ELSEVIER

Alzheimer's & Dementia: Translational Research & Clinical Interventions 5 (2019) 563-569

Alzheimer's & Dementia

Review Article

Statistical methods for dementia risk prediction and recommendations for future work: A systematic review

Jantje Goerdten[a,*], Iva Čukić[a], Samuel O. Danso[a], Isabelle Carrière[b], Graciela Muniz-Terrera[a]

[a]*Edinburgh Dementia Prevention & Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK*

[b]*INSERM, Neuropsychiatrie, Recherche Epidemiologique et Clinique, Montpellier, France*

Abstract

Introduction: Numerous dementia risk prediction models have been developed in the past decade. However, methodological limitations of the analytical tools used may hamper their ability to generate reliable dementia risk scores. We aim to review the used methodologies.

Methods: We systematically reviewed the literature from March 2014 to September 2018 for publications presenting a *dementia* risk prediction model. We critically discuss the analytical techniques used in the literature.

Results: In total 137 publications were included in the qualitative synthesis. Three techniques were identified as the most commonly used methodologies: machine learning, logistic regression, and Cox regression.

Discussion: We identified three major methodological weaknesses: (1) over-reliance on one data source, (2) poor verification of statistical assumptions of Cox and logistic regression, and (3) lack of validation. The use of larger and more diverse data sets is recommended. Assumptions should be tested thoroughly, and actions should be taken if deviations are detected.

Keywords: Dementia risk models; Methodological review; Logistic regression; Cox models; Machine learning

1. Introduction

The prevalence of dementia is increasing globally, because of the rapid aging of the population. In 2015, 47 million people were affected by dementia worldwide, whereas dementia prevalence is predicted to almost triple by 2050 [1]. There is no cure for dementia yet; hence, the early identification of individuals at higher risk of developing dementia becomes critical, as this may provide a window of opportunity to adopt lifestyle changes to reduce dementia risk [1,2].

Numerous dementia risk prediction models to identify individuals at higher risk have been developed in the past decade. Three systematic reviews and meta-analyses summarizing dementia risk prediction models were published over the past years [3–5]. Stephan et al. [3] and Tang et al. [4] mainly focused on the critique of the variables selected for inclusion and the assessment of models' prognostic performance, whereas Hou et al. [5] reviewed published dementia risk models in terms of sensitivity, specificity, and area under the curve from receiving operating characteristic analysis. Stephan et al. [3] and Tang et al. [4] concluded that none of the published models could be recommended for dementia risk prediction, largely because of multiple methodological weaknesses of the models or study designs for their derivation. Methodological limitations of the models reviewed included the lack of discrimination of dementia type, lack of internal and external validations of the models, the long interval elapsed between assessments of individuals at risk, and notably, concerns about the analytical techniques used were also highlighted. Hou et al. [5] recommended four

Conflict of Interest: The authors have declared that no conflict of interest exists.

*Corresponding author. Tel: +44 131 537 6530.

E-mail address: jantje.goerdten@ed.ac.uk

https://doi.org/10.1016/j.trci.2019.08.001

risk prediction models for different populations (midlife, late-life, patients with diabetes, or mild cognitive impairment (MCI)) with acceptable predictive ability (area under the curve ≥0.74), but still concluded that the models showed methodological limitations, such as lack of external validation.

To date, there is no systematic literature review focusing solely on the methodological approaches used in the dementia risk literature. In the present study, we aim to identify and critically discuss the analytical techniques used in the dementia risk literature and provide suggestions for future prediction model developments, to increase model reliability and accuracy.

2. Methods

2.1. Search strategy

We searched MEDLINE, Embase, Scopus, and ISI Web of Science for articles published from March 1, 2014 to September 17, 2018 using combinations of the following terms: "dementia," "prediction," "development," "receiver operating characteristic," "sensitivity," "specificity," "area under the curve," and "concordance statistic." When possible, terms were mapped to Medical Subject Headings. We searched relevant systematic literature reviews for additional references. March 1, 2014 was chosen as earliest date for this review as it is the upper limit of Tang et al.'s [4] dementia risk review (see Supplementary Material 1 for an example of the search strategy). An updated search was performed from September 17, 2018 to June 12, 2019.

2.2. Selection of studies

First, two independent reviewers (I.Č. and J.G.) screened titles and abstracts for suitable articles. Next, full-text articles were screened for eligibility by one reviewer (J.G.). The following eligibility criteria was used to select the relevant publications: (1) the study has to use a population-based sample or a sample restricted to individuals with MCI; (2) the article provided a model to predict dementia (all-type dementia) risk; (3) the article described the statistical technique that was used for the model development; (4) the article was written in English. Conference abstracts and validation studies were excluded from the review. Any disagreements were resolved by consensus between two authors (I.Č. and J.G.), or if necessary, by a third author (G.M.T.) if the disagreement could not be resolved.

2.3. Data extraction

Data were extracted by three authors (I.Č., J.G., and S.D.) from each article. Information collected included data source, sample size, country, study population, dementia type, length of follow-up, statistical technique used for model development, tested assumptions, and validation method. Only information relevant to our review was extracted from the articles, that is, when studies investigated several aims, we only reported the statistical method and sample that were used for the prediction of dementia risk. In one case a reference is counted twice in the results, as it reports risk models developed from two separate techniques (see Supplementary Material 2 for tables describing publications extracted for review).

3. Results

A total of 2600 nonduplicated articles were identified from the database search and additional relevant references. During the title and abstract review phase 2328 articles were excluded. Full texts were screened for 272 articles, of which 137 were found to be eligible for inclusion in the quantitative synthesis. The most frequent reasons for exclusion were the article was a conference abstract, no prediction model for dementia risk was provided, outcomes other than dementia risk were predicted (e.g., combination of MCI and Alzheimer's disease (AD)), and nonpopulation-based samples or samples not consisting of MCI individuals were used (e.g., sample consisting of menopausal women) (see Fig. 1, for a flow chart of the review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart diagram template was used [6]).

3.1. Outcomes and populations

Population-based samples were used in 31 (31/138, 22.5%) publications and 107 (107/138, 77.5%) publications used samples comprising MCI individuals for the development of a dementia risk prediction model. In total, 137 study populations were used for the development of the models, of which 74 are unique. In total 60 (60/138, 43.5%) samples were drawn from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The sample size of the studies reviewed ranged from 22 to 331,126 individuals, whereas 17 (17/138, 12.3%) studies had a sample size smaller than 100 participants. The follow-up time ranged from 1 to >30 years. In 103 (103/138, 74.6%) publications AD was the primary outcome. Other dementia types or a combination of dementia types with AD were regarded as the outcome in 35 (35/138, 25.4%) publications, including: dementia any type/not otherwise specified, vascular dementia, mixed dementia, frontotemporal dementia, Huntington disease, Lewy body dementia, multi-infarct type dementia, and Parkinson disease dementia. In 22 (22/138, 15.9%) publications, risk models were externally validated, whereas 46 (46/138, 33.3%) publications did not mention any validation procedure.

3.2. Analytical approaches

Machine learning (n = 55) was the most used technique for the development of dementia risk prediction models. In the publications selected for review, the support vector machine classifier (n = 17) was the most commonly used

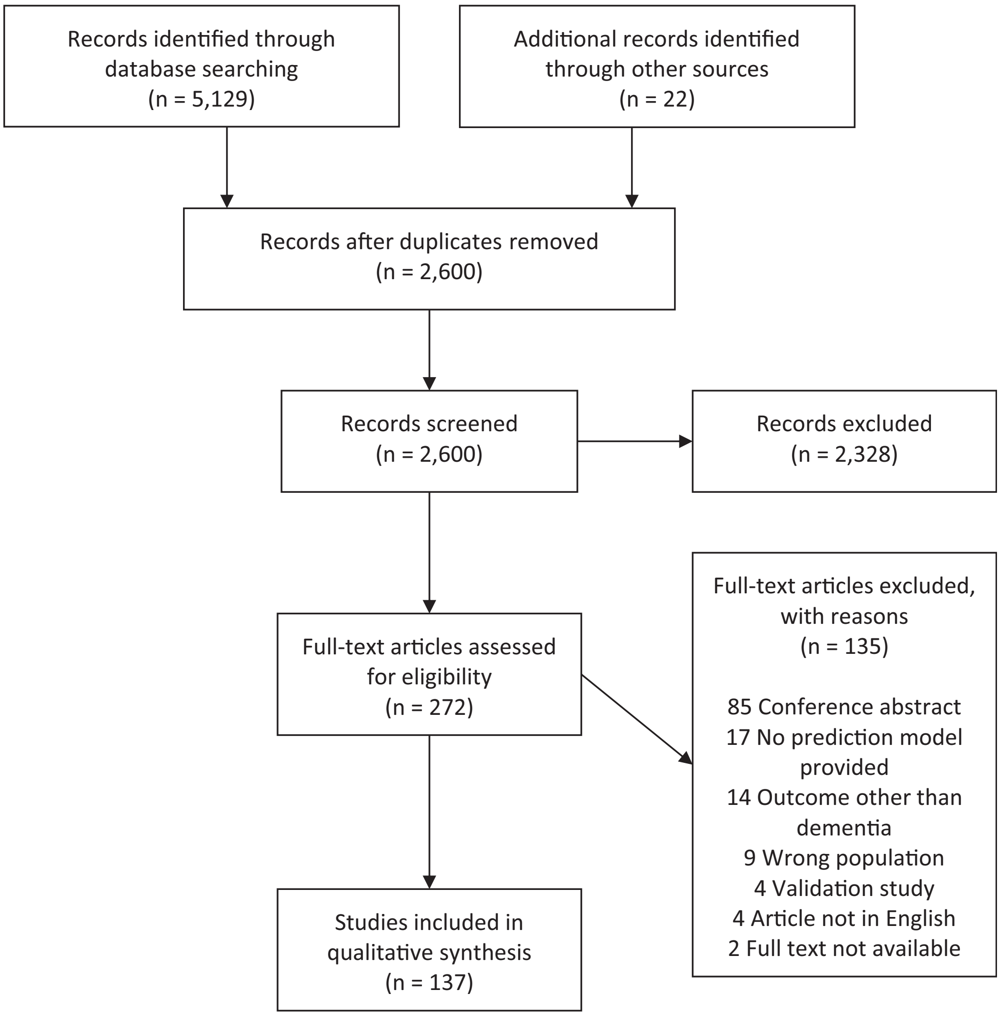

Records identified through database searching (n = 5,129)
Additional records identified through other sources (n = 22)
Records after duplicates removed (n = 2,600)
Records screened (n = 2,600)
Records excluded (n = 2,328)
Full-text articles assessed for eligibility (n = 272)
Full-text articles excluded, with reasons (n = 135)
85 Conference abstract
17 No prediction model provided
14 Outcome other than dementia
9 Wrong population
4 Validation study
4 Article not in English
2 Full text not available
Studies included in qualitative synthesis (n = 137)

