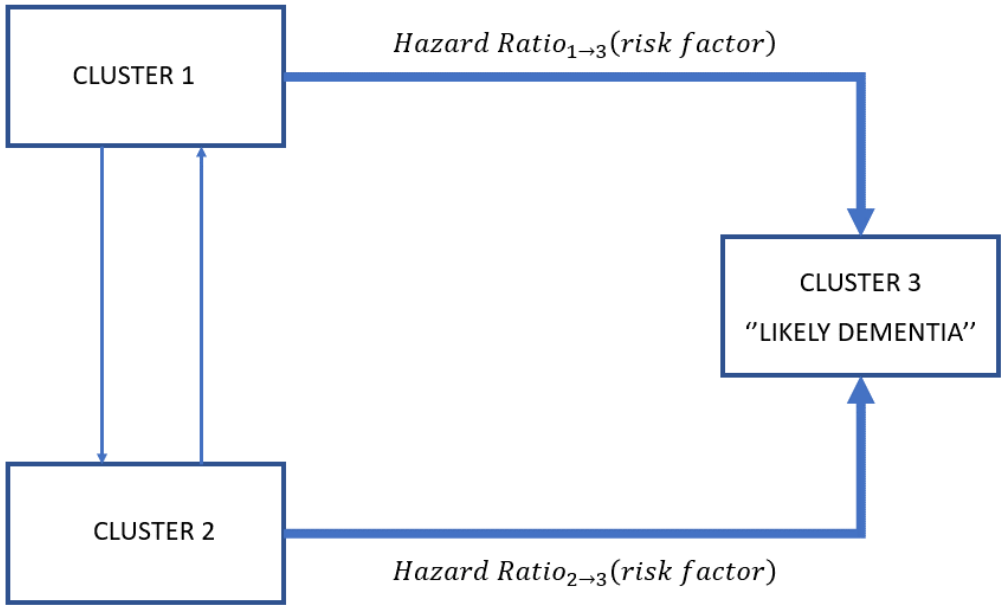
SHARE										
		Clusters			Self-reported dementia			Metrics		
Wave	Number of participants	Cluster 1	Cluster 2	Cluster 3 (Likely Dementia)	Missing	No	Yes	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Wave 1 (2004–05)	15,278	11,369 (74.4%)	3,374 (22.1%)	535 (3.5%)	NA	NA	NA	NA	NA	NA
Wave 2 (2006–07)	15,278	11,433 (74.8%)	2,832 (18.5%)	1,013 (6.6%)	40 (0.3%)	14,960 (97.9%)	278 (1.8%)	0.805 (0.776–0.835)	0.665 (0.610–0.721)	0.945 (0.942–0.949)
Wave 4 (2010–11)	10,008	7,911 (79%)	1,406 (14%)	691 (7%)	21 (0.2%)	9,735 (97.3%)	252 (2.5%)	0.825 (0.794–0.855)	0.702 (0.646–0.759)	0.947 (0.943–0.952)
Wave 5 (2012–13)	8,418	6,429 (76.4%)	1,312 (15.6%)	677 (8%)	10 (0.1%)	8,129 (96.6%)	279 (3.3%)	0.794 (0.763–0.825)	0.649 (0.593–0.705)	0.939 (0.934–0.944)
Wave 6 (2014–15)	6,485	4,913 (75.8%)	987 (15.2%)	585 (9%)	8 (0.1%)	6,204 (95.7%)	273 (4.2%)	0.755 (0.723–0.787)	0.579 (0.520–0.637)	0.931 (0.925–0.938)
Wave 7 (2016–17)	5,533	3,991 (72.1%)	1,028 (18.6%)	514 (9.3%)	5 (0.1%)	5,252 (94.9%)	276 (5%)	0.745 (0.712–0.777)	0.558 (0.499–0.617)	0.931 (0.925–0.938)

Table 2: Comparison of self-reported dementia cases and Cluster 3 "Likely Dementia" cases; Abbreviations: AUC, Area Under the Curve; CI, confidence interval; NA, not available

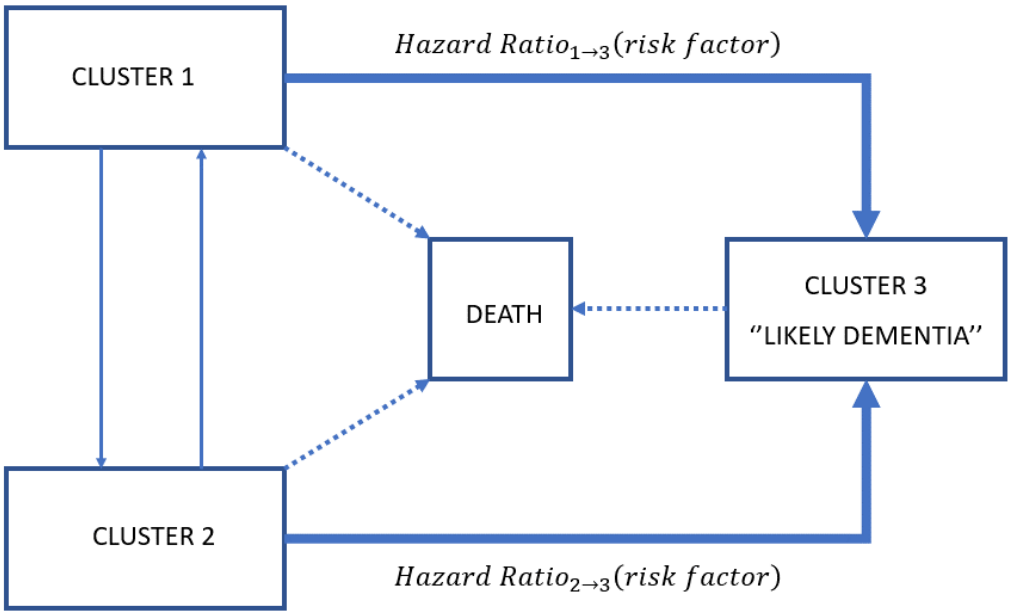


<u>Figure 2:</u> Prevalence of participants of the ''Likely dementia'' cluster by sex (2.A), and by age (2.B)

