Identifying dementia from cognitive footprints in hospital records among Chinese older adults: a machine-learning study

Jiayi Zhou,[a] Wenlong Liu,[b] Huiquan Zhou,[c] Kui Kai Lau,[d] Gloria H. Y. Wong,[a] Wai Chi Chan,[c] Qingpeng Zhang,[b,e] Martin Knapp,[f] Ian C. K. Wong,[b,g,h] and Hao Luo[a,i,*]

[a]Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong SAR, China
[b]Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[c]Department of Psychiatry, The University of Hong Kong, Hong Kong SAR, China
[d]Department of Medicine, The University of Hong Kong, Hong Kong SAR, China
[e]Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China
[f]Care Policy and Evaluation Centre (CPEC), The London School of Economics and Political Science, London, UK
[g]Laboratory of Data Discovery for Health (D24H), Hong Kong Science and Technology Park, Sha Tin, Hong Kong SAR, China
[h]Aston Pharmacy School, Aston University, Birmingham B4 7ET, UK
[i]Department of Computer Science, The University of Hong Kong, Hong Kong SAR, China

The Lancet Regional Health - Western Pacific 2024;46: 101060

Published Online xxx https://doi.org/10.1016/j.lanwpc.2024.101060

Summary

Background By combining theory-driven and data-driven methods, this study aimed to develop dementia predictive algorithms among Chinese older adults guided by the cognitive footprint theory.

Methods Electronic medical records from the Clinical Data Analysis and Reporting System in Hong Kong were employed. We included patients with dementia diagnosed at 65+ between 2010 and 2018, and 1:1 matched dementia-free controls. We identified 51 features, comprising exposures to established modifiable factors and other factors before and after 65 years old. The performances of four machine learning models, including LASSO, Multilayer perceptron (MLP), XGBoost, and LightGBM, were compared with logistic regression models, for all patients and subgroups by age.

Findings A total of 159,920 individuals (40.5% male; mean age [SD]: 83.97 [7.38]) were included. Compared with the model included established modifiable factors only (area under the curve [AUC] 0.689, 95% CI [0.684, 0.694]), the predictive accuracy substantially improved for models with all factors (0.774, [0.770, 0.778]). Machine learning and logistic regression models performed similarly, with AUC ranged between 0.773 (0.768, 0.777) for LASSO and 0.780 (0.776, 0.784) for MLP. Antipsychotics, education, antidepressants, head injury, and stroke were identified as the most important predictors in the total sample. Age-specific models identified different important features, with cardiovascular and infectious diseases becoming prominent in older ages.

Interpretation The models showed satisfactory performances in identifying dementia. These algorithms can be used in clinical practice to assist decision making and allow timely interventions cost-effectively.

Funding The Research Grants Council of Hong Kong under the Early Career Scheme 27110519.

Keywords: Dementia; Cognitive footprints; Machine learning; Electronic medical records

Introduction

Dementia is a syndrome characterized by progressive impairment in various cognitive domains that interfere with an individual's independence and daily functioning. The negative impact of dementia extends beyond patients themselves, affecting their families and the society as a whole.[1] Given that dementia remains incurable, early screening of individuals at increased risk and timely diagnosis of dementia patients is crucial for risk management and targeted interventions. However, identifying people with dementia is challenging. A meta-analysis estimated that 61.7% of patients with

*Corresponding author. The Jockey Club Tower, The Centennial Campus, HKU, 5/F, Pokfulam, Hong Kong SAR, China.
E-mail address: haoluo@hku.hk (H. Luo).

Research in context

Evidence before this study

We searched PubMed for studies about dementia prediction model published up to June 2023, using the search terms "(dementia OR Alzheimer's Disease) AND (predict*) AND (model OR risk score)", with the language restricted to English and Chinese. Previous dementia prediction models were constructed with either a clinical-driven selection of established risk factors using traditional statistical models, or a data-driven selection of features using machine learning algorithms. Most models that achieved high performance included cognitive assessments or *apolipoprotein E [ApoE]*, which are often not available in real-world data. Existing models not including cognitive assessments showed moderate performance with AUC/C-statistic ranging from 0.55 to 0.71. Most models did not consider the timing and duration of specific exposures, which could be associated with different risks of dementia. Moreover, previous models have predominately been developed from Western populations. Effective and practical models to predict dementia diagnosis among Chinese older adults remain lacking.

Added value of this study

Using electronic health records of 159,920 individuals from public hospitals in Hong Kong, this study developed general and age-specific clinical algorithms to identify dementia. Various prediction algorithms, including traditional logistics regression and four state-of-the-art machine learning models, were experimented based on the cognitive footprint theory. Comparably satisfactory performances were observed between models, with the best AUC of 0.780 observed in the general population and 0.830 among people younger than 80. Antipsychotics, education, antidepressants, head injury, and stroke were identified as the most important predictors in the total sample. Additionally, selected important predictors varied in age-specific models with cardiovascular diseases and infectious diseases becoming prominent as age increased, implying the need for implementing age-specific models to improve prediction accuracy.

Implications of all the available evidence

The prediction algorithms developed in this study can serve as an early screening tool for identifying individuals at increased risk of dementia or already with undiagnosed dementia. These prediction algorithms can be integrated into current information system to facilitate clinical decision-making in a cost-effective manner without collecting additional information. Important predictors identified from our models may enhance the current understanding of modifiable risk factors of dementia that are potentially unique to the Chinese population, which allows the generation of novel hypotheses.

dementia have not been formally diagnosed.[2] In Chinese societies where the largest number of people living with dementia reside, the diagnostic rate is even lower with an earlier estimate of 93.1%.[3] Therefore, there is an urgent need to develop effective and practical models to predict dementia diagnosis among Chinese older adults. The model should be able to be embedded into the current health information system without costing extra resources. Additionally, important predictors identified from the models can enhance the current understanding of modifiable risk factors of dementia that are potentially unique to the Chinese population, which allows the generation of novel hypotheses.

