Approximating dementia prevalence in population-based surveys of aging worldwide: An unsupervised machine learning approach

Laurent Cleret de Langavant, Eléonore Bayen, Anne-catherine Bachoud-Lévi, Kristine Yaffe

Received: 30 March 2020 | Revised: 28 July 2020 | Accepted: 28 July 2020 | Published online: 26 August 2020

DOI: 10.1002/trc2.12074

Alzheimer's & Dementia
Translational Research & Clinical Interventions

RESEARCH ARTICLE

Approximating dementia prevalence in population-based surveys of aging worldwide: An unsupervised machine learning approach

Laurent Cleret de Langavant[1,2,3,4,5] | **Eléonore Bayen**[5,6] | **Anne-Catherine Bachoud-Lévi**[1,2,3,4] | **Kristine Yaffe**[5,7]

[1] Département d'Etudes Cognitives, École normale supérieure, PSL University, Paris, France

[2] Faculté de medicine, Université Paris-Est Créteil, Créteil, France

[3] Equipe E01 NeuroPsychologie Interventionnelle, Inserm U955, Institut Mondor de Recherche Biomédicale, Créteil, France

[4] AP-HP, Centre de référence Maladie de Huntington, Service de Neurologie, Hôpital Henri Mondor-Albert Chenevier, Créteil, France

[5] UCSF, Global Brain Health Institute, San Francisco, California, USA

[6] Médecine Physique et de Réadaptation, Faculté de Médecine, Sorbonne Université, Paris, France

[7] Center for Population Brain Health, Department of Psychiatry, Neurology and Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA

Correspondance
Dr Laurent Cleret de Langavant, Service de Neurologie, CHU Henri Mondor, Créteil F-94010, France.
Email: laurent.cleret@gbhi.org

Funding information
Alzheimer's Association, Grant/Award Number: GBHI_ALZ-18-544248; Laurent Cleret de Langavant and Anne-Catherine Bachoud-Lévi, Grant/Award Number: ANR-17-EURE-0017; NIA, Grant/Award Number: K24 AG031155; National Institute on Aging, Grant/Award Number: U01AG009740; European Commission, Grant/Award Numbers: QLK6-CT-2001-00360, RII-CT-2006-062193, CIT5-CT-2005-028857, N°211909, N°227822, N°261982; National Institutes of Health; National Institute on Aging, Grant/Award Number: R01AG018016; Federal University of Minas Gerais; Brazilian Ministry of Health; Ministry of Science, Technology, Innovation and Communication; National Natural Science Foundation of China; National Institute on Aging; National Institute on Aging; National Institute on Aging; OGHA, Grant/Award Numbers: 04034785, YA1323-08-CN-0020, Y1-AG-1005-01, R01-AG034479, IR21-AG034263-0182; South African National Department of Health; National Institute

Abstract

Introduction: Ability to determine dementia prevalence in low- and middle-income countries (LMIC) remains challenging because of frequent lack of data and large discrepancies in dementia case ascertainment.

Methods: High likelihood of dementia was determined with hierarchical clustering after principal component analysis applied in 10 population surveys of aging: HRS (USA, 2014), SHARE (Europe and Israel, 2015), MHAS (Mexico, 2015), ELSI (Brazil, 2016), CHARLS (China, 2015), IFLS (Indonesia, 2014–2015), LASI (India, 2016), SAGE-Ghana (2007), SAGE-South Africa (2007), SAGE-Russia (2007–2010). We approximated dementia prevalence using weighting methods.

Results: Estimated numbers of dementia cases were: China, 40.2 million; India, 18.0 million; Russia, 5.2 million; Europe and Israel, 5.0 million; United States, 4.4 million; Brazil, 2.2 million; Mexico, 1.6 million; Indonesia, 1.3 million; South Africa, 1.0 million; Ghana, 319,000.

Discussion: Our estimations were similar to prior ones in high-income countries but much higher in LMIC. Extrapolating these results globally, we suggest that almost 130 million people worldwide were living with dementia in 2015.

KEYWORDS
dementia, low- and middle-income countries, machine learning, population survey, prevalence

on Aging, Grant/Award Number: 2R01 AG026676-05; National Institute for Child Health and Human Development, Grant/Award Number: 2R01 HD050764-05A1; World Bank, Indonesia and GRM International; Department of Foreign Affairs and Trade; National Institute of Aging, Grant/Award Numbers: R01 AG030153, RC2 AG036691, R03 AG043052

1 | INTRODUCTION

Persons with dementia have acquired cognitive and behavioral disorders leading to progressive functional impairment.[1] Global dementia prevalence has been estimated between 43.8 and 46.8 million persons in 2015–2016[2,3] and is expected to increase at a fast pace, especially in low- and middle-income countries (LMIC) where life expectancy is rising faster than in high-income countries (HIC).[2]