A



B



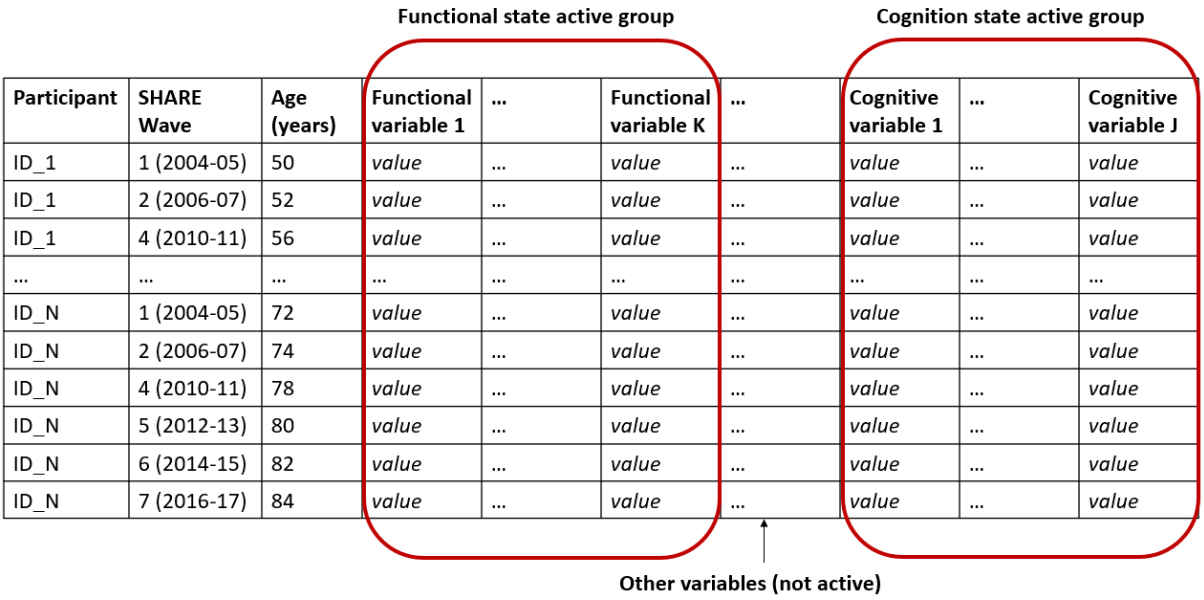
<u>Figure 3:</u> Three-state model (A) Multistate model (B) Multistate survival model

	SHARE					
	Main analysis		Sensitivity analyses			
	Model 1 (n=15,278)		Model 2 (n=15,173)		Model 3 (n=14,278)	
	HR (95% CI) (1 → 3)	HR (95% CI) (2 → 3)	HR (95% CI) (1 → 3)	HR (95% CI) (2 → 3)	HR (95% CI) (1 → 3)	HR (95% CI) (2 → 3)
Education	**1.92 (1.58-2.33)**	1.32 (0.91-1.9)	**1.86 (1.6-2.17)**	1.18 (0.86-1.61)	**1.77 (1.52-2.07)**	0.97 (0.67-1.4)
Hearing	**1.74 (1.45-2.09)**	1.23 (0.85-1.79)	**1.38 (1.2-1.59)**	1.03 (0.75-1.42)	**1.2 (1.03-1.4)**	0.88 (0.59-1.32)
Hypertension	**1.35 (1.14-1.16)**	1.24 (0.9-1.72)	**1.36 (1.2-1.55)**	1.14 (0.86-1.5)	**1.34 (1.17-1.53)**	1.09 (0.76-1.55)
Drinking (> 21 units)	0.79 (0.55-1.14)	0.42 (0.16-1.09)	1.25 (0.99-1.58)	0.36 (0.13-1.04)	**1.37 (1.09-1.73)**	0.54 (0.21-1.41)
Smoking	**1.45 (1.13-1.87)**	1.29 (0.73-2.29)	**1.64 (1.36-1.99)**	**2.23 (1.57-3.16)**	**1.7 (1.39-2.07)**	**2.68 (1.79-4.03)**
Depression	**2.51 (1.06-3.07)**	**2.39 (1.62-3.53)**	**2.05 (1.76-2.4)**	**1.98 (1.42-2.77)**	**1.78 (1.51-2.11)**	**1.65 (1.07-2.54)**
Social isolation	**1,66 (1,39-1,98)**	**2.31 (1.51-3.53)**	**1.61 (1.4-1.86)**	**1.6 (1.15-2.24)**	**1.56 (1.35-1.81)**	1.15 (0.79-1.68)
Physical inactivity	**3.66 (2.97-4.51)**	**3,21 (2.12-4.87)**	**2.48 (2.07-2.97)**	**2.89 (2-4.17)**	**2.09 (1.72-2.54)**	1.33 (0.68-2.6)
Diabetes	**2.4 (1.94-2.96)**	1.32 (0.79-2.21)	**2.15 (1.82-2.54)**	1.32 (0.85-2.05)	**2.22 (1.88-2.62)**	0.88 (0.45-1.73)
Obesity	**1.7 (1.39-2.07)**	**1.58 (1.08-2.32)**	**1.65 (1.41-1.93)**	**1.43 (1.01-2.01)**	**1.76 (1.5-2.06)**	1.32 (0.83-2.1)
Air Pollution	0.84 (0.7-1.02)	1.26 (0.83-1.9)	0.92 (0.79-1.07)	1.32 (0.92-1.89)	0.94 (0.81-1.1)	1.32 (0.83-2.08)

Table 3: Multistate models for the transition to cluster 3 ("Likely dementia") Analyses using age as time-scale. All transitions were adjusted for sex. Transition towards the third cluster ("Likely dementia") was further adjusted for age and each risk factor individually. All risk factors were taken at baseline. Main analysis was based on a multistate model (Model 1). Sensitivity analyses were based on a multistate survival model with death as an absorbing state. First, 105 participants were removed because of inconsistencies of dates (Model 2). Second, cases identified either at the first or the second wave were removed (Model 3). Abbreviations: HR, hazard ratio; CI, confidence interval.