Numerous prediction algorithms have been developed for all-cause dementia, Alzheimer's disease, and cognitive impairment. Earlier algorithms typically adopted a knowledge-driven approach and were constructed with a limited number of established risk factors (e.g., age, education, *apolipoprotein E [ApoE]*, and cardiovascular diseases) based on Cox or logistic regression models. Examples of these knowledge-driven algorithms include Cardiovascular Risk Factors, Aging, and Incidence of Dementia (CAIDE) Risk Score,[4] the Brief Dementia Screening Indicator (BDSI),[5] the Australian National University Alzheimer's Disease Risk Index (ANU-ADRI),[6] and the 'Lifestyle for Brain Health' (LIBRA) score.[7] Nevertheless, algorithms constructed following the simplicity principle ignored the complicated interactions of genetic and environmental factors across the lifespan, as well as the timing and duration of specific exposures, which limits their predictive abilities. Albeit simple, these risk scores may still be difficult to calculate since key predictors such as cognitive function, lifestyle factors (physical activity, smoking, and alcohol consumption), social network, and *ApoE* may not be available in existing data sources. Unlike conventional modeling approaches that heavily rely on parametric methods with strict assumptions, machine-learning models have the capacity to address complex interactions, non-linear and high-order effects,[8] and to mitigate multicollinearity.[9,10] More recently, researchers have experimented with machine learning models that incorporated a vast number, sometime hundreds, of candidate predictors. Following a data-driven approach, these newer models utilized as much information as possible in a research cohort or real-world database, but often without differentiating between exposures at different age periods. Although the predictive power can be high, this approach can sometimes lead to the selection of relatively nuanced predictors that are symptoms and signs of critical modifiable factors (e.g., leg fat percentage).[11]

Good evidence exists that certain risk factors contribute to increased dementia risk at different life periods.[12,13] The 2020 report of the Lancet commission summarized 12 modifiable risk factors for dementia

(including less education, hypertension, hearing impairment, smoking, obesity, depression, physical inactivity, diabetes, low social contact, excessive alcohol consumption, traumatic brain injury, and air pollution) and suggested that exposures at earlier, mid- and late-life may exert different effects on dementia.[14] This concurs with the cognitive footprint theory, which suggests that people's cognitive development is influenced by a series of events as they progress from birth to death.[15] Age-related exposures constitute a subset of cognitive footprints through the life span that may interact and cumulatively affect cognition. It is, therefore, reasonable to hypothesize that considering the timing of the cognitive footprints may provide additional information regarding the complex interplay between risk factors and dementia, and improve the predictive power of prediction models for dementia. This study aimed to utilize electronic health records to develop dementia prediction algorithms following the cognitive footprint framework, using both conventional statistical and machine learning methods. The prediction algorithm will be used to provide decision support to identify patients with dementia cost-effectively.

Methods

Data source

We extracted electronic medical records from the Clinical Data Analysis and Reporting System (CDARS), a territory-wide database managed by the Hong Kong Hospital Authority (HA). As a statutory organization, the HA provides around 80% of inpatient services for over 7.4 million Hong Kong residents. Patient data in the CDARS include basic demographic characteristics, diagnoses, treatments, procedures, laboratory test results and admission/discharge information. Data from the CDARS have been used in previous dementia-related studies and have been reported to be reliable.[16] Medical records between 2000 and 2018 were employed in this study. Data were pseudo-anonymized to protect patient privacy and no patients were contacted. The study was approved by the institutional review board of The University of Hong Kong/Hospital Authority Hong Kong West Cluster (UW 18–225).

Study population

A retrospective case–control study design was adopted in this study. We included patients who visited the accident and emergency (A&E) departments or were admitted to hospitals between 2010 and 2018. The dementia group was defined as patients who 1) received the first diagnosis of dementia of any kind between 2010 and 2018, including Alzheimer's disease, vascular dementia, Lewy body dementia or other kinds of dementia, and 2) were over 65 years at the time of diagnosis. The look-back period was set to ten years and patients with any record of dementia between 2000 and 2009 were excluded. The dementia diagnosis in CDARS was made according to International Statistical Classification of Disease and Related Health Problem (ICD) criteria of the World Health Organization and ascertained by senior specialists. Although imaging tests are ordered for the majority of patients to support the diagnosis, they are not routinely conducted. In our study, the diagnosis of dementia was determined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code of 290, 294.1, 294.2, 294.8, 331.0, 331.1, and 331.82. The date of the first diagnosis of dementia was defined as the index date. Patients who had records of A&E or hospital attendance between 2010 and 2018 but did not have any diagnosis of dementia before the end of 2018 were identified as controls. Each dementia case was 1:1 randomly matched with controls by age, sex, and the date (month or year if controls were not available) of medical attendance.

Predictors

All clinical records between January 1, 2000, and the index date (fist diagnosis of dementia) were interrogated to identify exposures to risk and protective factors of interest. According to the cognitive footprint theory, age at exposures was classified into three age periods: earlier (21–45 years), mid (45–64), and late-life (≥65) as specified in the protocol published earlier.[17] Only mid and late life were considered in this study since the observation period before the dementia diagnosis was shorter than 20 years. All factors were broken down into mid and late-life factors, based on the exposure period, except education.

The risk factors were divided into established modifiable factors and other factors. Established modifiable factors were determined based on the 2020 report of Lancet commission, including diabetes mellitus, hypertension, obesity, depression, head injuries, hearing loss and less education.[14] Smoking, physical activity and social isolation, although included in the original report, were not included in this study since they were not available in hospital records. Education in the CDARS is recorded in 5 categories: less than primary, primary, secondary, tertiary education or above, and unknown. In our study, less education was defined as people who received less than primary or primary education. Our preliminary analysis identified a high rate of missing values in education, as this information is only collected in psychiatric units. Missing value in education was therefore coded as a separate category "Unknown", and therefore, less education was categorized as yes, no, and unknown.

Other factors included in the analysis were factors derived from established risk scores (stroke[5] and ischemic heart disease[7]) as well as exploratory factors selected based on the cognitive footprint theory.[17] These exploratory factors encompassed selected vascular diseases, infectious diseases, toxicity, nutrition deficiencies

and medication. Toxicity included poisoning by drugs, medicines, and biological substances, as well as toxic effects of substances of a mainly non-medicinal source. The disease diagnoses were determined by ICD-9-CM. Supplementary Table S1 summarized the diseases and their corresponding codes. Medication history of interest included antidepressants, antipsychotics, lipid-regulating drugs, anti-hypertensive drugs and diabetes medications. Polypharmacy was measured as a medication count of five or more drugs at the same day, which considered all medical prescriptions. Medication prescription was identified by British National Formulary (BNF) chapters (Supplementary Table S2). Supplementary Table S3 compares factors derived from CAIDE, BDSI, ANU-ADRI, LIBRA, and the Lancet 2020 report with factors included in our analysis.