Estimation of global dementia prevalence remains a challenge despite the considerable efforts achieved by large international consortiums.[2,3] Indeed, data about dementia remain scarce or nonexistent in many LMIC and when data are available, the sample is often not fully representative of the country but only targets one or two—often urban—areas.[2,3] Another key challenge is the variety of case definitions in dementia prevalence studies ranging from medical records, International Classification of Diseases[4] (ICD) diagnoses, and in-person interviews, all of which vary in sensitivity and specificity.[2,3] Comparison of the Diagnostic and Statistical Manual of Mental Disorders[1] (DSM) and 10/66 measures in Latin America, India, and China has shown that the DSM method tends to underestimate dementia prevalence.[5] The use of medical records is also problematic because up to 50% of persons with dementia might be undiagnosed in HIC[6] and up to 75% in LMIC.[7] Furthermore, although dementia age of onset is likely younger in LMIC compared to HIC,[2] most studied populations in LMIC are generally older than 60 or 65.[5]

To address the above challenges, we propose a new approach to estimate dementia prevalence based on our unsupervised machine learning classification among international population-based surveys of aging.[8] These surveys assess large representative samples of individuals of age 50 or older from multiple sites within the corresponding country and collect data related to cognition, neuropsychiatric symptoms (NPS), functional status, and health through face-to-face interviews at home. For each assessed individual, a personal weight is also available, allowing us to project from the sample cohort to the whole country. Because population-based surveys of aging do not report the diagnosis of dementia, with the exception of the Health and Retirement Study (HRS[9]) and its dementia substudy the Aging, Demographics, and Memory Study (ADAMS[10]), we have proposed a method to identify individuals with high likelihood of dementia.[8] The data used for this classification include approximately 100 measures assessing demographics, health, lifestyle, mobility, cognition, NPS, and functional status in primary respondents, helped if necessary by familial informant proxies. Three clusters of participants are created with this method: "Normal" (healthy aging), "Mobility Impairment" (without functional impairment), and "Likely Dementia" (with functional impairment). We previously have demonstrated this method identifies high likelihood of dementia compared to the clinical diagnosis of dementia derived from ADAMS and is flexible enough to be applied to different population surveys, such as the Survey of Health Ageing and Retirement in Europe (SHARE[11]).[8] Importantly, this identification mostly relies on functional assessment and yields similar results when cognitive measures are removed from the analysis suggesting it is not biased by cultural background or education, a critical bias in cross-country assessment of dementia prevalence.[8]

Here, we apply this method to 10 population-based surveys covering 27 countries from North, Central, and South America; Sub-Saharan Africa; Western, Central, and Eastern Europe; and East, South-East, and South Asia. These studied countries represent 64.5% of the worldwide population over age 50 and 68.3% of the LMIC population over age 50 (after exclusion of U.S. and European cohorts). Because we expected dementia prevalence to be underestimated in LMIC, we sought to estimate the dementia burden worldwide using an innovative approach to better inform public health policy makers and trigger new interventions.

2 | METHODS

2.1 | Population surveys of aging

We selected data from 10 nationally representative longitudinal surveys, linked to the HRS family of studies. These surveys study representative samples of aging populations in each country. Participants from these surveys are carefully selected following socio-demographic methods in the different regions of the country to avoid selection biases and to provide the best nationwide estimations. The Gateway to Global Aging platform[12] provided data for China (China Health and Retirement Longitudinal Study, CHARLS,[13] wave 4, 2015), India (Longitudinal Aging Study in India, LASI,[14] wave 0, 2016), Mexico (Mexican Health and Aging Study, MHAS,[15] wave 4, 2015), the United States (HRS,[9] wave 12,2014), and Europe (SHARE,[11] wave 6, 2015). The SHARE data set included participants from several European countries (Austria, Belgium, Croatia, Czech Republic, Denmark, Estonia, France, Germany, Greece, Italy, Luxembourg, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland) and Israel. In addition, we used data from Brazil (Brazilian Longitudinal Study of Aging, ELSI,[16] wave 1, 2016), Indonesia (Indonesia Family Life Survey, IFLS,[17] wave 5, 2014–2015), Ghana (Study on global AGEing and adult health, SAGE[18]-Ghana, 2007),

South Africa (SAGE[18]-South-Africa, 2007), and Russia (SAGE[18]-Russia, 2007–2010). All primary respondents aged 50 and over (total 146,694 participants) from these 27 countries were included in this study.