Supplementary documents

Multiple Factor Analysis (MFA) is a principal component method similar to Principal Component Analysis (PCA) but, compared to PCA, it balances the influence of groups of variables. In our case, it balances the influence of two components, participants' function and cognition, which have equal importance in the clinical definition of dementia (data structure for SHARE is given in Supplementary Figure 1). Then, as PCA, it gives a representation of individuals in such a way that individuals are close on the representation if they have close values from the point of view of all the variables of all the groups. Subsequently, clustering can be performed on the MFA results. Full description of the variables used for SHARE and ELSA studies are available in Supplementary Table 1 and 2, respectively.



Participant	SHARE Wave	Age (years)	Functional variable 1	...	Functional variable K	...	Cognitive variable 1	...	Cognitive variable J
ID_1	1 (2004-05)	50	*value*	...	*value*	...	*value*	...	*value*
ID_1	2 (2006-07)	52	*value*	...	*value*	...	*value*	...	*value*
ID_1	4 (2010-11)	56	*value*	...	*value*	...	*value*	...	*value*
...	...	...	...	...	...	...	...	...	...
ID_N	1 (2004-05)	72	*value*	...	*value*	...	*value*	...	*value*
ID_N	2 (2006-07)	74	*value*	...	*value*	...	*value*	...	*value*
ID_N	4 (2010-11)	78	*value*	...	*value*	...	*value*	...	*value*
ID_N	5 (2012-13)	80	*value*	...	*value*	...	*value*	...	*value*
ID_N	6 (2014-15)	82	*value*	...	*value*	...	*value*	...	*value*
ID_N	7 (2016-17)	84	*value*	...	*value*	...	*value*	...	*value*

Supplementary Figure 1: Data structure in SHARE Each participant (from 1 to N) was seen multiple times (at least twice). Variables informative of functional status (from 1 to K) and variables informative of cognition (from 1 to J) formed the two active groups in the Multiple Factor Analysis (MFA).

SHARE

Group	Variable	Description	Categories	Active groups: Imputation	Active groups: MFA
Wave	wave	Wave of participation	1. Wave 1 (2004–05) 2. Wave 2 (2006–07) 3. Wave 4 (2010–11) 4. Wave 5 (2012–13) 5. Wave 6 (2014–15) 6. Wave 7 (2016–17)	X	
Sex	ragender	Respondent gender	1. Male 2. Female	X	
Age	age	(month/year of interview) - (month/year of birth)	.	X	
Education	raeducl	Respondent harmonized education level	1. Less than upper secondary 2. Upper secondary and vocational training 3. Tertiary	X	
Autonomy (qualitative variables)	hlthlma	Respondent has health problem that limits activities	0. Not limited 1. Limited	X	X
	walkra	Respondent has some difficulty walking across the room	0. No 1. Yes		
	dressa	Respondent has some difficulty dressing	0. No 1. Yes		
	batha	Respondent has some difficulty bathing/taking a shower	0. No 1. Yes		
	eata	Respondent has some difficulty eating	0. No 1. Yes		
	beda	Respondent has some difficulty getting in/out of the bed	0. No 1. Yes		
	toilta	Respondent has some difficulty using the toilet	0. No 1. Yes		
	phonea	Respondent has some difficulty using the telephone	0. No 1. Yes		
	medsa	Respondent has some difficulty taking medications	0. No 1. Yes		
	moneya	Respondent has some difficulty managing money	0. No 1. Yes		
	shopa	Respondent has some difficulty buying grocery	0. No 1. Yes		
	mealsa	Respondent has some difficulty preparing hot meal	0. No 1. Yes		
	mapa	Respondent has some difficulty using a map	0. No 1. Yes		

	housewka	Respondent has some difficulty doing household work around house	0. No 1. Yes		
	walk100a	Respondent has some difficulty walking 100m	0. No 1. Yes		
	sita	Respondent has some difficulty sitting for 2 hours	0. No 1. Yes		
	chaira	Respondent has some difficulty getting up from chair	0. No 1. Yes		
	climsa	Respondent has some difficulty climbing several flights of stairs	0. No 1. Yes		
	clim1a	Respondent has some difficulty climbing one flight of stairs	0. No 1. Yes		
	lifta	Respondent has some difficulty lifting/carrying 10lbs	0. No 1. Yes		
	stoopa	Respondent has some difficulty stooping/kneeling/crouching	0. No 1. Yes		
	armsa	Respondent has some difficulty reaching/extending arms up	0. No 1. Yes		
	pusha	Respondent has some difficulty pushing/pulling long object	0. No 1. Yes		
	dimea	Respondent has some difficulty picking up a small coin	0. No 1. Yes		
Autonomy (quantitative variables)	adltot_s	Some difficulty in Activities of Daily Living	.	X	X
	iadltot1_s	Any difficulty in total Instrumental Activities of Daily Living	.		
Cognition (conditions)	cogimp	Whether factors impaired cognitive tests	0. No 1. Yes	X	X
	cogothp	Whether other people present during cognitive tests	0. No 1. Yes		
Cognition (tests)	imrc	Immediate word recall	.	X	X
	dlrc	Delayed word recall	.		
	tr20	Respondent recall summary score	.		
	verbf	Respondent verbal fluency score	.		

<u>Supplementary Table 1:</u> Summary of variables used for both imputation and Multiple Factor analysis (MFA) in SHARE

ELSA

Group	Variable	Description	Categories	Active groups: Imputation	Active groups: MFA
Wave	wave	Wave of participation	1. Wave 1 (2002–03) 2. Wave 2 (2004–05) 3. Wave 3 (2006–07) 4. Wave 4 (2008–09) 5. Wave 5 (2010–11) 6. Wave 6 (2012–13) 7. Wave 7 (2014–15) 8. Wave 8 (2016–17) 9. Wave 9 (2018–19)	X	
Sex	ragender	Respondent gender	1. Male 2. Female	X	
Age	agey	Respondent age (years) at interview	.	X	
Education	raeducl	Respondent harmonized education level	1. Less than upper secondary 2. Upper secondary and vocational training 3. Tertiary	X	
Autonomy (qualitative variables)	hlthlma	Respondent has health problem that limits activities	0. Not limited 1. Limited	X	X
	walkra	Respondent has some difficulty walking across the room	0. No 1. Yes		
	dressa	Respondent has some difficulty dressing	0. No 1. Yes		
	batha	Respondent has some difficulty bathing/taking a shower	0. No 1. Yes		
	eata	Respondent has some difficulty eating	0. No 1. Yes		
	beda	Respondent has some difficulty getting in/out of the bed	0. No 1. Yes		
	toilta	Respondent has some difficulty using the toilet	0. No 1. Yes		
	phonea	Respondent has some difficulty using the telephone	0. No 1. Yes		
	medsa	Respondent has some difficulty taking medications	0. No 1. Yes		
	moneya	Respondent has some difficulty managing money	0. No 1. Yes		
	shopa	Respondent has some difficulty buying grocery	0. No 1. Yes		
	mealsa	Respondent has some difficulty preparing hot meal	0. No 1. Yes		