Fig. 1. Flow chart of review phases.

algorithm to predict dementia, followed by the disease state index ($n = 6$) and the random forest classifier ($n = 5$). Three studies used neural network algorithms to construct risk prediction models. Several different feature selection methods are used (including least absolute shrinkage and selection operator, recursive feature elimination, or correlation-based feature selection). In 48 (48/55, 87.3%) of the studies that used machine learning algorithms, prediction models were developed for individuals with MCI, whereas seven (7/55, 12.7%) studies developed models for individuals without clinically impaired cognition. Thirty-four (34/55, 61.8%) publications used the ADNI database. Twelve (12/55, 21.8%) models are externally validated with independent samples, of which 10 were MCI populations and two population-based, whereas 49 (49/55, 89.1%) models are internally validated using different cross-validation methods (e.g., 10-fold cross-validation), two (2/55, 3.6%) models are neither externally nor internally validated.

Logistic regression was used in 31 publications for the development of dementia risk prediction models. One study fitted a multinomial logistic regression including mortality as a third outcome and another study included follow-up time in the model. Eight (8/31, 25.8%) samples are drawn from the ADNI database. Five (5/31, 16.1%) studies checked for multicollinearity among the independent variables and two studies additionally checked the linearity assumption. Two (2/31, 6.5%) studies checked if the data were normally distributed. None of the dementia risk prediction models derived from logistic regression are externally validated. Eleven (11/31, 35.5%) models are internally validated using a cross-validation method, bootstrapping, or by splitting the sample into a testing and validation set, whereas 20 (20/31, 64.5%) models were not validated.

Cox proportional hazards regression (Cox regression) models were used for the development of 25 dementia risk prediction models. Of these, one study used time-dependent covariates in the Cox model, one study included death as a competing risk, one further study used a penalized Cox regression, and another study used age as the time axis. Seven (7/25, 28%) models are developed based on data from the ADNI database. Eight (8/25, 32%) studies verified the proportional hazard assumption

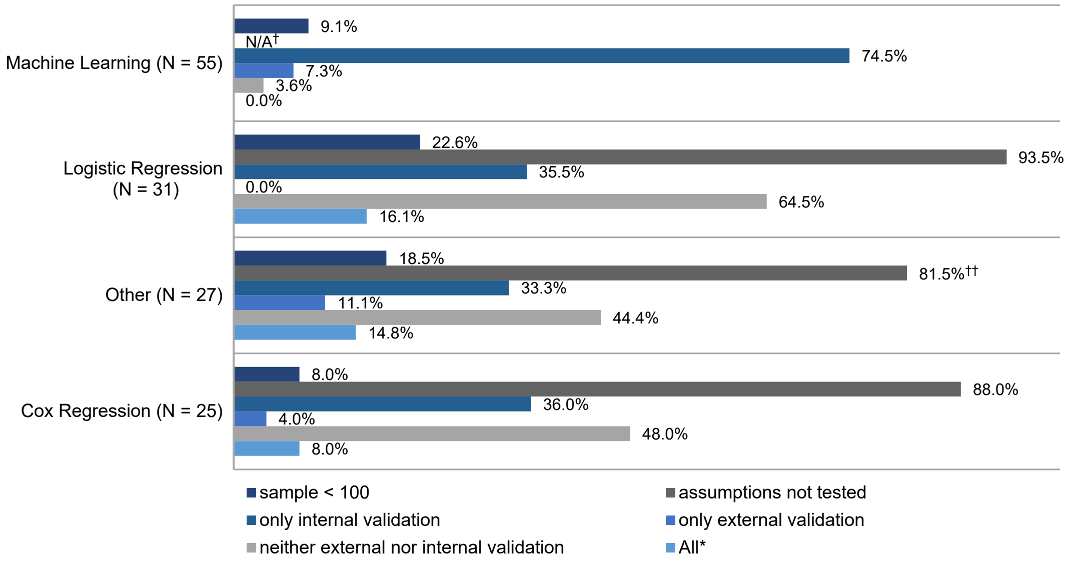


Fig. 2. Limitations of included studies stratified by technique used. *Study used a sample <100 individuals, did not test the assumptions (if applicable), and did not validate the results internally or externally. †Not applicable (N/A). ††Percentage calculated for whole group (22/27); however, only 22 studies needed to test assumptions (22/22, 100%).

and three studies additionally checked the linearity assumption. Four (4/25, 16%) risk models were externally validated with independent samples, of which one was an MCI population and three were population-based. Twelve (12/25, 48%) risk models are validated internally, using cross-validation methods or bootstrapping, whereas 12 (12/25, 48%) risk models are neither externally nor internally validated.

Five studies used a combination of a machine learning approach and a regression analysis (e.g., disease state index and Cox regression), four a combination of two regressions (e.g., logistic and Cox regression), two a joint longitudinal survival model, two an analysis of variance, two a bilinear regression, and two a receiver operating characteristic curve analysis to develop a dementia risk prediction model.

Less frequently used techniques were linear regression ($n = 1$), polynomial regression ($n = 1$), χ^2 test and Kruskal-Wallis test ($n = 1$), power of the t-sum score ($n = 1$), Poisson regression ($n = 1$), illness-death model ($n = 1$), multivariate ordinal regression ($n = 1$), event-based probabilistic model ($n = 1$), mixed linear model ($n = 1$), and general linear model ($n = 1$). An overview of the limitations found in the studies is provided in Fig. 2.

4. Discussion

Our review of analytical approaches in dementia risk prediction identified three techniques as the most commonly used methodologies: machine learning, Cox regression, and logistic regression models.

4.1. Machine learning

A growing number of dementia risk prediction models have been developed using machine learning algorithms [7,8]. Machine learning consists of computational methods, which are able to find meaningful patterns in the data [9], while using experience to improve and make predictions [10]. This means machine learning techniques can explore the structure of the data, in terms of associations between the variables, without having a theory of how the structure looks like. This might make them better suited to detect associations between variables than logistic or Cox regression [11]. However, as discussed by Pellegrini et al. [8], in a published systematic literature and meta-analyses of machine learning techniques in neuroimaging for cognitive impairment and dementia, studies using machine learning algorithms also show limitations. Generalizability of results generated from the application of these techniques and their transfer to clinical use are likely to be constrained because of their over-reliance on one data source, the fact that they commonly use data from populations with greater proportions of cases (i.e., individuals with the diseases) and lower proportions of control subjects, they are usually derived using only one machine learning method and the application of varying validation methods [7,8]. Furthermore, although machine learning methods perform well (accuracy ≥0.8) in differentiating healthy control subjects from individuals with dementia, their performance when identifying individuals at high risk of developing dementia is poorer (accuracy from 0.5 to 0.85) [8]. Similar methodological limitations were found in the present review: more than half the studies used the same data source (34/55, 61.8%), only six (7/55, 12.7%) studies investigated the prediction abilities of a machine learning method in a non-MCI population and only 12 of 55 (21.8%) studies externally validated their model. Relying mainly on one data source and focusing on individuals already at a higher risk of developing dementia results in limitations of clinical relevance and generalizability. Furthermore, without additional studies externally validating these prediction models, it is not clear if these models are overfitted and further limits the generalizability of findings.

4.2. Logistic and Cox regression

Cox and logistic regression, two traditional statistical techniques, are used frequently in dementia risk prediction [4].

Despite their popularity, several features of logistic and traditional Cox regression need to be reflected on when using these methods in dementia risk prediction (for a comparison of these approaches in general settings see Ingram and Kleinman [12] and Peduzzi et al. [13], but it is also worth remembering that although logistic regression aims at the estimation of odds ratios, Cox modeling aims to estimate hazard ratios over time). First, an aspect of both approaches that is relevant to note is that they both generate static risk predictions as they are based on a designated time 0 and on data (baseline covariates) collected at a single time point (time 0 or before). Extensions to time-dependent Cox models exist that are appropriate for use when risk factors themselves change over time [14]. Although not implemented yet in dementia risk prediction, these extended models are likely to be informative in the context of dementia risk prediction as change in predictors over time is likely to be more informative than a single value. However, if prediction is short term, models with time-dependent variables may not be necessary.

Second, although our review identified only one publication where a Cox model based on age was used, the choice of the time axis in Cox modeling is a methodological aspect that also needs consideration as different choices hamper the comparison of results across (and within) studies. When age is used as time axis, the analysis needs to be corrected for delayed entry. A discussion of this issue in the methodological literature and empirical demonstrations showing high sensitivity of results to different choices can be found in Pencina et al. [15].

Third, both methods assume a data structure that may not be adequately fulfilled when used in dementia prediction. Both techniques assume linear relationships between the independent and the dependent variables, that is, a linear relationship between the log of the odds (logistic regression) or the relative risk (Cox regression) and the covariates is assumed. Yet, this assumption is likely not to hold for critical variables used as input in the model (e.g., biomarker) [16]. Notably, our review identified only five of 56 (8.9%) studies that explicitly tested the linearity assumption. Cox regression additionally assumes proportional hazards, which postulates that the impact of a prognostic factor on dementia remains constant over the entire follow-up. Only a third (8/25, 32%) of the identified studies in our review that used Cox regression tested the proportional hazards assumption. Violations of the underlying assumptions in logistic and Cox regression result in biased estimates [16,17]. Furthermore, merely five (5/56, 8.9%) studies incorporated interactions in their regression model. Although an interaction makes it harder to interpret the estimates, it is still relevant and potentially informative to test these.