We included variables about demographics (including age, sex, education, working status, family structure), health (comorbidities), functional status (activities of daily living such as dressing, eating, cooking, handling money), mobility (eg, walking, climbing), cognition (including orientation, immediate and delayed word recall), and NPS (depression and anxiety) in the different cohorts (see Table S1 in supporting information for an overview of the measures used in each cohort). The selected variables were similar to those selected in a preceding study in HRS and SHARE cohorts.[8] Variables for which >33% of the data was missing were discarded. The remaining missing values were imputed with the regularized iterative principal component analysis (PCA) algorithm.[8,19]

2.2 | Unsupervised machine learning classification

We ran a PCA in each country study separately and then applied an agglomerative hierarchical clustering on the 10 first principal components resulting from each PCA.[8,20] This data-driven and automated method allows the classification of participants into different clusters, so that participants of the same cluster share similar PCA characteristics. In each cohort, three clusters were created following the method in our previous study.[8] See also Text S1 in supporting information for additional details about methods.

2.3 | Sensitivity analysis

Cognitive measures used in the surveys may be culturally sensitive, influenced by education, and often missing in participants with cognitive impairment.[8] For each country, we repeated the unsupervised machine learning classification after removing the cognitive measures from each data set. We then compared the outcomes classifications with and without cognitive measures in each cohort. This sensitivity analysis was first intended to test the impact of cultural and educational biases on cognitive assessment and to address the potential biases of the imputation of missing values. This analysis was also expected to test the scalability of this unsupervised machine learning classification, that is, whether it could be applied to population data sets lacking cognitive measures and nevertheless provide a fair estimation of dementia prevalence.

2.4 | Estimation of dementia prevalence

We used the available personal weight for each participant in each cohort to adjust the results from the study sample to better fit the actual distribution of the whole country population.[21] To estimate the number of persons with dementia in each country, we applied the estimated percentage of dementia prevalence to the number of individuals above age 50 in the corresponding country and year.[22] To allow cross-country comparisons despite different population distributions and life expectancies, we standardized our estimations of dementia prevalence by applying age- and sex- specific prevalence for each country to a single reference country distribution, the U.S. population in 2014. Given the 10 cohorts we analyzed are representative of 64.5% of the worldwide population over age 50, we extrapolated the dementia prevalence estimated in those studies to approximate worldwide dementia prevalence.

RESEARCH IN CONTEXT

1. **Systematic review**: Global dementia prevalence reached 46.8 million in 2015 according to the World Alzheimer Report, while the Global Burden of Diseases group provided an estimation of 43.8 million in 2016. Yet, both groups acknowledged limitations in their studies, because of lack of data from certain areas, variations in dementia diagnosis tools, and underdiagnosis of dementia in medical claims.
2. **Interpretation**: We identified participants over age 50 with high likelihood of dementia based on similar measures and similar unsupervised machine learning classification in 10 population surveys of aging from 27 countries. We found that global dementia prevalence has been underestimated and was close to 130 million in 2015.
3. **Future directions**: Further studies must confirm our provocative estimation of global prevalence of dementia and test: (1) whether dementia age of onset is really younger in low- and middle-income countries than in high-income countries and (2) which risk factors of dementia should be targeted for prevention.

3 | RESULTS

Variable domains influencing the five first components of the PCA of each cohort are described in Table S2 in supporting information. In every cohort, the first component represents mobility, ADL, and IADL measures which are critical for dementia assessment.

3.1 | Two detailed examples: the CHARLS and ELSI cohorts

To demonstrate our results in detail, we chose two cohorts, CHARLS and ELSI, from two LMIC, China and Brazil, in two continents. Similar approaches were conducted on each cohort. For all cohorts, three clusters of individuals were identified. Details about demographics, clinical profile, and associated comorbidities for the clusters in

TABLE 1 Demographics and profile in machine learning clusters for dementia likelihood created in CHARLS (China) and ELSI (Brazil) cohorts