	mapa	Respondent has some difficulty using a map	0. No 1. Yes		
	housewka	Respondent has some difficulty doing household work	0. No 1. Yes		
	walk100a	Respondent has some difficulty walking 100m	0. No 1. Yes		
	sita	Respondent has some difficulty sitting for 2 hours	0. No 1. Yes		
	chaira	Respondent has some difficulty getting up from chair	0. No 1. Yes		
	climsa	Respondent has some difficulty climbing several flights of stairs	0. No 1. Yes		
	clim1a	Respondent has some difficulty climbing one flight of stairs (straight)	0. No 1. Yes		
	lifta	Respondent has some difficulty lifting/carrying 10lbs	0. No 1. Yes		
	stoopa	Respondent has some difficulty stooping/kneeling/crouching	0. No 1. Yes		
	armsa	Respondent has some difficulty reaching/extending arms up	0. No 1. Yes		
	pusha	Respondent has some difficulty pushing/pulling long object	0. No 1. Yes		
	dimea	Respondent has some difficulty picking up a small coin	0. No 1. Yes		
Autonomy (quantitative variables)	adltot_s	Some difficulty in Activities of Daily Living	.	X	X
	iadltot1_s	Any difficulty in total Instrumental Activities of Daily Living	.		
Cognition (conditions)	cogimp	Whether factors impaired cognitive tests	0. No 1. Yes	X	X
	cogothp	Whether other people present during cognitive tests	0. No 1. Yes		
Cognition (tests)	imrc	Immediate word recall	.	X	X
	dlrc	Delayed word recall	.		
	tr20	Respondent recall summary score	.		
	verbf	Respondent verbal fluency score	.		

<u>Supplementary Table 2:</u> Summary of variables used for both imputation and Multiple Factor analysis (MFA) in ELSA

SHARE

Country	Wave	Number of participants	Clusters			Self-reported dementia			Metrics		
			Cluster 1	Cluster 2	Cluster 3	Missing	No	Yes	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Austria	Wave 1 (2004–05)	888	686 (77.3%)	177 (19.9%)	25 (2.8%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	888	704 (79.3%)	132 (14.9%)	52 (5.8%)	1 (0.1%)	870 (98%)	17 (1.9%)	0.707 (0.580–0.833)	0.471 (0.233–0.708)	0.943 (0.927–0.958)
	Wave 4 (2010–11)	549	436 (79.4%)	84 (15.3%)	29 (5.3%)	4 (0.7%)	525 (95.6%)	20 (3.7%)	0.746 (0.627–0.865)	0.550 (0.332–0.768)	0.943 (0.923–0.963)
	Wave 5 (2012–13)	455	365 (80.2%)	52 (11.4%)	38 (8.4%)	4 (0.9%)	423 (93%)	28 (6.1%)	0.729 (0.624–0.834)	0.536 (0.351–0.720)	0.922 (0.896–0.948)
	Wave 6 (2014–15)	382	306 (80.1%)	40 (10.5%)	36 (9.4%)	0 (0%)	348 (91.1%)	34 (8.9%)	0.757 (0.660–0.853)	0.588 (0.423–0.754)	0.925 (0.898–0.953)
	Wave 7 (2016–17)	312	248 (79.5%)	34 (10.9%)	30 (9.6%)	0 (0%)	287 (92%)	25 (8%)	0.700 (0.586–0.814)	0.480 (0.284–0.676)	0.920 (0.888–0.951)
Belgium	Wave 1 (2004–05)	2,471	1 740 (70.4%)	653 (26.4%)	78 (3.2%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	2,471	1 793 (72.6%)	567 (22.9%)	111 (4.5%)	1 (0%)	2,442 (98.8%)	28 (1.2%)	0.798 (0.705–0.891)	0.643 (0.465–0.820)	0.953 (0.945–0.961)
	Wave 4 (2010–11)	1,779	1 380 (77.6%)	306 (17.2%)	93 (5.2%)	0 (0%)	1,748 (98.3%)	31 (1.7%)	0.781 (0.690–0.872)	0.613 (0.441–0.784)	0.949 (0.939–0.959)
	Wave 5 (2012–13)	1,546	1 149 (74.3%)	316 (20.5%)	81 (5.2%)	2 (0.1%)	1,502 (97.2%)	42 (2.7%)	0.710 (0.628–0.791)	0.476 (0.325–0.627)	0.943 (0.932–0.955)
	Wave 6 (2014–15)	1,357	1 026 (75.6%)	240 (17.7%)	91 (6.7%)	1 (0.1%)	1,305 (96.2%)	51 (3.7%)	0.690 (0.615–0.765)	0.451 (0.314–0.588)	0.929 (0.915–0.943)
	Wave 7 (2016–17)	1,169	866 (74.1%)	231 (19.8%)	72 (6.1%)	1 (0.1%)	1,116 (95.5%)	52 (4.4%)	0.688 (0.614–0.763)	0.442 (0.307–0.577)	0.935 (0.920–0.949)
Denmark	Wave 1 (2004–05)	1,073	830 (77.4%)	216 (20.1%)	27 (2.5%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	1,073	889 (82.9%)	125 (11.6%)	59 (5.5%)	0 (0%)	1,058 (98.6%)	15 (1.4%)	0.876 (0.769–0.984)	0.811 (0.620–1.000)*	0.953 (0.940–0.966)