4.3. Validation, sample size, and data source

External and internal validations are crucial steps when developing a reliable prediction model. Although internal validation ensures the robustness of the findings, that is, there are no alternative explanations for the findings, external validation provides information to which extent results can be generalized, that is, the model can be applied to a wider population than the one from which it was developed [18,19]. Although, the validation phase is highly recommended in prediction models [20], a third of the studies did not perform internal or external validation (46/138, 33.3%). Of the 138 prediction models reviewed, only 14 (10.1%) studies validated their model internally and externally. Too many studies did not perform any validation, whereas too few studies performed both internal and external validations. This is a poor state and the field would benefit from a change in practice.

The data used for the development of a prediction model are as important as the technique used for the derivation of the model. Several studies (17/138, 12.3%) used a sample smaller than 100 participants, which likely is a limitation of these studies. There are no recommendations for a specific sample size, as it dependents on various factors (e.g., which technique is used, number of cases, and number of predictors). Nevertheless, the sample size should be considered when planning a study. The studies reviewed here used 74 unique study populations. The ADNI database was used frequently: 60 (60/138, 43.5%) models were derived using information from subsamples drawn from the ADNI database. Although this overlap makes the results more comparable, it renders at the same time generalizability and might inflate accuracy. For instance, predictions for individuals in ethnic minorities are unlikely to be accurate if models are derived from the ADNI database, where ethnic minorities are largely under-represented, as inferences will be based on a low number of cases and the studies underpowered. Furthermore, the generalizability and replication of findings generated from populations of different sociodemographic characteristics (age distribution, for instance) and study design (years of follow-up) are likely to be hampered as left censoring will almost certainly operate differently.

4.4. Recommendations

In this review, we identified three major methodological weaknesses, which we encourage researchers to address in future dementia risk prediction work: (1) over-reliance on one data source, (2) the limited evaluation of analytical assumptions of the models used (Cox and logistic regression), and (3) poor internal and external validations of the prediction models. Hence, we suggest the following recommendations to improve the reliability and accuracy of dementia risk prediction models and provide researches with some guidance:

1. A broader selection of data sources should be considered when developing dementia prediction models, including more diverse samples. Although we acknowledge challenges for differentiation between the dementia types, the discrimination of individuals by dementia type will facilitate the identification of risk factors specific to each dementia type. Data sets with different lengths of follow-up time will permit the evaluation of risk progression over different time frames. We encourage researchers to perform where possible, subgroup analyses to evaluate consistency of results in subgroups of similar features.
2. When using regression analyses for dementia risk prediction model development the assumptions need to be tested thoroughly. If deviations from the (linearity) assumptions are detected, appropriate actions need to be taken. There are a number of more flexible nonparametric extensions for regression analyses, through which the linearity assumption can be relaxed: polynomials or restricted cubic splines can be added to the regression model or the predictor can be (log-) transformed [21]. Similarly, the proportional hazard assumption for Cox regression can be relaxed by implementing alternative formulations of the models (i.e., adding splines).
3. Internal and external validations are key steps during the development and implementation of a new prediction model. Internal validation provides insight to which extent the model is overfitted and whether the predictive ability is too optimistic, whereas external validation proves the ability of the prediction model to perform similarly well in a comparable population. There are different internal validation methods, such as splitting the data into two subsets (a development and a validation sample), leave one-out cross-validation or bootstrapping. Bootstrapping is a recommended internal validation method, also when a large number of predictors are used [18]. However, the method might be limited when used in a small sample. For external validation data with a similar but different population to its development population are needed. As mentioned in recommendation 1, more and easily accessible data are required to enable fast and uncomplicated external validation.
4. We encourage researchers to adopt innovative methodologies such as dynamic risk prediction models [22], as the incorporation of within person change in markers of disease progression is likely to be more informative of risk than data collected at a single point in time while also being more likely to reflect clinical practice.
5. We strongly suggest the adoption of Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guidelines [20] when developing and validating risk prediction models.

4.5. Strength and limitations

The strength of this review is the broad inclusion, allowing a good representation of all possible methodological approaches used in the dementia risk literature. However, we only included published articles and excluded conference abstracts. We also included only population-based samples or samples consisting of MCI individuals, prediction models for other populations (e.g., individuals with Parkinson) might have been developed with different methodological approaches. Furthermore, only dementia was used as a search term, whereby studies looking at specific types of dementia could have been missed.

5. Conclusion

Dementia is one of the leading causes of disability and dependence in late-life [2]. There is a great need to identify individuals at high risk of developing dementia early on. Therefore, the large reliance on one data source, poor validation of results, and limited verification of model assumptions when developing dementia risk prediction models are of concern. It has been shown by Abrahamowicz et al. [16] and Exalto et al. [23] that an application of a more accurate or different analytical technique can result in altered risk prediction. An inaccurate representation of the true relationship of a predictor variable with the outcome might cause false identification of high-risk groups and biased prognosis. To ensure valid conclusions and accurate risk prediction, prognostic studies should rely on statistical methods that correctly represent the actual structure of empirical data and the true complexity of the biological processes under study. Improved practice in data analysis and innovative data designs may advance derivation of dementia risk scores. Machine learning approaches are frequently used for dementia risk prediction model development. As machine learning approaches still need to improve prediction abilities, regression analyses are robust techniques for prediction model development when applied correctly. Compared with machine learning methods, regression analyses are cost effective and require less computational time.

Advanced and innovative dynamic methods already adopted in other research and clinical areas are likely to be the best choice for future dementia risk prediction developments for now. The community will also benefit from the adoption of new data collection modes to advance knowledge in the short term.

Acknowledgments

Funding for this study was provided by Merck Sharp and Dohme Ltd [grant number 2726958 RB0538].

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.trci.2019.08.001.

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed the literature using traditional sources and references from previous publications.
2. Interpretation: Our findings identified several methodological limitations in the existing literature on dementia risk prediction.
3. Future directions: Future research about dementia risk prediction should use a more thorough methodological approach and devote efforts to ensure fulfilment of assumptions, explore interactions, and validation of results.

References

[1] Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, et al. Dementia prevention, intervention, and care. Lancet 2017;390:2673–734.

[2] Robinson L, Tang E, Taylor JP. Dementia: timely diagnosis and early intervention. BMJ 2015;350:h3029.

[3] Stephan BC, Kurth T, Matthews FE, Brayne C, Dufouil C. Dementia risk prediction in the population: are screening models accurate? Nat Rev Neurol 2010;6:318–26.

[4] Tang EY, Harrison SL, Errington L, Gordon MF, Visser PJ, Novak G, et al. Current developments in dementia risk prediction modelling: an updated systematic review. PLoS One 2015;10:e0136181.

[5] Hou XH, Feng L, Zhang C, Cao XP, Tan L, Yu JT. Models for predicting risk of dementia: a systematic review. J Neurol Neurosurg Psychiatry 2019;90:373–9.

[6] Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.

[7] Dallora AL, Eivazzadeh S, Mendes E, Berglund J, Anderberg P. Machine learning and microsimulation techniques on the prognosis of dementia: a systematic literature review. PLoS One 2017; 12:e0179804.

[8] Pellegrini E, Ballerini L, Hernandez M, Chappell FM, Gonzalez-Castro V, Anblagan D, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. Alzheimers Dement (Amst) 2018;10:519–35.

[9] Shalev-Shwartz S. In: Ben-David S, ed. Understanding machine learning: from theory to algorithms. Cambridge: Cambridge University Press; 2014.

[10] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. MIT Press; 2012:480.

[11] Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? Heart 2018;104:1156–64.

[12] Ingram DD, Kleinman JC. Empirical comparisons of proportional hazards and logistic regression models. Stat Med 1989;8:525–38.

[13] Peduzzi P, Holford T, Detre K, Chan Y-K. Comparison of the logistic and Cox regression models when outcome is determined in all patients after a fixed period of time. J Chronic Dis 1987;40:761–7.

[14] Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health 1999;20:145–57.

[15] Pencina MJ, Larson MG, D'Agostino RB. Choice of time scale and its effect on significance of predictors in longitudinal studies. Stat Med 2007;26:1343–59.

[16] Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. Am J Epidemiol 1997;145:714–29.

[17] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.

[18] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683.

[19] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691.

[20] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015; 13:1.

[21] Nick TG, Campbell KM. Logistic regression. In: Ambrosius WT, ed. Topics in biostatistics. Totowa, NJ: Humana Press; 2007. p. 273–301.

[22] Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 2011; 67:819–29.

[23] Exalto LG, Quesenberry CP, Barnes D, Kivipelto M, Biessels GJ, Whitmer RA. Midlife risk score for the prediction of dementia four decades later. Alzheimers Dement 2014;10:562–70.

frontiers in Big Data

ORIGINAL RESEARCH
published: 26 May 2021
doi: 10.3389/fdata.2021.613047

Developing an Explainable Machine Learning-Based Personalised Dementia Risk Prediction Model: A Transfer Learning Approach With Ensemble Learning Algorithms

Samuel O. Danso[1]*, *Zhanhang Zeng*[2], *Graciela Muniz-Terrera*[1] and *Craig W. Ritchie*[1]

[1] Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh Medical School, Edinburgh, United Kingdom, [2] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

Alzheimer's disease (AD) has its onset many decades before dementia develops, and work is ongoing to characterise individuals at risk of decline on the basis of early detection through biomarker and cognitive testing as well as the presence/absence of identified risk factors. Risk prediction models for AD based on various computational approaches, including machine learning, are being developed with promising results. However, these approaches have been criticised as they are unable to generalise due to over-reliance on one data source, poor internal and external validations, and lack of understanding of prediction models, thereby limiting the clinical utility of these prediction models. We propose a framework that employs a transfer-learning paradigm with ensemble learning algorithms to develop explainable personalised risk prediction models for dementia. Our prediction models, known as *source models,* are initially trained and tested using a publicly available dataset ($n = 84{,}856$, mean age = 69 years) with 14 years of follow-up samples to predict the individual risk of developing dementia. The decision boundaries of the best source model are further updated by using an alternative dataset from a different and much younger population ($n = 473$, mean age = 52 years) to obtain an additional prediction model known as the *target model*. We further apply the SHapely Additive exPlanation (SHAP) algorithm to visualise the risk factors responsible for the prediction at both population and individual levels. The best source model achieves a geometric accuracy of 87%, specificity of 99%, and sensitivity of 76%. In comparison to a baseline model, our target model achieves better performance across several performance metrics, within an increase in geometric accuracy of 16.9%, specificity of 2.7%, and sensitivity of 19.1%, an area under the receiver operating curve (AUROC) of 11% and a transfer learning efficacy rate of 20.6%. The strength of our approach is the large sample size used in training the source model, transferring and applying the "knowledge" to another dataset from a different and undiagnosed population for the early detection and prediction of dementia risk, and the ability to visualise the interaction of the risk factors that drive the prediction. This approach has direct clinical utility.