	CHARLS (China)			ELSI (Brazil)		
	Normal	Mobility Impairment	Likely Dementia	Normal	Mobility Impairment	Likely Dementia
N	8904	6044	1597	5767	3158	487
Demographics Mean (SD) - N (%)						
Age (years)	60.0 (7.7)	65.2 (9.3)	69.0 (10.2)	61.3 (8.7)	66.1 (10.5)	73.7 (13.2)
Male sex	65%	28%	39%	53.0%	27.0%	38.6%
Education (years)	7.1 (3.9)	2.9 (3.4)	3.2 (3.8)	6.8 (4.84)	3.8 (3.8)	2.4 (3.2)
Married	94.6%	72.2%	66.6%	65.1%	48.2%	34.5%
Working	80.0%	59.2%	32.1%	48.9%	14.8%	2.9%
Clinical scales mean (SD)						
Delayed recall (0-10)	3.3 (1.9)	1.7 (1.8)	1.3 (1.7)	3.2 (1.8)	2.2 (1.7)	1.4 (1.6)
ADL[a]	0.0 (0.2)	0.2 (0.4)	2.4 (1.2)	5.1 (0.3)	5.8 (1.3)	12.3 (4.4)
IADL[b]	0.1 (0.3)	0.6 (0.9)	2.7 (1.5)	5.4 (1.0)	7.4 (2.8)	15.2 (4.1)
Mobility[c]	0.5 (0.8)	2.1 (1.5)	4.8 (1.4)	4.9 (1.6)	9.2 (3.4)	14.2 (2.6)
Depression scale[d]	5.1 (4.1)	11.2 (6.3)	15.1 (6.8)	1.9 (1.9)	4.5 (2.4)	4.5 (2.2)
Comorbidities[e] N (%)						
Stroke	173 (1.9)	264 (4.4)	257 (16.1)	127 (2.2)	266 (8.4)	143 (29.4)
Diabetes	709 (8.0)	751 (12.4)	280 (17.5)	756 (13.1)	644 (20.4)	125 (25.7)
Heart disease	1245 (13.4)	1420 (20.5)	486 (30.6)	194 (3.4)	313 (9.9)	49 (10.1)
Hypertension	2547 (28.6)	2554 (42.3)	953 (59.7)	2658 (46.1)	2054 (65.0)	306 (62.8)
Dyslipidemia	1482 (15.9)	1124 (18.6)	412 (25.8)	1607 (27.9)	1162 (36.8)	124 (25.5)
Lung disease	955 (10.7)	1212 (20.1)	382 (23.9)	226 (3.9)	281 (8.9)	56 (11.5)
Cancer	110 (1.2)	146 (2.4)	51 (3.2)	253 (4.4)	218 (6.9)	33(6.8)
Depression	85 (1.0)	228 (3.8)	92 (5.8)	597 (10.4)	991 (31.4)	128 (26.3)
Arthritis	2912 (32.7)	3710 (61.4)	1028 (64.4)	816 (14.1)	1091 (34.5)	129 (26.5)

Notes: In both cohorts, participants assigned to cluster 3 (Likely Dementia) show low memory performance, and both functional and mobility impairment.
[a] ADL = Activities of Daily Living on a 0-5 scale in CHARLS and 5-20 scale in ELSI.
[b] IADL = Instrumental Activities of Daily Living on a 0-5 scale in CHARLS and 5-20 scale in ELSI.
[c] Mobility range 0-7 in CHARLS, and 4-16 in ELSI.
[d] CESD (Center for Epidemiologic Studies-Depression) scale on a 0-30 in CHARLS and 0-10 in ELSI.
[e] Self-declared comorbidities: the participant was asked to answer the question: "Has your doctor ever told you suffered from…?"
Abbreviations: CHARLS, China Health and Retirement Longitudinal Study; ELSI, Brazilian Longitudinal Study of Aging; SD, standard deviation.

CHARLS and ELSI are described in Table 1. Other cohorts' clusters are detailed in Table S3 in supporting information. The smaller cluster in each cohort corresponds to participants with high likelihood of dementia (labelled "Likely Dementia"): they are older, have lower education, lower memory performance, more functional impairment, physical impairment, and higher rate of comorbidities compared to participants in other clusters. Participants in an intermediate cluster show significant mobility impairment but no clear functional impairment, suggesting they are free from dementia (labelled "Mobility Impairment"). Participants in the largest cluster show neither physical nor functional impairment (labelled "Normal"). For each cohort, omitting cognitive measures had a minor impact on classifications, with concordant classification of individuals in the cluster "Likely Dementia" >98%.

3.2 | Conditions associated with dementia in CHARLS and ELSI cohorts

Dementia is classically associated with a series of risk factors and conditions.[23,24] We built two logistic regression models in CHARLS and ELSI cohorts to assess other conditions associated with the risk of being classified into the cluster "Likely Dementia" (Table 2). We found that older age (Figure 1), lower education, diabetes, stroke, and depression were associated with higher risk of dementia in both cohorts. In China, male sex, high blood pressure, and heart disease were additional risk factors of dementia, while drinking alcohol daily and current smoking seemed protective factors. Yet, the latter two factors, drinking alcohol and smoking, should be examined with caution given the cross-sectional nature of this study and the possibility of a survival bias