Statistical analysis

We compared the sample characteristics between the dementia group and control group using the Chi-squared test for categorical variables. The sample was split into two datasets: a training set and a testing set. In the training set, 70% of dementia cases and the correspondingly matched controls were randomly selected. The remaining sample was assigned to the testing set. Two reference models and four machine learning models were developed using the training set. The reference models were logistic regressions with established modifiable factors only and with all factors. Machine learning models included least absolute shrinkage and selection operator (LASSO), eXtreme Gradient Boosting (XGBoost),[18] light gradient-boosting machine (LightGBM),[11,19] and Multilayer perceptron (MLP). All candidate factors were included in machine learning models. More detailed descriptions of the selected machine learning models are available in Supplementary Text S1. Ten-fold cross-validation was performed to tune the hyperparameters.[20] The list of hyperparameters tuned in the models was listed in Supplementary Tables S4 and S5.

Model performances in the testing set were evaluated by the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity and Cohen's kappa. Because of the case–control design, sensitivity, specificity, accuracy and Cohen's kappa were calculated using a default cut-off value of 0.5. Calibration was not conducted since our cohort was matched and the case prevalence does not reflect the true population prevalence of dementia. Feature importance ranking of each machine learning model was computed and the feature selection process was elaborated in Supplementary Text S1. The best model were chosen based on AUCs and Delong tests. SHapley Additive exPlanations (SHAP) plot was adopted to visualize the contribution of the 15 most important predictors from the best performing model. A SHAP value above 0 indicates an increased risk of dementia, and a SHAP value below 0 indicates a decreased risk.

The best performing model selected for all patients was further validated for patients in different sex and age groups. We conducted subgroup analyses that stratified participants by sex and age at diagnosis (<80, 80–84, 85–89 and ≥90 years). These age cutoffs were chosen to ensure comparable numbers of individuals in each subgroup (Supplementary Fig. S1). We classified age periods of exposure using a 5-year cut-off (<65, 65–69, 70–74, 75–79, and ≥80 years) to examine whether using more detailed timing of exposures can improve the prediction accuracy and to compensate for the limitation of not including age in the prediction model. We also conducted sensitivity analyses that 1) defined low education as people who received less than primary education, and 2) defined polypharmacy as a medication count of ten or more drugs. Data processing and logistic regression analyses were conducted by R version 3.4.1, and machine learning analyses were conducted by Python version 3.9.13. Apart from educational level, there were no missing values in the analytical sample. When comparing the sample characteristics, a two-sided p-value of <0.001 was considered as statistically significant, determined by adjusting the standard value of 0.05 using Bonferroni correction to account for the 51 features being explored.

Role of the funding source

This study is supported by the Research Grants Council of Hong Kong under the Early Career Scheme. The funder did not participate in the writing of the manuscript or the decision to submit it for publication.

Results

A total of 159,920 individuals with and without dementia were included. The sample characteristics are presented in Table 1 and Supplementary Tables S6 and S7. The mean age at dementia diagnosis was 83.97 (SD = 7.38); 40.5% of participants were male. Dementia patients had a higher prevalence of comorbidities (except ischemic heart diseases, heart failure and cardiac valve diseases), and prescribing rate of medications. The missing value rate was high in education, particularly for controls (92.5% in controls vs. 62.6% in cases).

Table 2 shows the performance of different prediction algorithms. Compared with the reference model that included only established risk factors (AUC: 0.689, 95% CI: 0.684, 0.694), the predictive power substantially improved in the logistic model that included all factors (AUC: 0.774, 95% CI: 0.770, 0.778). Machine learning models with selected factors showed comparable performance with the logistic model with all factors, with AUCs ranging between 0.773 for LASSO and 0.780 for MLP and XGBoost. The MLP model was chosen as our final model due to its slightly higher value of AUC, although the Delong test did not show significant superiority (Supplementary Table S8). Subgroup analyses

stratified by sex and sensitivity analyses showed similar results to the models for overall population (Supplementary Tables S9–S12). Fig. 1 shows the SHAP values estimated from the MLP model, the model with the highest AUC, for all patients. The SHAP value measures the marginal contribution of each variable to the predicted probability of having a dementia diagnosis for each individual and this contribution is evaluated in relation to values of other variables. The most important factor that contributed to increased probability of dementia was antipsychotic prescribing in late-life, followed by antidepressants prescribing, and head injury in late-life. Other important factors included less education, and stroke, hypertension, depression, and infectious disease in late life.

Fig. 2 and Supplementary Table S13 illustrate the performance of all predictive algorithms by age at dementia diagnosis (<80, 80–84, 85–89, and ≥90 years). The XGBoost model showed the best performance for patients diagnosed at age 80–84; while the LightGBM was the best for patients diagnosed at 80 or below, 85–89 and ≥ 90 years. Besides, all models showed markedly better performance in younger age groups, with the highest AUC value of 0.830 (95% CI: 0.823, 0.837) observed in patients aged <80 at diagnosis, followed by those aged 80–84 (AUC = 0.796, 95% CI: 0.788, 0.804), 85–89 (AUC = 0.767, 95% CI: 0.759, 0.776) and ≥90 (AUC = 0.749, 95% CI: 0.740, 0.758). These age-specific models all showed slightly higher predictive power than the MLP model developed from the total sample in different age groups (Supplementary Table S14). Fig. 3 and Supplementary Figs. S2–S5 summarize the top 15 important features identified in the best performing model for individuals diagnosed at different ages. Educational level, antipsychotic and antidepressant prescribing were consistently identified as the most important predictors. Among individuals diagnosed at a younger age, stroke, neurological or psychological diseases and relevant medication use were more important in identifying dementia; while as age increased, other cardiovascular diseases than stroke and infectious disease become more prominent.