Keywords: early detection, risk factors, Alzheimer's, personalised dementia risk, explainable AI model, ensemble-based learning

OPEN ACCESS

Edited by:
Viktor Wottschel,
Amsterdam University Medical
Center, Netherlands

Reviewed by:
Lyduine Collij,
Academic Medical
Center, Netherlands
Alle Meije Wink,
VU University Medical
Center, Netherlands

Correspondence:
Samuel O. Danso
samuel.danso@ed.ac.uk

Specialty section:
This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Big Data

Received: *01 October 2020*
Accepted: *13 April 2021*
Published: *26 May 2021*

Citation:
Danso SO, Zeng Z, Muniz-Terrera G and Ritchie CW (2021) Developing an Explainable Machine Learning-Based Personalised Dementia Risk Prediction Model: A Transfer Learning Approach With Ensemble Learning Algorithms. Front. Big Data 4:613047. doi: 10.3389/fdata.2021.613047

INTRODUCTION

Dementia is the consequence of a number of progressive neurodegenerative diseases with Alzheimer's disease (AD) accounting for ∼60–80% of all types of dementias (Gaugler et al., 2019). AD is considered to be one of the top 10 causes of death, globally. Due to the progressive nature of the disease, people with dementia have different degrees of deterioration in cognition, memory, mental, and other functions (Lyketsos et al., 2002). Moreover, the socioeconomic burden of the disease is estimated to be in the region of one trillion USD per year (World Health Organization, 2017). Dementia has no cure; however, with early detection and diagnosis, it may be possible to delay the onset, which will help reduce the economic burden it currently poses on the society (Prince et al., 2018).

A recent Lancet report has identified modifiable risk factors, which when well-managed could reduce the risk of dementia or delay its onset (Livingston et al., 2020). However, the complexity of the interaction among these risk factors requires computational approaches capable of detecting patterns from these complex interactions to be able to achieve accurate prediction. Meanwhile, machine-learning based approaches have successfully been employed to help identify complex relationships between risk factors and their effect on disease outcomes in various application areas within the care pathway of patients. Examples of such application areas include prediction of pneumonia risk and 30-days readmission in hospital (Caruana et al., 2015), a real-time prediction of patients at the risk of septic shock (Henry et al., 2015), and application of machine learning model in breast screening (Houssami et al., 2017).

Following the above success storeys in the non-dementia domain, numerous attempts are being made to develop machine-learning models for dementia risk prediction. For example, Skolariki et al. (2021) applied machine learning algorithms to predict the likelihood of people with mild cognitive impairment converting to dementia based on features extracted from brain scans. Cui et al. (2019) also applied a recurrent neural network to develop a dementia risk prediction model based on longitudinal features extracted from brain scans. Other studies have also explored features obtained from sources, such as neuropsychological assessments (Barnes et al., 2009; Johnson et al., 2009; Lee et al., 2018; Adam et al., 2020). While these attempts have shown promising results, the prediction algorithms are mostly trained with samples containing diagnosis information and therefore unable to predict beyond the critical window of diagnosis (Prince et al., 2018), making these models ungeneralizable to relatively younger populations (Goerdten et al., 2019). Furthermore, despite these promising results achieved by machine learning-based approaches for dementia, their utility in healthcare settings remains limited partly due to the difficultly in interpreting the outputs of these models (Pellegrini et al., 2018). Interpretable models offer users the confidence and the ability to understand why a certain prediction was made for an individual and the specific underlining factors that led to the prediction. Confidence in how the prediction is made would allow the clinician to communicate this optimally to the patient and intervene. However, lack of confidence on the part of clinicians has resulted in the limited use of powerful machine learning approaches, such as deep learning and ensemble-based learning in developing prediction models for decision support systems in the dementia care pathway. Meanwhile, the complex nature of dementia, which results in complex data structures, makes it imperative to continue to explore these powerful machine learning methods, where traditional approaches, despite their limitations in handling complex data structures (Breiman, 2001), have widely been employed (Goerdten et al., 2019).

We develop and evaluate two ensemble-based interpretable models capable of learning patterns from the complex interactions among risk factors to be able to predict dementia risk at both population and individual levels up to an average of 14 years in advance. Unlike the approaches described above, our final model predicts individual dementia risk based on the parent history of dementia and genetic information about the individual. The prediction models are built using Random Forest (RF) and XGboost algorithms. Briefly, RF like other ensembles of classification and regression trees employs a "divide-and-conquer" strategy in the process of learning by repeatedly partitioning the input data into a number of large classification trees and fitting a prediction model for each tree (Breiman et al., 1984). It then employs the non-parametric bootstrap method (Efron and Tibshirani, 1994) to build a prediction model for each tree. Similarly, the XGBoost also belongs to the family of classification and regression trees and adopts the RF approach to learning. However, XGBoost employs a step-wise, additive approach to sequentially build a prediction model for each tree, while taking into account the difficulties encountered in fitting previous models (Natekin and Knoll, 2013). It is worth noting that RF and XGboost both combine the predictions from weak learners to produce a final model—a process known as "voting." These algorithms have been demonstrated to be powerful when applied to various problems, such as risk prediction of hypoxaemia during general anaesthesia and surgery (Lundberg et al., 2018).

We argue that our proposed approach provides useful and actionable information to assist clinicians and other users in their decision-making process around diagnosis, prognosis, and management. We also believe that this is an important step for machine learning in neurodegenerative disease research and translation to clinical care. Our approach not only significantly improves the ability for the early detection of neurodegenerative disease but also the ability to explain the predictions from accurate and complex models in order to understand drivers of the prediction for important intervention strategies to be developed.

METHODS

Overview of the Research Framework

It is believed that dementia clinically manifests after decades of exposure to risk factors (Ritchie and Ritchie, 2012). Therefore, the aim of this project was to develop a machine learning model capable of predicting the risk of developing dementia decades prior to the onset of the dementia syndrome. To achieve this, the task was formulated as a transfer learning classification

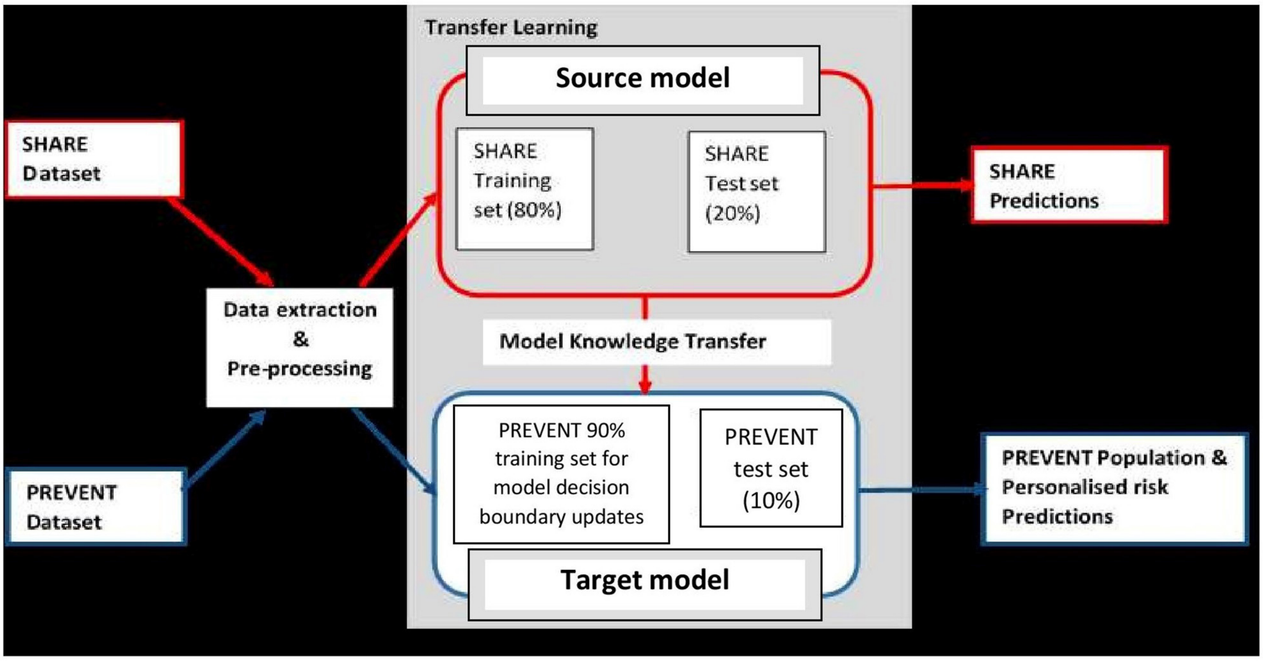


FIGURE 1 | Transfer learning process showing how data extraction and pre-processing procedures are applied to SHARE and PREVENT datasets. A prediction model (Source model) is built using the SHARE dataset with 80% of the data used for training and 20% held-out for testing for SHARE predictions. The Source model is updated with 90% PREVENT training and the updated prediction model (Target model) is applied to PREVENT 10% test set held-out for population as well as personalised risk prediction of dementia.

problem (Pan and Yang, 2009). This made it possible to develop the machine learning prediction model using the data drawn from different populations and applied the model to another population. **Figure 1** illustrates the methodology employed. As the figure shows, unlike traditional machine learning where a model is developed and applied to predict data from the same population, our model was developed using external data source and transferred the knowledge learned from the external population and applied it to data from population of different characteristics. The characteristics of the data sources are discussed in the next section.