TABLE 2 Conditions associated with the risk of dementia in China and in Brazil

	China (CHARLS)		Brazil (ELSI)	
	Odds ratio (CI 95%)	*P*-value	Odds ratio (CI 95%)	*P*-value
Age (years)	1.07 (1.06 – 1.08)	<.001	1.08 (1.07 – 1.09)	<.001
Sex (male)	1.33 (1.15 – 1.55)	<.001	1.07 (0.86 – 1.33)	.6
Education (years)	0.94 (0.92 – 0.95)	<.001	0.88 (0.85 – 0.91)	<.001
Hypertension	1.56 (1.36 – 1.79)	<.001	0.86 (0.69 – 1.08)	.2
Diabetes	1.46 (1.23 – 1.74)	<.001	1.74 (1.37 – 2.23)	<.001
Dyslipidemia	1.06 (0.90 – 1.24)	.5	0.72 (0.57 – 0.92)	.009
Body mass index (kg/m^2)	1.02 (1.00 - 1.03)	.068	0.91 (0.89 – 0.94)	<.001
Stroke	5.22 (4.29 – 6.35)	<.001	5.97 (4.61 – 7.74)	<.001
Heart disease	1.17 (1.02 – 1.35)	.029	1.07 (0.75 – 1.53)	.7
Drinking alcohol daily	0.67 (0.53 – 0.84)	<.001	0.36 (0.13 – 1.02)	.054
Current smoking	0.64 (0.54 – 0.77)	<.001	0.77 (0.57 - 1.04)	.087
Depression (CESD scale)	1.16 (1.15 – 1.17)	<.001	1.37 (1.31 – 1.44)	<.001

Abbreviations: CESD, Center for Epidemiologic Studies-Depression scale; CHARLS, China Health and Retirement Longitudinal Study; CI, confidence interval; ELSI, Brazilian Longitudinal Study of Aging.

that could be demonstrated with longitudinal studies. In Brazil, dyslipidemia and higher body mass index might be protective factors against dementia. Further longitudinal studies are, however, needed to establish the possible causality links between the above factors and the risk of dementia.

3.3 | Estimation of dementia prevalence for each country

Estimation of dementia prevalence for persons over age 50 in each cohort was computed based on our classifications with cognitive measures and on personal weights available in data sets (Table 3). The total estimated number of persons with dementia in the 27 countries assessed in this study reached 82.2 million, with a mean prevalence of 7.8%. If we extrapolate this prevalence to the whole world population over age 50,[22] the estimated number of persons with dementia in 2015 was 127.8 million. To better compare dementia prevalence between countries, we computed standardized dementia prevalence by applying each country-specific prevalence to the U.S. population distribution used as a reference. High standardized dementia prevalence was observed in LMIC such as China, India, Russia, Ghana, and South Africa (Table 3). In all countries, dementia affected more women than men. In the 27 assessed countries, we estimated 16.7 million persons between the age of 50 and 60 were living with dementia, a 3.5% prevalence, which would correspond to >25 million persons worldwide in 2015.

4 | DISCUSSION

In this study, we approximated dementia prevalence in 27 countries using unsupervised machine learning and data from 10 recent population-based cohorts. According to this approach, the total number of persons with dementia in the 27 studied countries amounted to 82.2 million in 2015. Extrapolating from this, the worldwide dementia prevalence in 2015 was approaching 130 million. Compared to previous studies of global dementia prevalence, our estimations of numbers of persons with dementia are much higher, especially in China, India, Russia, Ghana, and South Africa.

4.1 | Dementia prevalence in HIC versus LMIC and comparisons to prior estimates

Our high estimation dementia prevalence is mostly driven by the high numbers we found in LMIC. According to our method, dementia prevalence in China would reach 40.2 million while the Global Burden of Disease (GBD) group reports 10.4 million.[3] In India, our estimation reaches 18.0 million while the GBD group estimates 2.9 million.[3] However, our estimations are fairly similar to previous ones for HIC: in the United States, our estimation is 4.4 million compared to 4.0 million for the GBD group.[3] Similarly, in France, we estimate that 859,000 persons would have dementia while the GBD group proposes a prevalence of 870,000.[3] Our hypothesis is that dementia prevalence might have been severely underestimated in LMIC because of obstacles for dementia prevalence assessment including: scarcity of studies about dementia in LMIC, insensitive case ascertainment, studies of nongeneralizable cohorts combined with unawareness of dementia in many areas.[5] We suggest that because our method[8] is based on measures gathered through culturally adapted and language-specific interviews in representative samples of aging populations and combined with the power of data-driven analysis through unsupervised machine learning, many of these obstacles have been overcome. The relatively similar estimates for HIC from our approach with prior approaches further