Discussion

Using electronic health records in Hong Kong, this study constructed a group of dementia risk prediction algorithms based on the cognitive footprint of medical history spanning 19 years. Both conventional and machine learning models exhibited satisfactory performance in predicting dementia diagnosis. Age-specific models further enhanced the predictive accuracy of dementia, with distinct sets of important predictors identified for each age group. Notably, when incorporating other factors such as antipsychotic, antidepressant, and diabetic drug prescribing, into the models, the predictive accuracy markedly improved in comparison to

Variables	Control	Case	p	Variables	Control	Case	p	Variables	Control	Case	p
n	79,960	79,960		BCD at midlife	20 (0.0)	23 (0.0)	0.760	Infectious disease at late-life	18,594 (23.3)	21,190 (26.5)	<0.001
Male	32,356 (40.5)	32,356 (40.5)	-	BCD at late-life	1336 (1.7)	1134 (1.4)	<0.001	Toxicity at midlife	65 (0.1)	117 (0.1)	<0.001
Age	83.97 (7.38)	83.97 (7.38)	-	Cardiac dysrhythmias at midlife	397 (0.5)	416 (0.5)	0.527	Toxicity at late-life	1280 (1.6)	1613 (2.0)	<0.001
Low education			<0.001	Cardiac dysrhythmias at late-life	13,515 (16.9)	12,647 (15.8)	<0.001	Nutrition deficiency at midlife	5 (0.0)	10 (0.0)	0.302
Yes	1189 (1.5)	5437 (6.8)		Stroke at midlife	416 (0.5)	1062 (1.3)	<0.001	Nutrition deficiency at late-life	553 (0.7)	1058 (1.3)	<0.001
No	4787 (6.0)	24,492 (30.6)		Stroke at late-life	11,076 (13.9)	16,203 (20.3)	<0.001	Hearing loss at midlife	63 (0.1)	60 (0.1)	0.857
Unknown	73,984 (92.5)	50,031 (62.6)		Heart failure at midlife	198 (0.2)	278 (0.3)	<0.001	Hearing loss at late-life	1167 (1.5)	1185 (1.5)	0.724
Diabetes at midlife	985 (1.2)	1692 (2.1)	<0.001	Heart failure at late-life	12,070 (15.1)	11,394 (14.2)	<0.001	Antidepressants at midlife	622 (0.8)	1107 (1.4)	<0.001
Diabetes at late-life	15,884 (19.9)	18,353 (23.0)	<0.001	Other CVD at midlife	227 (0.3)	232 (0.3)	0.852	Antidepressants at late-life	9814 (12.3)	23,402 (29.3)	<0.001
Hypertension at midlife	1363 (1.7)	1891 (2.4)	<0.001	Other CVD at late-life	6211 (7.8)	5969 (7.5)	0.023	Antipsychotics at midlife	300 (0.4)	858 (1.1)	<0.001
Hypertension at late-life	34,016 (42.5)	34,423 (43.1)	0.040	Dyslipidemia at midlife	498 (0.6)	615 (0.8)	<0.001	Antipsychotics at late-life	5243 (6.6)	30,666 (38.4)	<0.001
Obesity at midlife	36 (0.0)	31 (0.0)	0.625	Dyslipidemia at late-life	9496 (11.9)	9282 (11.6)	0.098	Lipid drugs at midlife	1838 (2.3)	2013 (2.5)	0.005
Obesity at late-life	423 (0.5)	320 (0.4)	<0.001	Cardiac valve disease at midlife	102 (0.1)	87 (0.1)	0.308	Lipid drugs at late-life	30,129 (37.7)	31,066 (38.9)	<0.001
Depression at midlife	221 (0.3)	444 (0.6)	<0.001	Cardiac valve disease at late-life	1251 (1.6)	920 (1.2)	<0.001	Hypertension drugs at midlife	2804 (3.5)	3250 (4.1)	<0.001
Depression at late-life	2442 (3.1)	4596 (5.7)	<0.001	CBD at midlife	257 (0.3)	628 (0.8)	<0.001	Hypertension drugs at late-life	46,494 (58.1)	47,696 (59.6)	<0.001
Head injury at midlife	250 (0.3)	476 (0.6)	<0.001	CBD at late-life	9587 (12.0)	13,137 (16.4)	<0.001	Diabetes drugs at midlife	2004 (2.5)	2939 (3.7)	<0.001
Head injury at late-life	10,497 (13.1)	17,993 (22.5)	<0.001	ET at midlife	11 (0.0)	17 (0.0)	0.345	Diabetes drugs at late-life	21,059 (26.3)	23,849 (29.8)	<0.001
IHD at midlife	728 (0.9)	697 (0.9)	0.425	ET at late-life	474 (0.6)	380 (0.5)	0.001	Polypharmacy at midlife	5551 (6.9)	5924 (7.4)	<0.001
IHD at late-life	13,868 (17.3)	12,179 (15.2)	<0.001	Infectious disease at midlife	645 (0.8)	750 (0.9)	0.005	Polypharmacy at late-life	69,645 (87.1)	72,825 (91.1)	<0.001

IHD: Ischemic heart disease; BCD: Bradycardias and conduction disease; CVD: cardiovascular disease; CBD: Cerebrovascular disease; ET: Embolism and thrombosis. Note: 1) Age was summarized using mean and standard deviation (SD), and all the other factors were summarized using counts and percentages. 2) p value was obtained using Chi-squared test for categorical variables.

Table 1: **Characteristics between dementia case group and control group.**

Model	AUC (95% CI)	Sensitivity	Specificity	Accuracy	Kappa
Logistic model using established factors	0.6889 (0.6843, 0.6936)	0.511	0.813	0.662	0.324
Logistic model using all factors	0.7742 (0.7700, 0.7784)	0.591	0.859	0.725	0.450
LASSO	0.7726 (0.7684, 0.7768)	0.575	0.870	0.723	0.445
XGBoost	0.7801 (0.7760, 0.7843)	0.610	0.840	0.725	0.450
LightGBM	0.7775 (0.7733, 0.7816)	0.612	0.841	0.727	0.453
MLP	0.7803 (0.7762, 0.7844)	0.598	0.852	0.725	0.450

AUC: area under the curve; CI: confidence interval; LASSO, least absolute shrinkage and selection operator; XGBoost, eXtreme Gradient Boosting; lightGBM, light gradient-boosting; MLP, Multilayer perceptron.

Table 2: **Performance comparison of different prediction algorithms for identifying dementia in the test sample.**

models with established risk factors only. This underscores the importance of considering medication records when utilizing real-world data for dementia prediction. The derived algorithms can be used during any medical encounter of a patient to estimate the patient's likelihood of dementia using clinical records that are available prior to the date of the counter.