Data Description and Preprocessing

The data sources used in developing the models were obtained from the Survey of Health, Ageing, and Retirement in Europe (SHARE) study (Börsch-Supan et al., 2013) and the PREVENT Dementia programme (Ritchie and Ritchie, 2012). While both SHARE and PREVENT projects are related to dementia research, the rationale and aims of each of the studies vary resulting in differences in the datasets. **Table 1** shows a brief description of the datasets. While SHARE population covers 20 European countries with the mean age of 69 years, the PREVENT data, on the other hand, is a relatively younger cohort with the mean age of 52 years drawn from a population limited to the United Kingdom. Further, the SHARE cohort includes individuals with some having been diagnosed with dementia, while the PREVENT cohort contains healthy individuals without a diagnosis of dementia. However, the PREVENT study participants are children of individuals with or without a diagnosed dementia. The study also collects information about the apolipoprotein E (ApoE) genotype of each individual.

Even though both SHARE and PREVENT research programmes have different research aims and objectives, there was a high degree of overlap between the two datasets in terms of data collection. In order to make transfer learning possible, it was important to focus on common data items between the two datasets. **Table 2** shows the categories of common variables found in both datasets. We extracted data records from the SHARE dataset and merged the data of individuals across waves 1–6 which covers the period between 2004 and 2015. Therefore, from the SHARE cohort, it was possible to build a prediction model using a longitudinal dataset of 14 years of follow-up data. The PREVENT dataset on the other hand is the baseline data collected between February 2014 and October 2018.

The difference in data collection protocols used by the studies resulted in structural differences in data. To address these differences, we devised a pre-processing procedure to harmonise the representation of the data items, which were employed as features to train the learning algorithms. All medical history variables were processed to have binary feature representation based on the responses as either condition being present or not present, with a feature value of "1" and "0," respectively. The Body Mass Index (BMI) as per WHO classification was applied to obtain the following four categories: underweight (<18.5 kg/m^2), normal (18.5–24.9 kg/m^2), overweight (25–29.9 kg/m^2), and obese (>30 kg/m^2) with feature values of

"0," "1," "2," and "3," respectively Furthermore, "marital status" had categorical entries ("divorced," "married," "living with spouses," "married," "not living with spouse," "never married," and "registered partnership"), and each of these was separately represented as binary based on the response as either "yes" or "no," with a feature value of "1" and "0," respectively. The International Standard Classification of Education scheme was applied to "education level" variable to have seven categories with feature value representations (0 = none; 1 = first stage of basic education; 2 = lower secondary education or second stage of basic education; 3 = upper secondary education; 4 = post-secondary non-tertiary education; 5 = first stage of tertiary education; and 6 = second stage of tertiary education). The "daily activity" variables had two categories: "vigorous" and "moderate" sports with each having feature value representations (0 = hardly ever or never; 1 = one to three times a month; 2 = once a week; and 3 = more than once a week). We believe that this method of representation provides information on the activity as well as the intensity of the activity, which can be useful for the learning algorithms. The "smoking" variable was also processed to have a binary representation based on the responses with feature values (0 = never smoked and 1 = current or past smoker). Finally, the SHARE dataset contained data on whether a participant had been diagnosed with Alzheimer's disease (AD) and those without a diagnosis. This was therefore used as the class variable for the prediction model feature values representation (Non-AD = no diagnosis; AD = diagnosis of Alzheimer's dementia). However, in the absence of a diagnosis in the PREVENT dataset, and to facilitate the evaluation of our approach, we employed a classification scheme proposed by Ritchie and Ritchie (2012) to group the participants according to parental clinical status and ApoE genotype. Therefore, participants with a parental dementia diagnosis and ApoE 4 genotype were allocated to a "High-Risk" (HR) group as these individuals were considered to be at high risk of dementia. All other participants were allocated to a "Low-Risk" (LR) group. The final distribution of classes is as follows: SHARE dataset, Non-AD (95%) and AD (5%); PREVENT dataset HR (23%) and LR (77%).

TABLE 1 | Characteristics of SHARE and PREVENT datasets.

Data description	SHARE data	PREVENT data
Population	20 European countries	The United Kingdom
Number of samples	84,856	473
Mean age	69	52
Number of years of follow-ups	14 years (2004–2015), 2 years interval on average	Only used baseline data
Class distribution	Diagnosis • Diagnosis of Alzheimer's disease—"AD" (*n* = 4,157) • No diagnosis of-Alzheimer's disease diagnosis—"non-AD" (*n* = 80,699)	Parental diagnosis of AD and Apolipoprotein E4 allele (ApoE4) genotype status of individual • Parental diagnosis of AD + ApoE4 status—"High Risk" (*n* = 109) • No parental diagnosis of AD + No ApoE4 status of individual—"Low Risk" (*n* = 364)

Building the Prediction Model

We built four ensemble-based prediction models by training RF and XGBoost algorithms. The algorithms were trained by applying a hybrid approach that combines cross-validation and hold out, through a procedure we refer to as *cross-validation with hold out* (Pedregosa et al., 2011). This procedure involved splitting the SHARE data into training and test sets. The training set (D_train), which constituted 80% of the SHARE data, was used to train the algorithms including hyperparameters tuning. The 20% test set (D_eval) was held and used only for the model performance evaluation. Similarly, the PREVENT data was also split into 80% training set (PREV_train) and 20% test set (PREV_eval). The splits were stratified in order to ensure the equal proportion of class representation in both training and test sets. A summary of our cross-validation with hold out training of algorithms procedure is as follows:

- Step 1: We employed a 5-fold cross-validation during training, which randomly split the 80% training set into 5-folds each containing a subset of training (D_train$_{1-5}$) and validation (D_val$_{1-5}$) sets.
- Step 2: We applied a set of initial hyperparameters to train the algorithm to obtain five different models using D_train$_{1-5}$ and D_val$_{1-5}$, to obtain a number of potential hyperparameters from each cross-validation.
- Step 3: We then applied the random search optimization algorithm (Bergstra and Bengio, 2012), to search and choose from a set of potential number of hyperparameters derived

TABLE 2 | The common data items between SHARE and PREVENT datasets used to develop the prediction models.

Data category	Data items
Sociodemographic	• Gender • Age • Education level • Marital status • Had children? • BMI
Self-reported medical history	• Heart attack • Hypertension (high blood pressure) • High cholesterol • Diabetes • Lung disease • Peptic ulcer disease • Parkinson's disease • Emotional disorders • Osteoarthritis
Life style	• Daily activity • Smoking

from Step 2 to obtain the optimal set of hyperparameters based on the evaluation function of the optimization algorithm. **Table 3** shows the set of initial and optimal hyperparameter settings obtained.

- Step 4: Once the optimal hyperparameters are obtained, we then retrained the algorithm using the optimum hyperparameters on the entire training set, D_train.
- Step 5: We applied the procedures in Steps 2–4 for RF and XGBoost to obtain SHARE_RF_pred and SHARE_XGBoost_pred prediction models, respectively.
- Step 6: We evaluated the performance of the prediction models obtained in Step 5 by applying SHARE_XGBoost_pred and SHARE_RF_pred to the hold-out test set (D_eval).
- Step 7: We employed the method proposed by DeLong et al. (1988) to carry out a pairwise comparison of the receiver operating curve (ROC) to compare the performance difference between SHARE_XGBoost_pred and SHARE_RF_pred to determine the best model.
- Step 8: We randomly spit the PREVENT data into 80% training set (PREV_train) and 20% held out test set (PREV_eval). Again, the split was stratified in order to ensure an equal proportion of class representation in both the training and test sets.
- Step 9: We employed a parameter-transfer learning approach as described by Yao and Doretto (2010) to build a target model. This approach assumes that the target shares parameters with the best source model as determined in Step 7. The parameters of the best source model are further updated using the PREV train set. This process adjusted the decision boundaries of the source model to produce PREVENT_target prediction model.
- Step 10: We evaluated the performance of prediction models obtained in Step 9 by applying them to the hold-out test set (PREV_eval).
- Step 11: We trained the XGBoost algorithm using PREV_train and applied procedures into Steps 2–4 to obtain a prediction model (PREVENT_only).
- Step 12: We evaluated the performances of PREVENT_target and PREVENT_only by applying them to the hold-out test set (PREV_eval).
- Step 13: We finally applied the procedures in Step 7 to compare the performance difference between the PREVENT_target and PREVENT_only to determine the best model.

TABLE 3 | Hyperparameter settings for prediction models.