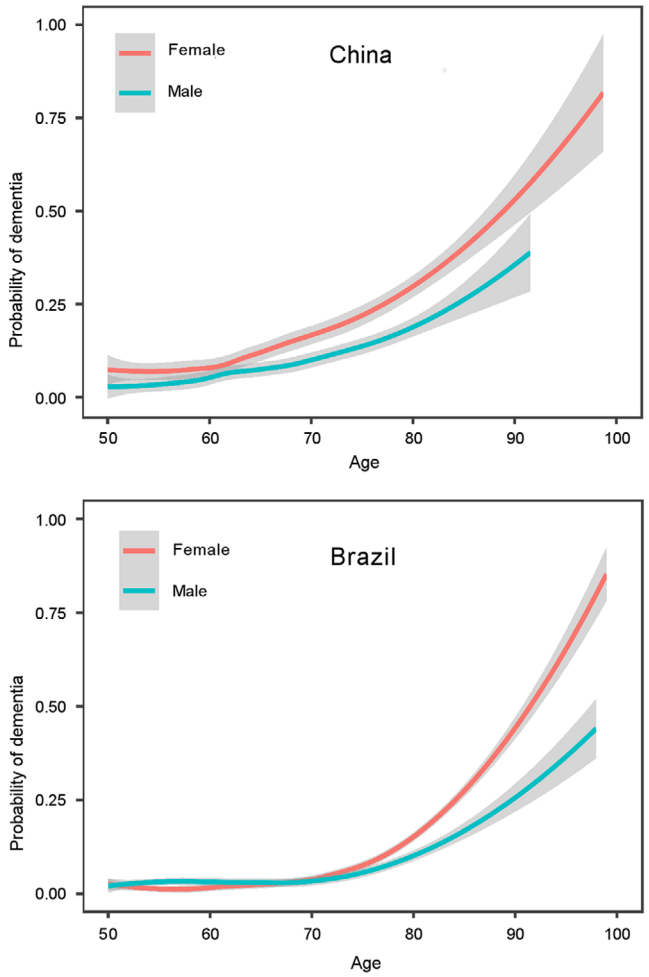

FIGURE 1 Probability of dementia according to age and sex in China (2015), panel (A) and Brazil (2016), panel (B). Shading indicates 95% confidence intervals

suggests that our method is reliable and may be more optimal for LMIC.

4.2 | Younger onset of dementia in LMIC

Part of our important estimation of dementia prevalence includes studied participants older than age 50, while previous studies have focused on persons older than age 60 or 65.[2,3,5] This may be especially important in LMIC in which life expectancy is rising but environmental stressors, education, and health conditions exert a toll in such a way that dementia incidence likely occurs at a younger age than in HIC.[2] According to our estimation, 16.7 million persons were living with dementia before the age of 60 in the 27 studied countries, with a dementia prevalence more than 5% in people between 50 and 60 in China, South Africa, and Ghana. Although dementia is a condition associated with older age, the young onset of dementia we observe in many LMIC might reflect health disparities between LMIC and HIC. This result was anticipated in the World Alzheimer Report in 2015: "an indicator of successful dementia risk reduction is deferral of dementia incidence to older ages."[2] Efficient dementia prevention policies would yield "the compression of cognitive morbidity"[25] with an older age of onset of dementia and a shorter disease duration.

4.3 | Factors associated with dementia

We found that individuals with high likelihood of dementia not only show clinical features observed in dementia such as memory loss, functional disability, and mobility difficulties, but also exhibit conditions or comorbidities often associated with dementia risk such as stroke or diabetes (Table 1). Older age remains the most important risk factor associated with dementia whatever the country (see China and Brazil, Figure 1, Table 2). Given that life expectancy is rising faster in LMIC compared to HIC, a steep increase in dementia prevalence is expected in LMIC in the next future.[2,22] Several other factors were associated with dementia risk in both China and Brazil, such as lower education, diabetes, stroke, or depression. Some of these factors are potentially modifiable through social and health policies and could be the basis of ambitious programs of dementia prevention.[24] Because other factors are not associated with dementia risk similarly in both China and Brazil, prevention strategies could benefit from a customization according to the country.[26]

4.4 | Limitations

Although we acknowledge unsupervised machine learning cannot provide a definite diagnosis of dementia, the individuals classified within the Likely Dementia cluster show the clinical characteristics typical of dementia (Table 1) and the estimation of dementia prevalence is close to previous estimates in HIC, further suggesting the accuracy of our method. However, clinical assessment of dementia in a sample of participants of these cohorts would be useful to validate our estimations. Another possible limitation of our approach is the cognitive assessment included in the datasets because it might not be appropriate for testing persons from different cultures and different educational background, especially in LMIC. Yet, our classification method does not solely rely on cognitive function; removing cognitive measures from the datasets before unsupervised classification does not greatly influence the outcomes.[8] Although cognitive measurements are required for a proper clinical assessment of dementia, we suggest that these measures are not necessary at a population level when provided with sufficient information about functional abilities and demographics. In this sense, our method could be very scalable and be used to identify dementia using electronic health records or other datasets without cognitive assessments. The imputation technique and the weighting methods might have introduced slight uncertainties in the results. Another limitation could be that data for SAGE cohorts were older than 2015. Yet, given the global aging tendency and its well-known impact on increasing dementia prevalence, this might only yield an underestimation of 2015 dementia prevalence in these