The prediction accuracies of our models were slightly lower than previous machine-learning studies that utilized data from sources such as the OptumLabs Data Warehouse, UK Biobank, and the Korean Longitudinal Study of Aging.[11,21,22] This difference may be attributable to the lack of laboratory tests, cognitive assessments, neurological examinations, imaging data or whole-genome sequencing in our dataset. However, it is important to note that our study was designed to develop a practical dementia prediction algorithm that can be implemented in existing health information system to detect suspected dementia as early as possible without collecting additional information. This is different from the goal of improving diagnostic accuracy when patients are already in the process of seeking help for dementia. Based solely on easily accessible predictors that are available in health records, our models exhibited comparable or superior performance to existing studies that utilized primary care data. For example, a German study that used claims data showed a C-statistic of 0.71[23]; and a US study which used diagnosis records of comorbidities and symptoms obtained a C-statistic of 0.63 for predicting dementia among those aged ≥ 65.[24]

Age is strongly associated with the onset of dementia and was shown to be the most important predictor in existing dementia prediction models.[25,26] Limited by the case–control design where the cases and controls were already matched by age, we did not include age as a predictor in our models. To account for the potential impact of age, we carried out subgroup analysis stratified by the age of dementia diagnosis. This study design built upon the fact that individuals do not age at the same pace, which led to the distinction of chronological and biological age; and that diseases and medications may mediate the association between age and dementia, allowing them to be surrogate markers for age.[27,28] Our

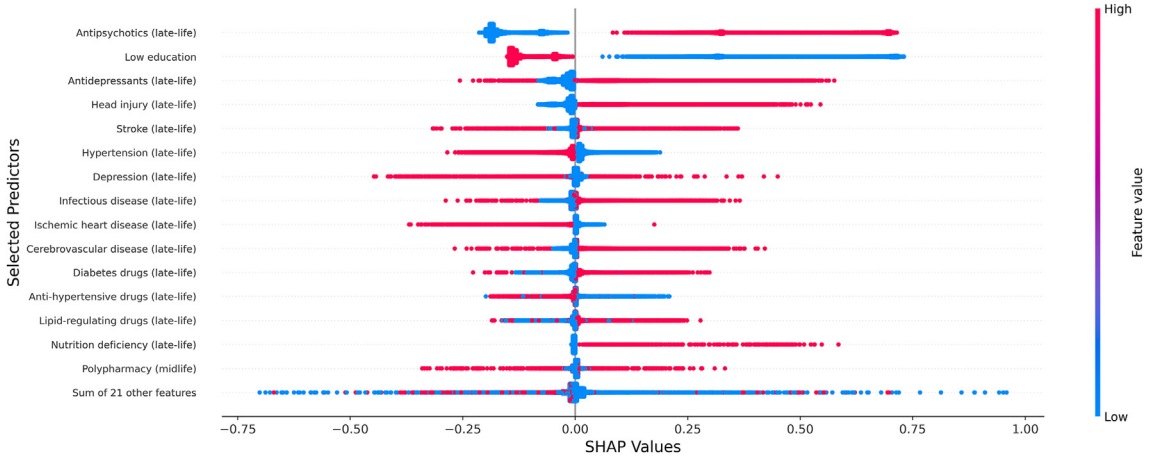

Fig. 1: **SHAP plot of the MLP prediction algorithm for dementia generated from the training sample.** The higher the SHAP values, the larger the probability of developing dementia. A dot was created for each patient and coded with gradient color representing the magnitude of predictors: red represented higher feature value, and blue represented lower feature value.

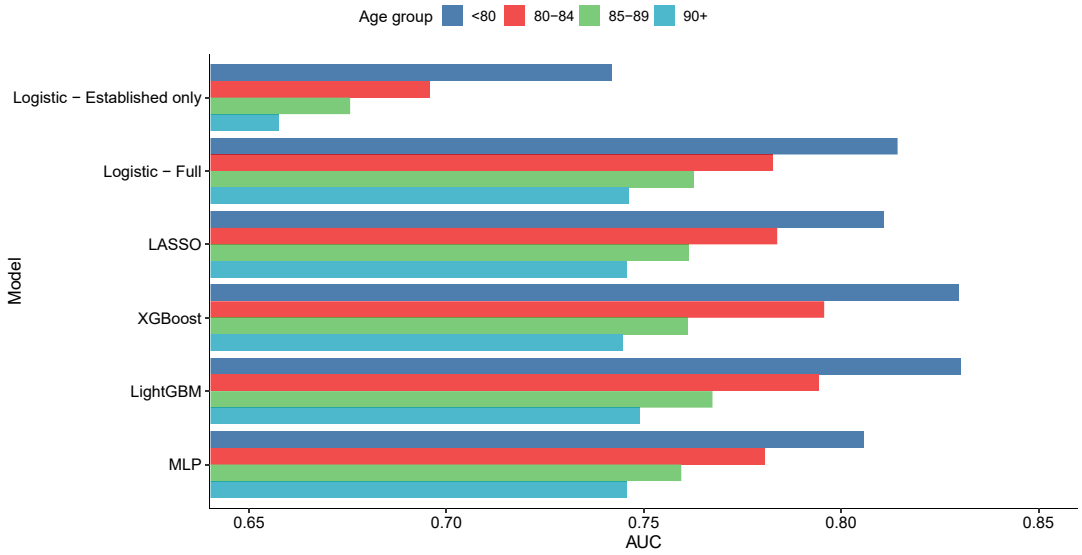

Fig. 2: **Area under the receiver operating characteristic curve (AUC) of age-specific models in the test samples.**

models showed comparable predictive power compared to other models including age.[23,29] The findings suggested that incorporating age-related predictors may provide a more nuanced understanding of dementia risk across the lifespan. Additionally, our subgroup analysis showed that the discriminatory ability decreased as the increase of age, which concurs with previous studies.[26,30] There were two possible explanations for the results. First, the very-old cohort might have missing information regarding diagnoses and medications at younger ages. Second, there are competing risks of dementia to consider. In other words, the role of predictors on dementia may diminish with age, as these factors were significantly associated with other fatal diseases and people did not survive until the diagnosis of dementia.[31] Further research might need to consider the competing risk of dementia.