Algorithm	Initial parameters	Optimal hyperparameter settings
Random Forest	n_estimators = range (5, 40), max_features = ['auto', 'sqrt', 'log2'], max_depth = range (10, 25), criterion = [gini, entropy]	Bootstrap = True; ccp_alpha = 0.0; class_weight = None; criterion = entropy; max_depth = 24; max_features = sqrt; max_leaf_nodes = None; max_samples = None; min_impurity_decrease = 0.0; min_impurity_split = None; min_samples_leaf = 1; min_samples_split = 2; min_weight_fraction_leaf = 0.0; n_estimators = 33, n_jobs = None; oob_score = False; random_state = None; verbose = 0; warm_start = False
XGBoost	n_estimators = range (1, 20), max_depth = range (10, 25), learning_rate = [.1,.2,.4,.45,.5,.55,.6], colsample_bytree': [.6,.7,.8,.9, 1], booster = gbtree, min_child_weight = [0.001, 0.003, 0.01]	Objective = multi:softprob; base_score = 0.5; booster = gbtree; colsample_bylevel = 1; colsample_bynode = 1; colsample_bytree = 0.7; gamma = 0; gpu_id = −1; importance_type = gain; interaction_constraints = None; learning_rate = 0.5, max_delta_step = 0; max_depth = 24; min_child_weight = 0.003; missing = nan; monotone_constraints = None; n_estimators = 16; n_jobs = 0; num_parallel_tree = 1; random_state = 0; reg_alpha = 0; reg_lambda = 1; scale_pos_weight = None; subsample = 1; tree_method = None; validate_parameters = False; verbosity = None; num_class = 2

Performance Evaluation

We employed a series of metrics to evaluate the performance of the models based on the D_eval and PREV_eval unseen datasets. As already pointed out, D_eval contained "AD" and "No-AD" which served as the ground truth for the evaluation of SHARE_RF_pred and SHARE_XGBoost_pred models. PREV_eval on the other contained "HR" and "LR" as explained above, and this served as the ground truth for the evaluation of our PREVENT_target and PREVENT_only models. These metrics were primarily based on the following information obtained from the outputs of the prediction models: Refer False Positive (FP), False Negative (FN), True Positive (TP), and True Negative (TN) (Pollack, 1970) for details of these metrics. The comparison of the models was based on geometric accuracy (GA) as expressed in Equation (3) which is derived from Equations (1) and (2) which represent sensitivity and specificity, respectively. GA accounts for both majority and minority class error rates which makes it ideal for imbalanced problems (Kim et al., 2015).

$$Sensitivity = \frac{Number\ of\ TP}{Number\ of\ TP + Number\ of\ FN} \tag{1}$$

$$Specificity = \frac{Number\ of\ TN}{Number\ of\ TN + Number\ of\ FP} \tag{2}$$

$$Geometric\ Accuracy = \sqrt{(Sensitivity * Specificity)} \tag{3}$$

We also employed area under the receiver operating curve (AUROC) to further explore the robustness of our models, given the wide usage of this metric in medical applications (Mandrekar, 2010). Also, as already stated, a significant test was used to examine the performance differences between the prediction models.

Finally, we employed a method proposed by Taylor and Stone (2009) to examine the efficacy of our transfer learning approach based on a learning ratio as expressed in Equation (4).

$$ratio = \frac{area\ under\ curve\ with\ transfer - area\ under\ curve\ without\ transfer}{area\ under\ curve\ with\ transfer} \quad (4)$$

Feature Importance and Model Interpretability

An important advantage of tree-based algorithms is their ability to provide information on the decisions made around predictions. This information is provided in the form of weights that are assigned to the features as a result of the learning process. The value of weight assigned to a given feature is an indicator of the importance of that feature as determined by the prediction model, which enabled us to examine how each feature was ranked by the prediction models.

We further applied the SHapley Additive exPlanation (SHAP) algorithm to explore the interactions between the features (Lundberg et al., 2018). Briefly, the algorithm is inspired by game theory, where the interaction between features is considered as a "team" of features, with each feature being a member of the team responsible for driving the overall risk. An instance of the interaction between the features registers a set of predicted values produced by the prediction model. These values serve as input for the SHAP algorithm to generate another set of values known as "impact values." The SHAP values provide a dynamic view of the effects of the interaction between the features to determine the probability of risk and the role of each feature on the individual level. Furthermore, the SHAP algorithm offers the possibility to compare an individual predicted risk probability with a baseline prediction, which is the average predicted probability known as the "base value."

RESULTS

Model Performance Analyses

Figure 2 shows the confusion matrix of the results obtained when SHARE_RF_pred (**Figure 2A**) and SHARE_XGBoost_pred (**Figure 2B**) models were applied to 20% of SHARE unseen test set. The figure also shows the results when PREVENT_target (**Figure 2C**) and PREVENT_only (**Figure 2D**) models were applied to 20% of PREVENT unseen test set. **Table 4** further shows a summary of the performances obtained. As seen from the table, SHARE_XGBoost achieves a GA of 87%, specificity of 99%, sensitivity of 76%, and AUROC of 96%. In comparison, SHARE_RF_pred achieves a GA of 85%, specificity of 99%, sensitivity of 73%, and AUROC of 94%. **Figure 3A** shows an AUROC curve comparison between SHARE_RF_pred and SHARE_XGBoost, with SHARE_XGBoost showing a marginal difference in the performance between the two models. A pairwise comparison of the AUROC scores between the two prediction models demonstrates a significant difference in performance ($P < 0.0001$, 95% Confidence Interval: 0.01–0.02), suggesting SHARE_XGBoost as the best performing model.

Again, as seen from **Table 4**, PREVENT_target achieves a GA of 56.5%, specificity of 84.7%, sensitivity of 38.1%, and AUROC of 63%. In comparison, PREVENT_only achieves a GA of 39.6%, specificity of 82.0%, sensitivity of 19%, and AUROC of 51%. **Figure 3B** shows an AUROC curve comparison between PREVENT_target and PREVENT_only, with PREVENT_target showing a marginal difference in performance between the two models. Even though a pairwise comparison of the AUROC scores between PREVENT_target and PREVENT_only, no significant difference in performance is observed ($P = 0.2166$, 95% Confidence Interval: 0.07–0.325), the PREVENT_target model outperformed PREVENT_only model across all the performance metrics as shown in **Table 4**. There is an increase in the sensitivity of 19.1%, specificity of 2.7%, GA of 16.9%, AUROC of 11%, and a transfer-learning rate of 20.6%.

Feature Importance Analysis and Interpretability of Personalised Risk Prediction

Even though RF and XGboost are both considered ensemble-based algorithms, the learning strategy tends to differ as briefly discussed. From that score, we examine how both models assessed the importance of the features used in training the models. **Figures 4A,B** depict a comparison between SHARE_RF_pred and SHARE_XGBoost_pred prediction models on how features were ranked based on the weights assigned. As shown by **Figures 4A,B**, while significant similarities in the ranking of the features exist between the two models, some striking differences can also be observed. For example, the ranking of the top seven features of both RF and XGBoost appear to be in the same order, with "“age" being the most important feature followed by "moderate sport," "education," "vigorous sports," "BMI," "hypertension," and "esmoked." Some differences in rankings were observed. Where RF ranks "gender" and "emotional disorders" as the 8th and 9th most important features, XGBoost ranks "high cholesterol" and "osteoarthritis," respectively. Additionally, RF ranks "widowed" as the 10th most important feature, whereas XGBoost ranks "diabetes" as the 10th most important feature, and ranks "widowed" as one of the least important features (ranked 18th).

Similarly, a comparison between PREVENT_only and PREVENT_target shows how these prediction models ranked the features as shown in **Figures 4C,D**, respectively. Again, while there appear to be some overlaps in the order of feature rankings between the models, some differences can also be observed. For example, "age" remains the most important feature among the two models. A close examination of the top 10 features of the models show some differences in the order of rankings. For example, while PREVENT_only ranks "divorced," and "no_children" among the top 10, PREVENT_target also ranks "BMI" and "gender" among the top 10, but ranks "divorced," and "no_children" in the 11th and 13th positions, respectively. Even though these differences in feature rankings can be observed between these two models, the difference is not statistically significant. However, because our PREVENT_target demonstrated some marginal increase in the performance over PREVENT_only, our analysis will be based on the output of PREVENT_target model. A further comparison of the order

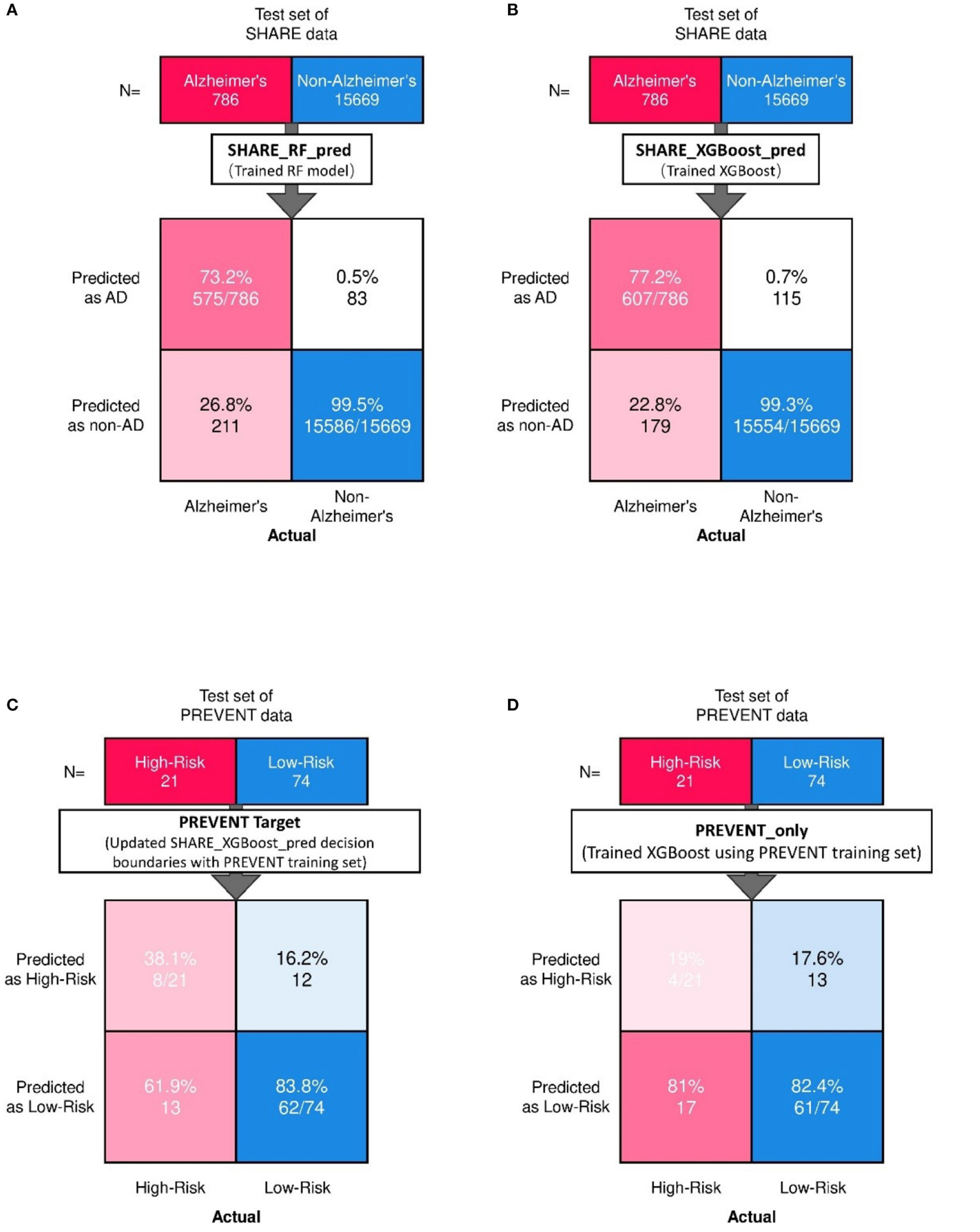


FIGURE 2 | Confusion matrix showing the prediction results from unseen 20% of SHARE test data as predicted by **(A)** Random Forest **(A,B)** XGBoost models. Also showing are the prediction results from 20% unseen PREVENT test data as predicted by **(C)** Updated SHARE_XGBoost_pred decision boundaries with PREVENT training set and **(D)** Trained XGBoost using PREVENT training set.