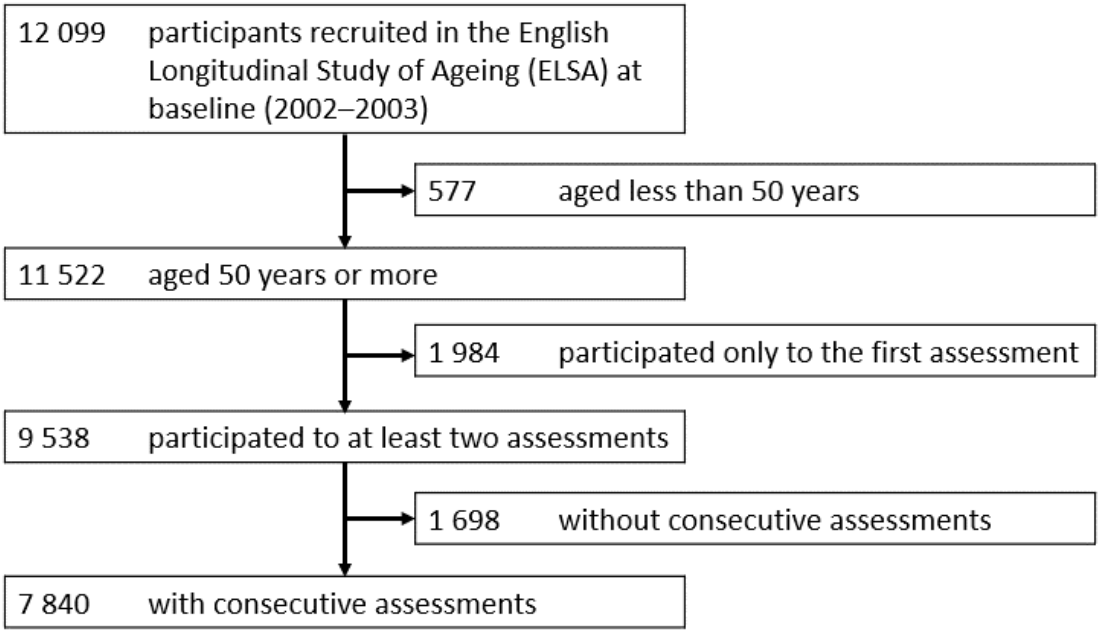
	Wave 4 (2010–11)	776	661 (85.2%)	84 (10.8%)	31 (4%)	2 (0.3%)	762 (98.2%)	12 (1.5%)	0.814 (0.674–0.954)	0.667 (0.400–0.933)	0.962 (0.948–0.976)
	Wave 5 (2012–13)	682	574 (84.2%)	75 (11%)	33 (4.8%)	0 (0%)	669 (98.1%)	13 (1.9%)	0.822 (0.689–0.956)	0.692 (0.441–0.943)	0.952 (0.936–0.968)
	Wave 6 (2014–15)	603	505 (83.8%)	73 (12.1%)	25 (4.1%)	2 (0.3%)	586 (97.2%)	15 (2.5%)	0.711 (0.577–0.846)	0.467 (0.214–0.719)	0.956 (0.939–0.972)
	Wave 7 (2016–17)	514	384 (74.7%)	101 (19.7%)	29 (5.6%)	0 (0.0%)	501 (97.5%)	13 (2.5%)	0.783 (0.641–0.925)	0.615 (0.351–0.880)	0.950 (0.931–0.969)
France	Wave 1 (2004–05)	1,661	1,182 (71.2%)	424 (25.5%)	55 (3.3%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	1,661	1,246 (75%)	334 (20.1%)	81 (4.9%)	22 (1.3%)	1,607 (96.8%)	32 (1.9%)	0.822 (0.736–0.907)	0.688 (0.527–0.848)	0.956 (0.946–0.966)
	Wave 4 (2010–11)	1,153	935 (81.1%)	141 (12.2%)	77 (6.7%)	10 (0.9%)	1,108 (96.1%)	35 (3%)	0.814 (0.730–0.898)	0.686 (0.532–0.840)	0.942 (0.929–0.956)
	Wave 5 (2012–13)	899	737 (82%)	102 (11.3%)	60 (6.7%)	2 (0.2%)	865 (96.2%)	32 (3.6%)	0.877 (0.801–0.952)	0.812 (0.677–0.948)	0.941 (0.925–0.957)
	Wave 6 (2014–15)	709	543 (76.6%)	132 (18.6%)	34 (4.8%)	1 (0.1%)	692 (97.6%)	16 (2.3%)	0.812 (0.689–0.935)	0.688 (0.460–0.915)	0.936 (0.918–0.955)
	Wave 7 (2016–17)	578	425 (73.5%)	128 (22.2%)	25 (4.3%)	1 (0.2%)	560 (96.9%)	17 (2.9%)	0.832 (0.715–0.948)	0.706 (0.489–0.922)	0.957 (0.940–0.974)
Germany	Wave 1 (2004–05)	1,509	1,169 (77.5%)	300 (19.9%)	40 (2.6%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	1,509	1,136 (75.3%)	303 (20.1%)	70 (4.6%)	2 (0.1%)	1,473 (97.6%)	34 (2.3%)	0.802 (0.716–0.887)	0.647 (0.486–0.808)	0.957 (0.946–0.967)
	Wave 4 (2010–11)	903	725 (80.3%)	130 (14.4%)	48 (5.3%)	0 (0.0%)	884 (97.9%)	19 (2.1%)	0.845 (0.739–0.951)	0.737 (0.539–0.935)	0.952 (0.938–0.967)
	Wave 5 (2012–13)	621	467 (75.2%)	118 (19%)	36 (5.8%)	0 (0.0%)	605 (97.4%)	16 (2.6%)	0.783 (0.655–0.911)	0.625 (0.388–0.862)	0.940 (0.922–0.959)
	Wave 6 (2014–15)	562	425 (75.6%)	106 (18.9%)	31 (5.5%)	0 (0.0%)	539 (95.9%)	23 (4.1%)	0.732 (0.620–0.844)	0.522 (0.318–0.726)	0.942 (0.923–0.962)
	Wave 7 (2016–17)	482	376 (78%)	81 (16.8%)	25 (5.2%)	0 (0.0%)	459 (95.2%)	23 (4.8%)	0.753 (0.641–0.865)	0.565 (0.363–0.768)	0.941 (0.920–0.963)
Greece	Wave 1 (2004–05)	742	593 (79.9%)	83 (11.2%)	66 (8.9%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	742	621 (83.7%)	46 (6.2%)	75 (10.1%)	3 (0.4%)	724 (97.6%)	15 (2%)	0.747 (0.612–0.882)	0.600 (0.352–0.848)	0.894 (0.871–0.916)