Additionally, our findings revealed variations in the factors selected among the models for different age groups, which highlighted the need for age-specific prediction algorithms for dementia. For individuals at a younger age, stroke, neurological or psychological diseases and medication use were more important in predicting dementia; while as age increased, cardiovascular diseases and infectious disease become more prominent. The increased importance of cardiovascular diseases might be attributed to different subtypes of dementia, where the very-old patients may be more likely to be diagnosed with vascular dementia, given their high risk and prevalence of cardiovascular

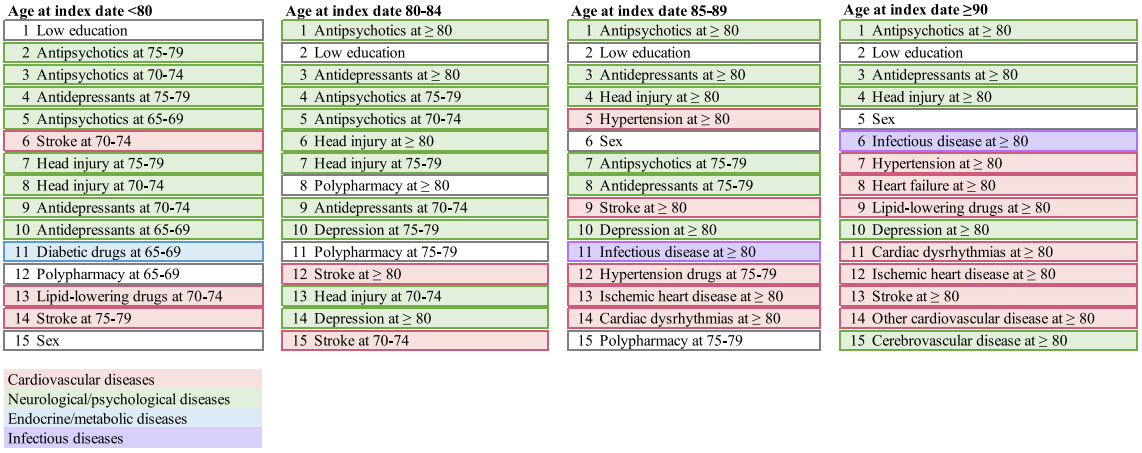

Fig. 3: **Top-15 important predictors identified in each age-specific model.**

diseases. The vascular factors have been linked to brain vascular lesions, which contribute to brain atrophy and trigger neurodegeneration process by resulting in amyloid deposition or activating an autoimmune response.[32,33] The finding underscores the vascular health in maintaining cognitive function among older people. In terms of infections, previous studies also found that the risk of dementia following infection increased with age.[34,35] This finding might be explained by immunosenescence and inflammaging, which results in an increased susceptibility of recurrent and more severe infections and in turn a greater chance of developing dementia caused by accumulated systematic inflammation or vascular damage.[36,37]

Our studies revealed that factors such as less education, depression, hypertension and head injury were significant associated with dementia diagnosis. This further verified previous findings on established modifiable factors for dementia.[14] Notably, we found hypertension at late-life was associated with a reduced risk of dementia, which contradicted the idea that hypertension is a risk factor for dementia. However, existing research remains inconsistent regarding the relationship between late-life hypertension and dementia.[38] One possible explanations of our finding might be an increased mortality in older adults who had hypertension, while case patients were not old enough to be diagnosed with dementia.[39] Depression at late-life did not show a clear pattern on the risk for dementia. The results might be attributed to the fact that depression is commonly underdiagnosed, particularly in Hong Kong where mental illness remains largely misunderstood, and may be part of the clinical profile of dementia.[40,41]

Several measures of medication history were found to significantly contribute to our dementia prediction algorithms. One of the most important features for identifying dementia was late-life antipsychotic prescribing, which may indicate prodromal symptoms of dementia, since antipsychotics are often used to alleviate behavioral and psychological symptoms associated with dementia.[42] Additionally, exposures to antidepressant, diabetic drugs and lipid-regulating drugs were found to be important, potentially serving as markers of the presence and/or the severity of diabetes, dyslipidemia and depression, both of which are established risk factors for dementia.[14] Anti-hypertensive drug contributed to decreased probability of dementia for most individuals. The results were consistent with a meta-analysis that found anti-hypertensive treatment was significantly associated with reduced dementia risk among patients with hypertensive level of blood pressure.[43] Exposure to polypharmacy at different age periods was also identified as a key feature in the total sample and three subgroups. A meta-analysis has shown that simultaneous use of multiple medications is associated with a higher risk of dementia,[44] which may be due to drug–drug interactions and additive side effects, or even drug–disease interactions where medication used to treat one disease may worsen the conditions of other diseases.[45] Researchers also noticed that patients with polypharmacy had a higher tendency for potentially inappropriate medication administration and a lower adherence to prescribed schedule, which may increase the risk for dementia.[46,47]

Late-life stroke and other cerebrovascular disease were identified as significant contributors to dementia prediction, consistent with earlier studies that found them to be main determinants of cognitive impairment.[48,49] Late-life ischemic heart disease was found to contribute to decreased probability of having dementia, which may be due to our controls being patients who had hospital or A&E attendance and who had a higher prevalence of IHD (17.3%) than dementia cases (15.2%). Lipid-regulating drug contributed to increased probability of dementia for most individuals, suggesting that the potential protective role of statin remains debatable.[50] Nutrition deficiency contributed to an increased probability of dementia. Some studies postulated that severe nutrition deficiency may be a manifestation of dementia,[51] as cognitive impairment might alter lifestyle behaviors several years prior to diagnosis of dementia.[52] In conclusion, further research is needed to better understand the impact of these factors on dementia.

Some limitations of our study should be noted. First, the case–control study design, in contrast to cohort design, may not provide an accurate reflection of the prevalence of dementia in the real-world. As age and sex were used to match the control cohort, they could no longer be included in the prediction model, despite being previously identified as important predictors of dementia. We opted for the case–control design due to practical barriers in accessing the entire CDARS database. However, the case–control design does allow for the estimation of the age-specific likelihood of having dementia at the time of a clinical encounter. Second, our study only included patients who had hospital or A&E attendance. Mild cases being underdiagnosed or attending community outpatient clinics and patients using private healthcare services may be incorrectly included in the control group. Besides, misclassification bias may exist as the diagnosis of morbidities and dementia relies on ICD-9-CM, and are subject to the clinical context of Hong Kong.[53] However, this would only lead to the underestimation of the effects of risk factors and decrease our discriminatory capacity. Third, medication history relied on medical prescription, while there is no information available regarding the actual intake of drugs. Fourth, our prediction algorithms were derived from Hong Kong older adults who used public healthcare services and had more severe medical problems which resulted in in-patient stays or A&E visits. They cannot be applied to a new patient who do not have any medical records, and the generalizability of our

findings in other populations needs to be tested. Fifth, our measure of midlife factors might be biased since for very-old patients, information regarding disease diagnosis and medication use at midlife might be missing. Therefore, further research with more robust and complete medical records in Chinese people and other populations is warranted. Sixth, although we tried to unpack the "black-boxes" of machine learning models by using SHAP plot to visualize the contribution of each feature to the likelihood of having dementia, it still does not provide information of the exact effect size as odds ratio. SHAP plot only provides locally accurate attribution values for each feature within the model.[54] Finally, it is important to note that dementia encompasses different subtypes, such as Alzheimer's disease and vascular dementia, among others, and subtype-specific models may have improved predictive accuracy. However, the subtype of dementia was not always recorded at the first diagnosis in real-world data. In our study sample, only 13.4% (N = 10,710) and 6.9% (N = 5480) of patients were coded as having Alzheimer's disease and vascular dementia, respectively, at the time of their initial diagnosis. Consequently, we considered only dementia of all kinds in this current analysis.