TABLE 4 | Summary of prediction models on the unseen test set.

Model	Sensitivity (%)	Specificity (%)	Geometric Accuracy (%)	AUROC (%)	*P*-value	Transfer learning efficacy (%)
SHARE_RF_pred	73	99	85	94	$P < 0.0001$	N/A
SHARE_XGBoost_pred	*76 (+3%)	*99 (0%)	*87 (2%)	*96 (2%)		
PREVENT_target	**38.1 (+19.1%)	**84.7 (+2.7%)	**56.5 (+16.9%)	**63 (+11%)	$P = 0.2166$	20.6%
PREVENT_only	19.0	82.0	39.6	51		

**Performance comparison in relation to SHARE_RF_pred.*

***Performance comparison in relation to PREVENT_only.*

of rankings of features between SHARE_XGBoost_pred as the source model and our PREVENT_target as the target model also shows 70% overlap among the top 10 features as ranked by both the models. The differences observed include: "emotional_disorders," "hypertension," and "diabetes" ranked among the top 10 by SHARE_XGBoost_pred, but ranked by PREVENT_target model at 12th, 14th, and 21st positions, respectively.

Furthermore, we examined the performance of the models at individual levels. **Figure 5** shows the visualisation of SHAP values of four randomly selected prediction outputs when SHARE_XGBoost_pred was applied to SHARE unseen test set. **Figure 5A** shows an individual with AD and correctly predicted by the model, with the probability of 80%. **Figure 5B** shows an individual with AD which is incorrectly predicted as a non-AD with the probability of 6%. **Figure 5C** shows an individual without AD predicted as AD with the probability of 66%. **Figure 5D** also shows an individual without AD and correctly predicted as a Non-AD with the probability of 4%. The figures also show the risk factors that drive each of the probabilities, with red indicating risk factors and blue suggesting protective factors. For example, **Figure 5A** shows a 69-year-old woman correctly predicted to be living with AD with the probability of 80%. While smoking, vigorous sports, education, BMI, and osteoarthritis appear to be playing a role in the prediction, the lack of moderate sports appears to be the most important risk factors as determined by the colour (red) and the length of the bar allocated to each risk factor. In contrast, as **Figure 5B** shows, age and the fact that the person engages in moderate sports appear to have significant impact on the prediction, which resulted in a relatively low risk of probability of 6%. Similarly, age and moderate sports appear to have a significant impact on the prediction of probabilities in both **Figures 5C,D**. However, while moderate sports appear to be protective for the individual as shown in **Figure 5C**, the relatively older age (80 years) and the lack of education appear to be the risk factors that have a significant impact on the prediction resulting in the probability of 66% of AD. In contrast, the individual shown in **Figure 5D** is relatively young and engages in moderate as well as vigorous sports, which appear to be the proactive factors driving the prediction with a relatively low probability of 4% risk of AD.

Examining our target model at the individual level, **Figure 6** shows randomly selected outputs when PREVENT_target model was applied to PREVENT unseen test set. **Figure 6A** shows a low-risk individual predicted as a high-risk with the probability of 70%. **Figure 6B** shows a high-risk individual correctly predicted with the probability of 7%. **Figure 6C** shows a high-risk individual predicted as low-risk with the probability of 19%. **Figure 6D** is also a low-risk individual correctly predicted as low-risk with the probability of 27%. As the figures show, while age appears to be the most protective factor for all the individuals, the lack of vigorous sports, relatively low education, and BMI appear to be the risk factors with the greatest impact. A closer look at **Figure 6A** shows a 60-year-old individual who has no education and lacks physical activity and therefore predicted by the model to be at high risk despite having been allocated to the low-risk group. Similarly, **Figure 6B** shows a 52-year-old individual belonging to the high-risk group and correctly predicted by the model with a probability of 63%. In this figure, individual age is the most protective factor, while education (3 = upper secondary level) and having a healthy weight (BMI = 1) appear to be risk factors. This may suggest that higher education may be critical for individuals with an APOE e4 gene and a parental history of dementia, compared to individuals without that fall outside the high-risk group.

DISCUSSION

This study developed an ensemble-based machine-learning model to predict Alzheimer's dementia risk at both population and individual levels based on the data drawn from two populations with different characteristics. Our models were built using large heterogeneous data drawn from a population of 20 European countries with up to 14 years of follow-up data. Our best model achieves high-performance accuracy, obtaining an AUROC score of 96% on the unseen test set. The decision boundaries of the best model were further updated through transfer learning. The update was done using data from a different population with different dementia risk profiles to produce a target model. The target model achieves an AUROC score of 63% and a transfer learning efficacy rate of 20%. It is also able to visualise the risk as well as protective factors that are responsible for the prediction at both population and individual levels.

To the best of our knowledge, this is the first approach that employs transfer learning with ensembles to develop dementia risk prediction models and visualisation of risk factors from an undiagnosed population in mid-life. Although numerous computational approaches have been developed, these methods

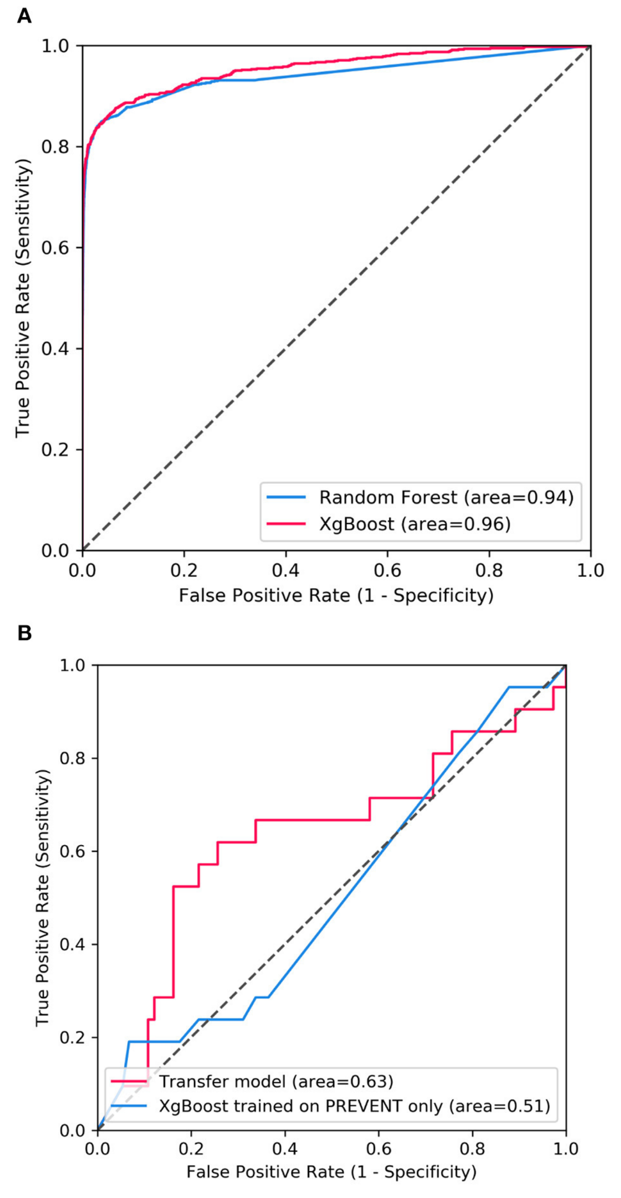


FIGURE 3 | Showing ROC curves with AUROC scores of **(A)** the performance difference between Random Forest and XGBoost prediction models when applied to 20% SHARE unseen test set and **(B)** the performance compassion between XGBoost model updated with PREVENT training set (Transfer model) and XGBoost trained with PREVENT only (PREVENT only model) and applied to 20% PREVENT unseen test set.

have been limited in terms of sample size and the over-reliance on a homogenous sample for validation (Goerdten et al., 2019). van Maurik et al. (2019) attempted to address this issue by combining data from older adults in different populations across Europe and North America to develop dementia-risk prediction models for people with mild cognitive impairment. They employed traditional statistical modelling approaches and biomarkers, such as cerebrospinal fluid and imaging data to develop the prediction models. While we are unable to compare our proposed approach to that of van Maurik et al. (2019) due to differences in data used, it would be interesting to compare the performance of the two modelling approaches on the same dataset in the future.

Even though the relative differences in feature rankings between the models may be hard to interpret relative to their importance in predicting the dementia risk, and given that XGBoost outperforms RF as our significant test suggests, it would be reasonable to conclude that the feature rankings of XGBoost model could be more accurate and therefore reliable. The prediction models developed here identified risk factors that agree with previous literature. We demonstrate this by examining the top 10 features as ranked by the XGboost prediction models. Numerous studies have concluded that age remains the single biggest risk factor (Song et al., 2014). This is consistent with our model, ranking age to be the most important risk factor. Even though age is considered a non-modifiable risk factor, the Lancet commission report on dementia prevention by Livingston et al. (2020) identified a number of risk factors which when modified could reduce the risk of dementia by 40%. The report identified less education, hypertension, hearing impairment, smoking, obesity, depression, physical inactivity, diabetes, and infrequent social contact as potentially modifiable risk factors. Seventy percent of these risk factors were ranked among the top 10 by the study's prediction model as shown in **Figure 4**.