Israel	Wave 1 (2004–05)	241	140 (58.1%)	74 (30.7%)	27 (11.2%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	241	125 (51.9%)	50 (20.7%)	66 (27.4%)	5 (2.1%)	214 (88.8%)	22 (9.1%)	0.806 (0.705–0.906)	0.873 (0.713–1.000)*	0.748 (0.689–0.806)
Italy	Wave 1 (2004–05)	1,471	1,048 (71.2%)	351 (23.9%)	72 (4.9%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	1,471	1,090 (74.1%)	273 (18.6%)	108 (7.3%)	0 (0%)	1 449 (98.5%)	22 (1.5%)	0.755 (0.646–0.865)	0.591 (0.385–0.796)	0.920 (0.906–0.934)
	Wave 4 (2010–11)	1,165	918 (78.8%)	153 (13.1%)	94 (8.1%)	1 (0.1%)	1 134 (97.3%)	30 (2.6%)	0.879 (0.804–0.953)	0.833 (0.700–0.967)	0.924 (0.909–0.940)
	Wave 5 (2012–13)	1,029	774 (75.2%)	171 (16.6%)	84 (8.2%)	0 (0%)	988 (96%)	41 (4%)	0.825 (0.749–0.902)	0.732 (0.596–0.867)	0.919 (0.902–0.936)
	Wave 6 (2014–15)	924	709 (76.7%)	145 (15.7%)	70 (7.6%)	1 (0.1%)	879 (95.1%)	44 (4.8%)	0.756 (0.674–0.837)	0.591 (0.446–0.736)	0.920 (0.902–0.938)
	Wave 7 (2016–17)	819	586 (71.6%)	159 (19.4%)	74 (9%)	0 (0%)	765 (93.4%)	54 (6.6%)	0.774 (0.700–0.848)	0.630 (0.501–0.758)	0.919 (0.900–0.938)
The Netherlands	Wave 1 (2004–05)	1,678	1,212 (72.2%)	439 (26.2%)	27 (1.6%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	1,678	1,206 (71.9%)	432 (25.7%)	40 (2.4%)	3 (0.2%)	1,655 (98.6%)	20 (1.2%)	0.687 (0.576–0.798)	0.400 (0.185–0.615)	0.974 (0.966–0.982)
	Wave 4 (2010–11)	1,124	905 (80.5%)	184 (16.4%)	35 (3.1%)	2 (0.2%)	1,101 (97.9%)	21 (1.9%)	0.700 (0.589–0.811)	0.429 (0.217–0.640)	0.972 (0.962–0.982)
	Wave 5 (2012–13)	951	721 (75.8%)	201 (21.1%)	29 (3.1%)	1 (0.1%)	926 (97.4%)	24 (2.5%)	0.780 (0.676–0.883)	0.583 (0.386–0.781)	0.976 (0.966–0.986)
Spain	Wave 1 (2004–05)	1,239	878 (70.9%)	283 (22.8%)	78 (6.3%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	1,239	832 (67.1%)	287 (23.2%)	120 (9.7%)	1 (0.1%)	1,204 (97.2%)	34 (2.7%)	0.860 (0.788–0.933)	0.824 (0.695–0.952)	0.897 (0.880–0.914)
	Wave 4 (2010–11)	942	601 (63.8%)	219 (23.2%)	122 (13%)	1 (0.1%)	892 (94.7%)	49 (5.2%)	0.893 (0.840–0.946)	0.898 (0.813–0.983)	0.888 (0.867–0.909)
	Wave 5 (2012–13)	848	506 (59.7%)	218 (25.7%)	124 (14.6%)	0 (0%)	795 (93.8%)	53 (6.2%)	0.825 (0.758–0.891)	0.792 (0.683–0.902)	0.857 (0.832–0.881)
	Wave 6 (2014–15)	725	470 (64.8%)	150 (20.7%)	105 (14.5%)	0 (0%)	669 (92.3%)	56 (7.7%)	0.811 (0.742–0.879)	0.768 (0.657–0.878)	0.854 (0.827–0.880)
	Wave 7 (2016–17)	619	382 (61.7%)	152 (24.6%)	85 (13.7%)	2 (0.3%)	562 (90.8%)	55 (8.9%)	0.743 (0.664–0.821)	0.655 (0.529–0.780)	0.831 (0.800–0.862)

Sweden	Wave 1 (2004–05)	1,652	1,345 (81.4%)	272 (16.5%)	35 (2.1%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	1,652	1,337 (80.9%)	252 (15.3%)	63 (3.8%)	2 (0.1%)	1,618 (98%)	32 (1.9%)	0.831 (0.747–0.915)	0.688 (0.527–0.848)	0.974 (0.966–0.982)
	Wave 4 (2010–11)	1,131	978 (86.5%)	104 (9.2%)	49 (4.3%)	1 (0.1%)	1,106 (97.8%)	24 (2.1%)	0.838 (0.742–0.934)	0.708 (0.526–0.890)	0.967 (0.957–0.978)
	Wave 5 (2012–13)	964	866 (89.8%)	64 (6.7%)	34 (3.5%)	1 (0.1%)	939 (97.4%)	24 (2.5%)	0.774 (0.669–0.878)	0.583 (0.386–0.781)	0.964 (0.952–0.976)
	Wave 6 (2014–15)	847	711 (84%)	107 (12.6%)	29 (3.4%)	3 (0.3%)	817 (96.5%)	27 (3.2%)	0.741 (0.640–0.842)	0.519 (0.330–0.707)	0.963 (0.950–0.976)
	Wave 7 (2016–17)	713	555 (77.9%)	135 (18.9%)	23 (3.2%)	1 (0.1%)	685 (96.1%)	27 (3.8%)	0.705 (0.604–0.805)	0.444 (0.257–0.632)	0.965 (0.951–0.979)
Switzerland	Wave 1 (2004–05)	653	546 (83.6%)	102 (15.6%)	5 (0.8%)	NA	NA	NA	NA	NA	NA
	Wave 2 (2006–07)	653	560 (85.8%)	80 (12.3%)	13 (2%)	0 (0%)	646 (98.9%)	7 (1.1%)	0.777 (0.589–0.965)	0.571 (0.205–0.938)	0.983 (0.973–0.993)
	Wave 4 (2010–11)	486	450 (92.6%)	26 (5.3%)	10 (2.1%)	0 (0%)	475 (97.7%)	11 (2.3%)	0.766 (0.614–0.919)	0.545 (0.251–0.840)	0.987 (0.977–0.997)
	Wave 5 (2012–13)	423	389 (92%)	28 (6.6%)	6 (1.4%)	0 (0%)	417 (98.6%)	6 (1.4%)	0.575 (0.420–0.730)	0.237 (0.076–0.544)*	0.983 (0.971–0.996)
	Wave 6 (2014–15)	376	330 (87.8%)	38 (10.1%)	8 (2.1%)	0 (0%)	369 (98.1%)	7 (1.9%)	0.632 (0.457–0.807)	0.341 (0.075–0.611)*	0.978 (0.963–0.993)
	Wave 7 (2016–17)	327	281 (85.9%)	39 (11.9%)	7 (2.1%)	0 (0%)	317 (96.9%)	10 (3.1%)	0.689 (0.529–0.849)	0.400 (0.096–0.704)	0.978 (0.962–0.994)