TABLE 3 Estimated dementia prevalence and numbers of persons with dementia

Countries	Year	Dementia prevalence %	N thousand persons	% Male	Standardized dementia prevalence
China	2015	10.4	40,227	34.4	15.5
India	2016	7.5	17,956	30.4	13.7
Russia	2007	11.6	5228	22.4	14.9
United States	2014	4.0	4384	38.1	4.0
Mexico	2015	7.1	1626	34.3	8.5
Brazil	2016	4.7	2211	42.3	8.0
Germany	2015	6.3	2208	34.1	5.6
Italy	2015	6.7	1738	25.2	5.1
Indonesia	2015	2.7	1287	39.3	5.2
Spain	2015	6.0	1010	31.2	4.6
South Africa	2007	14.3	1004	33.3	19.2
France	2015	3.5	859	32.4	2.6
Poland	2015	4.8	683	36.3	4.9
Portugal	2015	9.3	393	34.1	7.7
Ghana	2010	13.4	319	43.9	19.2
Belgium	2015	4.8	205	32.1	4.0
Greece	2015	4.5	201	37.8	3.8
Czech Republic	2015	4.9	194	33.0	4.7
Austria	2015	4.3	145	30.7	3.7
Israel	2015	6.7	136	37.3	6.4
Sweden	2015	3.2	117	31.7	2.8
Switzerland	2015	3.2	99	33.4	2.8
Denmark	2015	4.2	89	44.0	3.9
Croatia	2015	5.2	89	27.3	4.6
Slovenia	2015	4.7	39	40.1	4.4
Estonia	2015	6.3	32	31.6	5.6
Luxembourg	2015	5.0	9	37.9	4.9

Notes: Estimates are for persons over the age of 50 in the 27 assessed countries. To allow cross-country comparisons, we provide standardized dementia prevalence using the U.S. population distribution in 2014 as a reference.

countries. Also, despite the high quality of the datasets used in this study, some cohorts such as that from India has small sample size, which could affect the precision of our estimation of dementia prevalence in such a populated country. Similarly, extrapolating globally our results from the 27 studied countries representing 64.5% of the worldwide population remains an approximation. Finally, we acknowledge some circularity in the logistic regression models because the same variables (including those about comorbidities) were used for machine-learning classification and for the latter models. However, the primary goal of this research which was to identify participants with dementia using any relevant measures in datasets and provide the best estimation of dementia prevalence. We show in Table S2 that variables related to comorbidities are not the most influent on the first components of the PCA. Consequently, the results of the logistic regression models are not modified highly when clustering is achieved without variables about comorbidities (data not shown).

4.5 | Conclusions and perspectives

Our unsupervised machine learning classification method provides provocative results of dementia prevalence worldwide. Although our identification of high likelihood of dementia cannot be considered as an equivalent of clinical diagnosis of dementia, we suggest that this approach is less biased, truly population based, and more capable of comparing dementia prevalence between countries all over the world than previous dementia prevalence studies. Health policy makers should acknowledge the threat and face the challenge of dementia worldwide.

ACKNOWLEDGMENTS

Laurent Cleret de Langavant and Eléonore Bayen are senior Atlantic Fellows for Equity in Brain Health at the Global Brain Health Institute (GBHI). Laurent Cleret de Langavant is supported with funding from GBHI and the Alzheimer's Association (GBHI_ALZ-18-544248). Eléonore Bayen is supported with funding from GBHI and the Alzheimer's Association. Laurent Cleret de Langavant and Anne-Catherine Bachoud-Lévi are supported by the ANR-17-EURE-0017. Kristine Yaffe is supported by the NIA K24 AG031155. The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. The SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857), and FP7 (SHARE-PREP: N°211909, SHARE-LEAP: N°227822, SHARE M4: N°261982). The MHAS (Mexican Health and Aging Study) is partially sponsored by the National Institutes of Health/National Institute on Aging (grant number NIH R01AG018016) and the INEGI in Mexico. ELSI-Brazil is coordinated by the Oswaldo Cruz Foundation–Minas Gerais (FIOCRUZ-MG) and the Federal University of Minas Gerais (UFMG). The baseline survey was funded by the Brazilian Ministry of Health and the Ministry of Science, Technology, Innovation and Communication. CHARLS is supported by Peking University, the National Natural Science Foundation of China, the National Institute on Aging, and the World Bank. The LASI project is supported by an R01 grant from the National Institute on Aging (NIA), one of the 27 institutes and centers of the National Institutes of Health (NIH) and an equivalent grant from the Government of India. The Study on Global Ageing and Adult Health (Russia, Ghana, South Africa) was supported by the US National Institute on Aging with financial support through Interagency Agreements (OGHA 04034785; YA1323-08-CN-0020; Y1-AG-1005-01) and Grants (R01-AG034479; IR21-AG034263-0182). The Study on Global Ageing and Adult Health-South Africa was also supported by the South African National Department of Health (NDOH). Funding for IFLS5 was provided by the National Institute on Aging (NIA), grant 2R01 AG026676-05, the National Institute for Child Health and Human Development (NICHD), grant 2R01 HD050764-05A1 and grants from the World Bank, Indonesia and GRM International, Australia from DFAT, the Department of Foreign Affairs and Trade, Government of Australia. The Gateway to Global Aging Data is funded by the National Institute of Aging (R01 AG030153, RC2 AG036691, R03 AG043052) and developed and maintained by the Program on Global Aging, Health, and Policy, the USC Dornsife Center for Economic and Social Research. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The first and corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Laurent Cleret de Langavant performed the literature search, analyzed the data, created the figures, and wrote the first draft of the manuscript. All authors interpreted the results, edited the first draft of the manuscript, and approved the final version of the manuscript.