In conclusion, we developed prediction algorithms of dementia for overall population and by age groups based on cognitive footprint of medical history. The prediction algorithms varied by age groups, in terms of the best machine learning model and important predictors, indicating the need for tailored dementia prediction models in different subpopulations. The models may aid physicians and health service planners with risk detection of dementia in a cost-effective manner.

Contributors

J. Zhou conducted a literature review, led the data analysis and interpretation, and prepared the first draft of the manuscript. H. Luo led the study, developed the research proposal and study protocol, secured funding, and contributed to the data analysis, interpretation, and writing. W. Liu and H. Zhou contributed to data curation, analysis, and interpretation. K. K. Lau, G. Wong, W. Chan, Q. Zhang, M. Knapp, and I. Wong contributed to the study design and conceptualization and reviewed and edited the draft. H. Luo, J. Zhou, and W. Liu directly accessed and verified the underlying data reported in the manuscript. The corresponding authors confirm that all co-authors meet the authorship criteria.

Data sharing statement

The data contains confidential information and hence cannot be shared with the public due to third-party use restrictions.

Declaration of interests

ICK Wong received research funding from Amgen, Bristol-Myers Squibb, Pfizer, Janssen, Bayer, GSK, Novartis, Takeda, the Hong Kong Research Grants Council, the Hong Kong Health and Medical Research Fund, National Institute for Health Research in England, European Commission, National Health and Medical Research Council in Australia, and the European Union's Seventh Framework Programme for research and technological development, unrelated to this work. He has also received consulting fees from IQVIA, the WHO and expert testimony for Appeal Court in Hong Kong over the past three years. He is an advisory member of Pharmacy and Poisons Board, Expert Committee on Clinical Events Assessment Following COVID-19 Immunization, and the Advisory Panel on COVID-19 Vaccines of the Hong Kong Government. He is also a non-executive director of Jacobson Medical Hong Kong, founder and director of the Therakind Limited (United Kingdom), Adanced Data Analytics for Medical Science (ADAMS) Limited and OCUS Innovation Limited (Hong Kong, Ireland and United Kingdom). WC Chan received consulting fees from Eisai (Hong Kong). All other authors declare no financial or non-financial competing interests.

Acknowledgements

Nil.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.lanwpc.2024.101060.

References

1 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(1):88–106.
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(2):e011146.
3 Chen R, Hu Z, Chen RL, Ma Y, Zhang D, Wilson K. Determinants for undetected dementia and late-life depression. *Br J Psychiatry*. 2013;203(3):203–208.
4 Kivipelto M, Ngandu T, Laatikainen T, Winblad B, Soininen H, Tuomilehto J. Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *Lancet Neurol*. 2006;5(9):735–741.
5 Barnes DE, Beiser AS, Lee A, et al. Development and validation of a brief dementia screening indicator for primary care. *Alzheimers Dement*. 2014;10(6):656–665.e1.
6 Anstey KJ, Cherbuin N, Herath PM. Development of a new method for assessing global risk of Alzheimer's disease for use in population health approaches to prevention. *Prev Sci*. 2013;14(4):411–421.
7 Schiepers OJ, Köhler S, Deckers K, et al. Lifestyle for Brain Health (LIBRA): a new model for dementia prevention. *Int J Geriatr Psychiatr*. 2018;33(1):167–175.
8 Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177(5):443–452.
9 Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms. From machine learning to statistical modelling. *Methods Inf Med*. 2014;53(6):419–427.
10 Altelbany SI. Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: a simulation study. *J Appl Econ Bus Stud*. 2021;5(1):131–142.
11 You J, Zhang YR, Wang HF, et al. Development of a novel dementia risk prediction model in the general population: a large, longitudinal, population-based machine-learning study. *EClinicalMedicine*. 2022;53:101665.
12 Dekhtyar S, Wang HX, Scott K, Goodman A, Koupil I, Herlitz A. A life-course study of cognitive reserve in dementia–from childhood to old age. *Am J Geriatr Psychiatry*. 2015;23(9):885–896.
13 Whalley LJ, Dick FD, McNeill G. A life-course approach to the aetiology of late-onset dementias. *Lancet Neurol*. 2006;5(1):87–96.
14 Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*. 2020;396(10248):413–446.
15 Rossor M, Knapp M. Can we model a cognitive footprint of interventions and policies to help to meet the global challenge of dementia? *Lancet*. 2015;386(9997):1008–1010.
16 Cheung ECL, Leung MTY, Chen K, et al. Risk of adverse events and delirium after COVID-19 vaccination in patients living with dementia. *J Am Med Dir Assoc*. 2023;24(6):892–900.e12.
17 Luo H, Lau KK, Wong GHY, et al. Predicting dementia diagnosis from cognitive footprints in electronic health records: a case-control study protocol. *BMJ Open*. 2020;10(11):e043487.
18 James C, Ranson JM, Everson R, Llewellyn DJ. Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA Netw Open*. 2021;4(12):e2136553.