<u>Supplementary Table 3:</u> Self-reported dementia cases / Cluster 3 "Outcome" Comparison by country (SHARE), Abbreviations: NA, not available

*Values obtained using bootstrapping

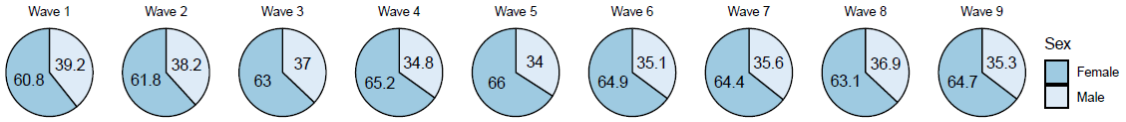


Supplementary Figure 2: ELSA flow chart

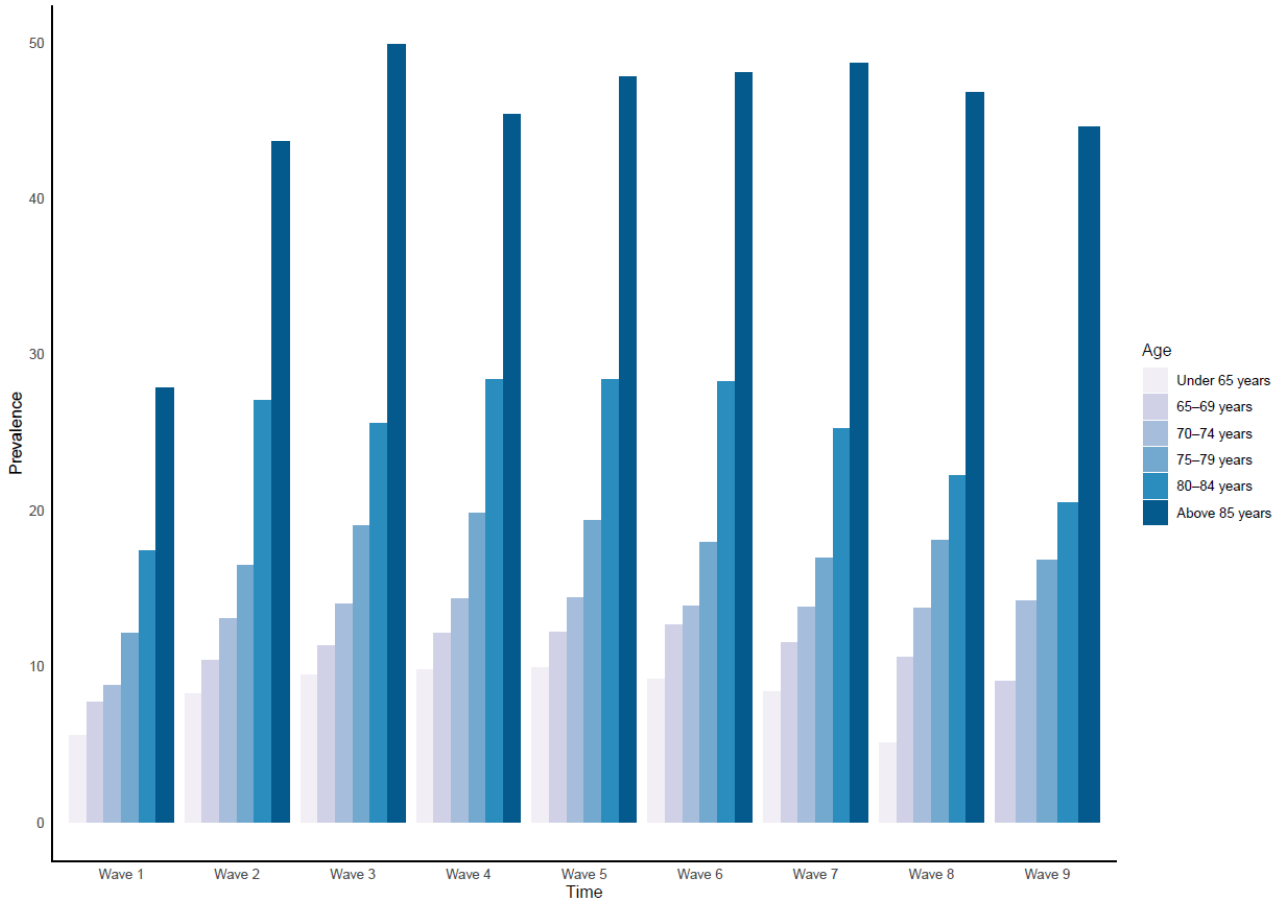
	ELSA		
	Cluster 1 (n=6,556)	Cluster 2 (n=625)	Cluster 3 (n=659)
Age, Years	64.1 (9.7)	65.8 (10.4)	69.8 (11)
Sex			
Female	3,584 (54.6%)	322 (51.5%)	401 (60.8%)
Male	2,976 (45.4%)	303 (48.5%)	258 (39.2%)
Education			
Less than upper secondary education	2,646 (40.4%)	345 (55.2%)	423 (64.2%)
Upper secondary and vocational training	3,010 (45.9%)	247 (39.5%)	208 (31.6%)
Tertiary education	900 (13.7%)	33 (5.3%)	28 (4.2%)
Mobility difficulty score [0–7]	1.1 (1.4)	1.6 (1.8)	4.9 (1.4)
Autonomy			
ADL score [0–6]	0.1 (0.4)	0.3 (0.7)	2.7 (1.5)
IADL score [0–7]	0.1 (0.4)	0.3 (0.7)	2.5 (1.4)
Cognition			
Immediate Word Recall [0–10]	5.7 (1.7)	5.2 (1.8)	4.6 (1.9)
Verbal Fluency [0–49]	20.2 (6.1)	18.7 (6.4)	16.5 (6)