REFERENCES

1. Diagnostic and Statistical Manual of Mental Disorders | DSM Library. Available at: http://dsm.psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596 Accessed July 14, 2017.
2. Prince MJ. World Alzheimer Report 2015: The Global Impact of Dementia 2015.
3. Nichols E, Szoeke CEI, Vollset SE, et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2019;18:88-106.
4. WHO | International Classification of Diseases, 11th Revision (ICD-11).
5. Llibre Rodriguez JJ, Ferri CP, Acosta D, etal. Prevalence of dementia in Latin America, India, and China: a population-based cross-sectional survey. *Lancet*. 2008;372:464-474.
6. Connolly A, Gaehl E, Martin H, Morris J, Purandare N. Underdiagnosis of dementia in primary care: variations in the observed prevalence and comparisons to the expected prevalence. *Aging Ment Health*. 2011;15:978-984.
7. Nakamura AE, Opaleye D, Tani G, Ferri CP. Dementia underdiagnosis in Brazil. *Lancet*. 2015;385:418-419.
8. Cleret de Langavant L, 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.
9. Welcome to the Health and Retirement Study. http://hrsonline.isr.umich.edu/ (accessed July 13, 2017).
10. Langa KM, Plassman BL, Wallace RB, et al. The aging, demographics, and memory study: study design and methods. *Neuroepidemiology*. 2005;25:181-191.
11. The Survey of Health, Ageing and Retirement in Europe (SHARE): Home. http://www.share-project.org/ (accessed July 13, 2017).
12. Gateway to Global Aging Data. https://g2aging.org/ (accessed July 13, 2017).
13. Home | CHARLS. http://charls.pku.edu.cn/en (accessed September 6, 2019).
14. Longitudinal Aging Study in India (LASI). Program Glob Demogr Aging Harv Univ 2014. https://www.hsph.harvard.edu/pgda/major-projects/lasi-2/ (accessed September 6, 2019).
15. MHAS. http://mhasweb.org/ (accessed September 6, 2019).
16. ELSI-Brazil [English] – Brazilian Longitudinal Study of Aging. http://elsi.cpqrr.fiocruz.br/en/ (accessed September 6, 2019).
17. RAND IFLS-5 Survey Description. https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/ifls5.html (accessed September 6, 2019).
18. WHO | WHO Study on global AGEing and adult health (SAGE). http://www.who.int/healthinfo/sage/en/ (accessed September 6, 2019).
19. Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. *J Stat Softw*. 2016). 70(1), .
20. Lê Sébastien, Josse Julie, Husson François (2008). FactoMineR: AnR-Package for Multivariate Analysis. *Journal of Statistical Software*, 25, (1), https://doi.org/10.18637/jss.v025.i01.
21. Valliant R, Dever JA, Kreuter F. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer-Verlag; 2013.
22. World Population Prospects - Population Division - United Nations. https://population.un.org/wpp/ (accessed September 6, 2019).

23. Livingston G, Sommerlad A, Orgeta V, et al. Dementia prevention, intervention, and care. *Lancet*. 2017;390(10113):2673-2734.
24. Norton S, Matthews FE, Barnes DE, Yaffe K, Brayne C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *Lancet Neurol*. 2014;13:788-794.
25. Langa KM, Larson EB, Karlawish JH, et al. Trends in the prevalence and mortality of cognitive impairment in the United States: is there evidence of a compression of cognitive morbidity?. *Alzheimers Dement J Alzheimers Assoc*. 2008;4:134-144.
26. Mukadam N, Sommerlad A, Huntley J, Livingston G. Population attributable fractions for risk factors for dementia in low-income and middle-income countries: an analysis using cross-sectional survey data. *Lancet Glob Health*. 2019;7:e596-e603.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: de Langavant LC, Bayen E, Bachoud-Lévi A-C, Yaffe K. Approximating dementia prevalence in population-based surveys of aging worldwide: An unsupervised machine learning approach. *Alzheimer's Dement*. 2020;6:e12074. https://doi.org/10.1002/trc2.12074