19 Hane CA, Nori VS, Crown WH, Sanghavi DM, Bleicher P. Predicting onset of dementia using clinical notes and machine learning: case-control study. *JMIR Med Inform*. 2020;8(6):e17819.
20 Rennie JDM, Shih L, Teevan J, Karger DR. Tackling the poor assumptions of naive bayes text classifiers. *Int Conf Mach Learn*. 2003;2003:616–623.
21 Nori VS, Hane CA, Crown WH, et al. Machine learning models to predict onset of dementia: a label learning approach. *Alzheimers Dement (N Y)*. 2019;5:918–925.
22 Na KS. Prediction of future cognitive impairment among the community elderly: a machine-learning based approach. *Sci Rep*. 2019;9(1):3335.
23 Reinke C, Doblhammer G, Schmid M, Welchowski T. Dementia risk predictions from German claims data using methods of machine learning. *Alzheimers Dement*. 2023;19(2):477–486.
24 Albrecht JS, Hanna M, Kim D, Perfetto EM. Predicting diagnosis of Alzheimer's disease and related dementias using administrative claims. *J Manag Care Spec Pharm*. 2018;24(11):1138–1145.
25 Licher S, Yilmaz P, Leening MJG, et al. External validation of four dementia prediction models for use in the general community-dwelling population: a comparative analysis from the Rotterdam Study. *Eur J Epidemiol*. 2018;33(7):645–655.
26 Licher S, Leening MJG, Yilmaz P, et al. Development and validation of a dementia risk prediction model in the general population: an analysis of three longitudinal studies. *Am J Psychiatry*. 2019;176(7):543–551.
27 Hamczyk MR, Nevado RM, Barettino A, Fuster V, Andrés V. Biological versus chronological aging: JACC focus seminar. *J Am Coll Cardiol*. 2020;75(8):919–930.
28 Belsky DW, Caspi A, Houts R, et al. Quantification of biological aging in young adults. *Proc Natl Acad Sci USA*. 2015;112(30):E4104–E4110.
29 Honda T, Ohara T, Yoshida D, et al. Development of a dementia prediction model for primary care: the Hisayama Study. *Alzheimers Dement (Amst)*. 2021;13(1):e12221.
30 Walters K, Hardoon S, Petersen I, et al. Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data. *BMC Med*. 2016;14:6.
31 Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? *Stat Med*. 2012;31(11–12):1089–1097.
32 Qiu C, Fratiglioni L. A major role for cardiovascular burden in age-related cognitive decline. *Nat Rev Cardiol*. 2015;12(5):267–277.
33 de Roos A, van der Grond J, Mitchell G, Westenberg J. Magnetic resonance imaging of cardiovascular function and the brain: is dementia a cardiovascular-driven disease? *Circulation*. 2017;135(22):2178–2195.
34 Muzambi R, Bhaskaran K, Smeeth L, Brayne C, Chaturvedi N, Warren-Gash C. Assessment of common infections and incident dementia using UK primary and secondary care data: a historical cohort study. *Lancet Healthy Longev*. 2021;2(7):e426–e435.
35 Muzambi R, Bhaskaran K, Brayne C, Davidson JA, Smeeth L, Warren-Gash C. Common bacterial infections and risk of dementia or cognitive decline: a systematic review. *J Alzheimers Dis*. 2020;76(4):1609–1626.
36 Sipilä PN, Heikkilä N, Lindbohm JV, et al. Hospital-treated infectious diseases and the risk of dementia: a large, multicohort, observational study with a replication cohort. *Lancet Infect Dis*. 2021;21(11):1557–1567.
37 Sterling K, Xing M, Song W. Do systemic infections contribute to the pathogenesis of dementia? *Neurosci Bull*. 2022;38(3):331–333.
38 Walker KA, Power MC, Gottesman RF. Defining the relationship between hypertension, cognitive decline, and dementia: a review. *Curr Hypertens Rep*. 2017;19(3):24.
39 Castilla-Guerra L. Late-life hypertension as a risk factor for cognitive decline and dementia. *Hypertens Res*. 2022;45(10):1670–1671.
40 Leung DKY, Chan WC, Spector A, Wong GHY. Prevalence of depression, anxiety, and apathy symptoms across dementia stages: a systematic review and meta-analysis. *Int J Geriatr Psychiatry*. 2021;36(9):1330–1344.
41 Chin WY, Chan KT, Lam CL, et al. Detection and management of depression in adult primary care patients in Hong Kong: a cross-sectional survey conducted by a primary care practice-based research network. *BMC Fam Pract*. 2014;15:30.
42 Gareri P, Segura-García C, Manfredi VG, et al. Use of atypical antipsychotics in the elderly: a clinical review. *Clin Interv Aging*. 2014;9:1363–1373.
43 Ding J, Davis-Plourde KL, Sedaghat S, et al. Antihypertensive medications and risk for incident dementia and Alzheimer's disease: a meta-analysis of individual participant data from prospective cohort studies. *Lancet Neurol*. 2020;19(1):61–70.
44 Leelakanok N, D'Cunha RR. Association between polypharmacy and dementia - a systematic review and metaanalysis. *Aging Ment Health*. 2019;23(8):932–941.
45 Chippa V, Roy K. Geriatric cognitive decline and polypharmacy. StatPearls. Treasure island (FL) ineligible companies. In: *Disclosure: kamalika Roy declares no relevant financial relationships with ineligible companies*. StatPearls Publishing; 2023. Copyright © 2023, StatPearls Publishing LLC.
46 Park HY, Park JW, Song HJ, Sohn HS, Kwon JW. The association between polypharmacy and dementia: a nested case-control study based on a 12-year longitudinal cohort database in South Korea. *PLoS One*. 2017;12(1):e0169463.
47 El-Saifi N, Moyle W, Jones C, Tuffaha H. Medication adherence in older patients with dementia: a systematic literature review. *J Pharm Pract*. 2018;31(3):322–334.
48 Verdelho A, Wardlaw J, Pavlovic A, et al. Cognitive impairment in patients with cerebrovascular disease: a white paper from the links between stroke ESO Dementia Committee. *Eur Stroke J*. 2021;6(1):5–17.
49 Gardener H, Wright CB, Rundek T, Sacco RL. Brain health and shared risk factors for dementia and stroke. *Nat Rev Neurol*. 2015;11(11):651–657.
50 Olmastroni E, Molari G, De Beni N, et al. Statin use and risk of dementia or Alzheimer's disease: a systematic review and meta-analysis of observational studies. *Eur J Prev Cardiol*. 2022;29(5):804–814.
51 Kimura A, Sugimoto T, Kitamori K, et al. Malnutrition is associated with behavioral and psychiatric symptoms of dementia in older women with mild cognitive impairment and early-stage Alzheimer's disease. *Nutrients*. 2019;11(8):1951.
52 Lefèvre-Arbogast S, Wagner M, Proust-Lima C, Samieri C. Nutrition and metabolic profiles in the natural history of dementia: recent insights from systems biology and life course epidemiology. *Curr Nutr Rep*. 2019;8(3):256–269.
53 Henderson JN, Traphagan JW. Cultural factors in dementia: perspectives from the anthropology of aging. *Alzheimer Dis Assoc Disord*. 2005;19(4):272–274.
54 Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. preprint arXiv:180203888 *arXiv*. 2018. https://doi.org/10.48550/arXiv.1802.03888.