Supplementary Table 4: Baseline characteristics of the ELSA study participants according to the three clusters identified by the algorithm

A



B



<u>Supplementary Figure 3:</u> Prevalence of participants from the ''Likely dementia'' cluster by sex (3.A), and by age (3.B)

		ELSA								
		Clusters			Self-reported dementia			Metrics		
Wave	Number of participants	Cluster 1	Cluster 2	Cluster 3 (Likely Dementia)	Missing	No	Yes	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Wave 1 (2002–03)	7,840	6,556 (83.6%)	625 (8%)	659 (8.4%)	3 (0%)	7,813 (99.7%)	24 (0.3%)	0.646 (0.546–0.746)	0.375 (0.181–0.569)	0.917 (0.911–0.923)
Wave 2 (2004–05)	7,840	6,210 (79.2%)	566 (7.2%)	1 064 (13.6%)	0 (0%)	7,778 (99.2%)	62 (0.8%)	0.749 (0.685–0.813)	0.629 (0.509–0.749)	0.868 (0.861–0.876)
Wave 3 (2006–07)	6,662	5,204 (78.6%)	380 (5.7%)	1,038 (15.7%)	0 (0%)	6,550 (98.9%)	72 (1.1%)	0.814 (0.762–0.866)	0.778 (0.682–0.874)	0.850 (0.841–0.859)
Wave 4 (2008–09)	5,720	4,468 (78.1%)	292 (5.1%)	960 (16.8%)	0 (0%)	5,630 (98.4%)	90 (1.6%)	0.832 (0.788–0.877)	0.822 (0.743–0.901)	0.843 (0.833–0.852)
Wave 5 (2010–11)	5,125	3,956 (77.2%)	257 (5%)	912 (17.8%)	0 (0%)	5,018 (97.9%)	107 (2.1%)	0.834 (0.793–0.874)	0.832 (0.761–0.903)	0.836 (0.826–0.846)
Wave 6 (2012–13)	4,642	3,525 (75.9%)	267 (5.8%)	850 (18.3%)	0 (0%)	4,544 (97.9%)	98 (2.1%)	0.829 (0.786–0.872)	0.827 (0.752–0.901)	0.831 (0.820–0.842)
Wave 7 (2014–15)	4,067	3,127 (76.9%)	190 (4.7%)	750 (18.4%)	0 (0%)	3,963 (97.4%)	104 (2.6%)	0.805 (0.759–0.851)	0.779 (0.699–0.859)	0.831 (0.820–0.843)
Wave 8 (2016–17)	3,551	2,764 (77.8%)	125 (3.5%)	662 (18.7%)	0 (0%)	3,459 (97.4%)	92 (2.6%)	0.767 (0.714–0.820)	0.707 (0.613–0.800)	0.827 (0.815–0.840)
Wave 9 (2018–19)	3,033	2,314 (76.3%)	135 (4.4%)	584 (19.3%)	0 (0%)	2,939 (96.9%)	94 (3.1%)	0.785 (0.734–0.836)	0.745 (0.657–0.833)	0.825 (0.811–0.839)

Supplementary Table 5: Comparison of self-reported dementia cases and Cluster 3 "Likely dementia" cases. Abbreviations: AUC, Area under the Curve; CI, confidence interval.

	ELSA			
	Main analysis		Sensitivity analysis	
	Model 1 (n=7,840)		Model 2 (n=6,784)	
	HR (95% CI) (1 → 3)	HR (95% CI) (2 → 3)	HR (95% CI) (1 → 3)	HR (95% CI) (2 → 3)
Education	**1.89 (1.64-2.17)**	1.19 (0.77-1.82)	**1.6 (1.37-1.87)**	1.07 (0.62-1.85)
Hearing	**1.75 (1.52-2.03)**	1.35 (0.86-2.12)	**1.63 (1.39-1.91)**	1.27 (0.7-2.29)
Hypertension	**1.42 (1.24-1.62)**	**1.64 (1.13-2.38)**	**1.39 (1.2-1.61)**	1.3 (0.78-2.16)
Drinking (> 21 units)	**0.6 (0.43-0.83)**	1.23 (0.68-2.24)	0.79 (0.58-1.08)	1 (0.37-2.7)
Smoking	**1.93 (1.6-2.33)**	1.63 (0.98-2.71)	**1.18 (1.79-2.64)**	1.7 (0.87-3.33)
Depression	**2.03 (1.74-2.37)**	**2 (1.26-3.17)**	**1.86 (1.57-2.19)**	1.3 (0.69-2.42)
Social isolation	**1.61 (1.38-1.86)**	1.52 (0.93-2.47)	**1.41 (1.2-1.65)**	1.61 (0.85-3.06)
Physical inactivity	**2.65 (2.27-3.1)**	**2.69 (1.73-4.18)**	**1.74 (1.42-2.13)**	**2.02 (1.1-3.69)**
Diabetes	**1.77 (1.38-2.26)**	**2.23 (1.26-3.95)**	**1.62 (1.24-2.13)**	1.84 (0.87-3.9)
Obesity	**1.53 (1.32-1.77)**	0.97 (0.63-1.48)	**1.62 (1.38-1.9)**	0.8 (0.44-1.46)
Pollution	NA	NA	NA	NA

Supplementary Table 6: Multistate models for the transition to cluster 3 ("Likely dementia") Analyses using age as time-scale. All transitions were adjusted for sex. Transition towards the third cluster ("Likely dementia") was further adjusted for age and each risk factor individually. All risk factors were taken at baseline. Main analysis was based on a multistate model (Model 1). In sensitivity analysis, cases identified either at the first or the second wave were removed (Model 2). Abbreviations: HR, hazard ratio; CI, confidence interval; NA, not available