Furthermore, the interaction effects identified by the study's models are also in accordance with the existing evidence. For example, low education level is known to account for up to 8% and physical inactivity accounts for up to 3% of the dementia risk (Livingston et al., 2017). Again, both education and physical activity are associated with cognitive reserves and improvement in mental functions, suggesting that these could act as protective factors (Sharp and Gatz, 2011). Therefore, poorly educated individuals with a sedentary lifestyle could have an increased risk of dementia. This phenomenon is consistent with what is observed in **Figures 5**, **6**. As **Figure 5A** demonstrates, the relatively low education and low levels of physical activity (moderate/vigorous sports) were the two major risk factors among the (non-age) other risk factors that increased the risk of dementia up 80% of this individual. This is consistent with what is observed in **Figure 6A** which shows an individual considered to be at low risk but due to lack of education and physical activity, the risk profile of this individual is predicted with 70% probability, with age being the only protective factor.

While the majority of the top 10 risk factors ranked by the study's prediction model were part of those identified by the recent Lancet Commission report, there are a few that appear to be playing a major role in the risk prediction but not currently part of the report. **Figure 6B** demonstrates the effect of emotional disorder on the risk of dementia at the individual level. Again, while age and physical activity remain significant protective factors, emotional disorder appears to be playing a significant role in the 7% risk of Alzheimer's Dementia for this individual. Therefore, any intervention in the emotional health of this participant chosen for illustrative purposes could further reduce their risk. This approach is exactly what is envisaged in the Brain Health Clinics being developed across Europe (Frisoni et al., 2020) based on a consensus led by our group in how to change clinical services for dementia prevention (Ritchie et al.,

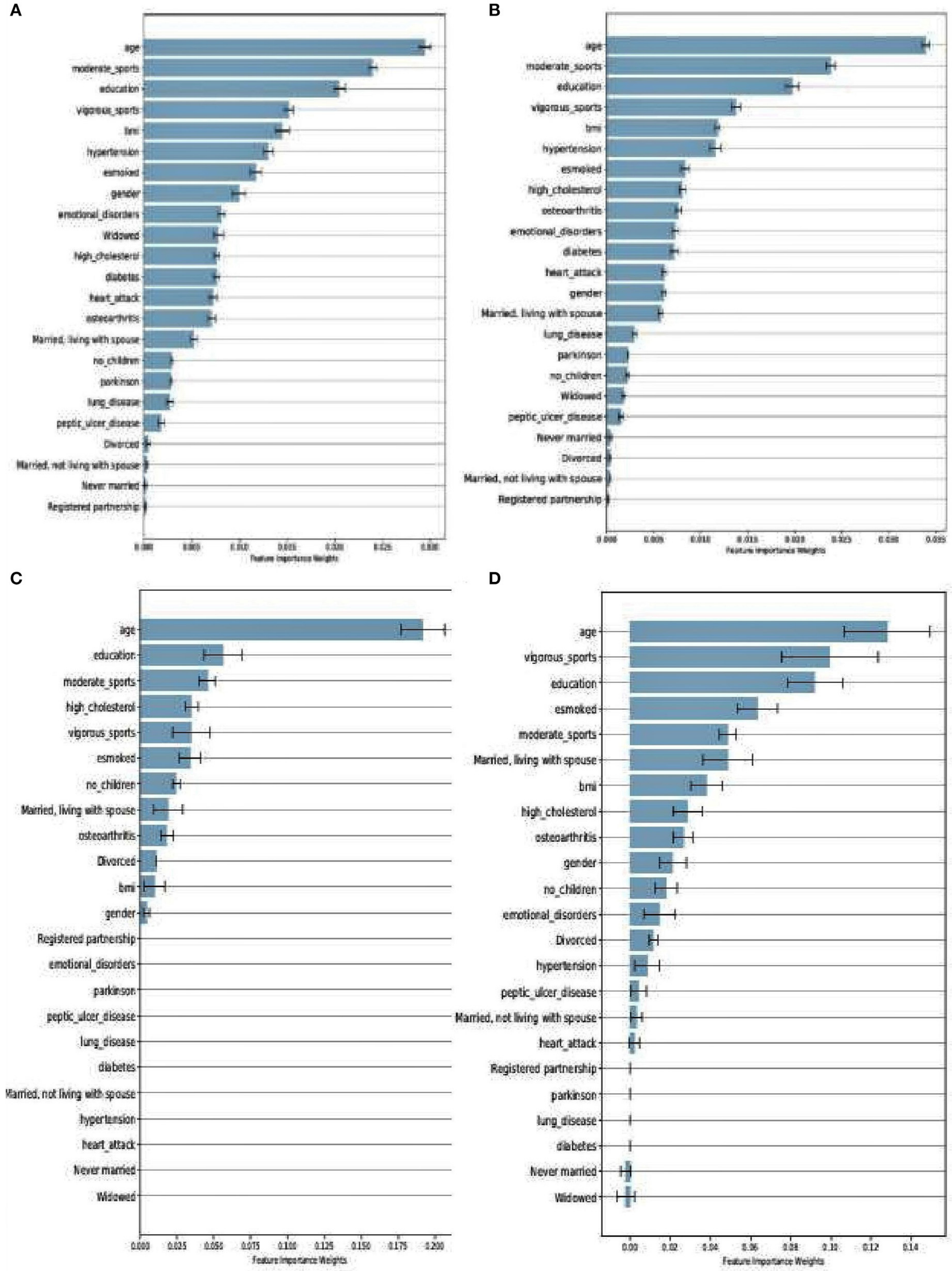
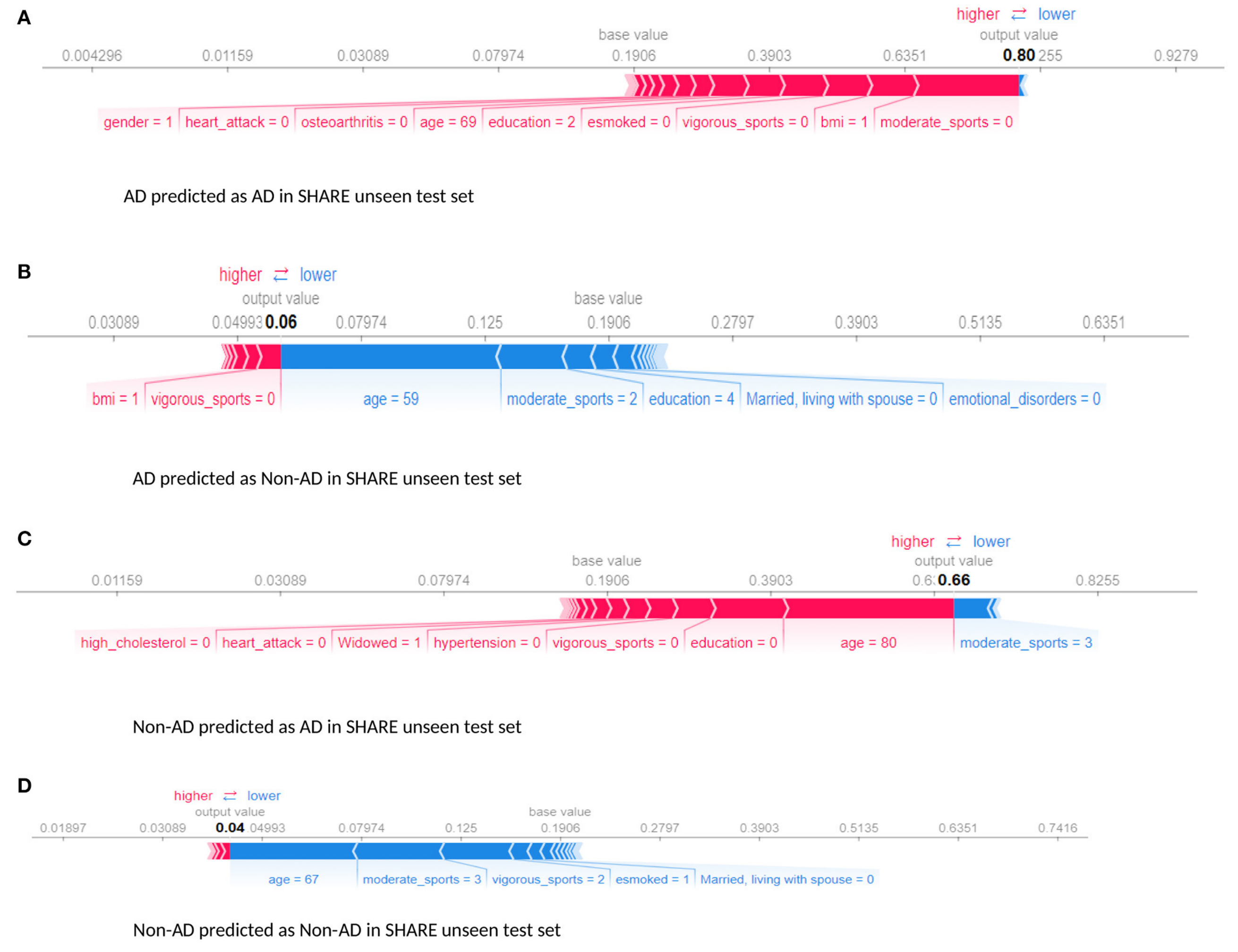


FIGURE 4 | Feature importance as ranked by the weights derived from SHARE_RF_pred **(A)** and SHARE_XGBoost_pred **(B)** prediction models that were trained using SHARE dataset. It also shows the ranking of features of PREVENT only **(C)** and PREVENT target **(D)**.



AD predicted as AD in SHARE unseen test set

AD predicted as Non-AD in SHARE unseen test set

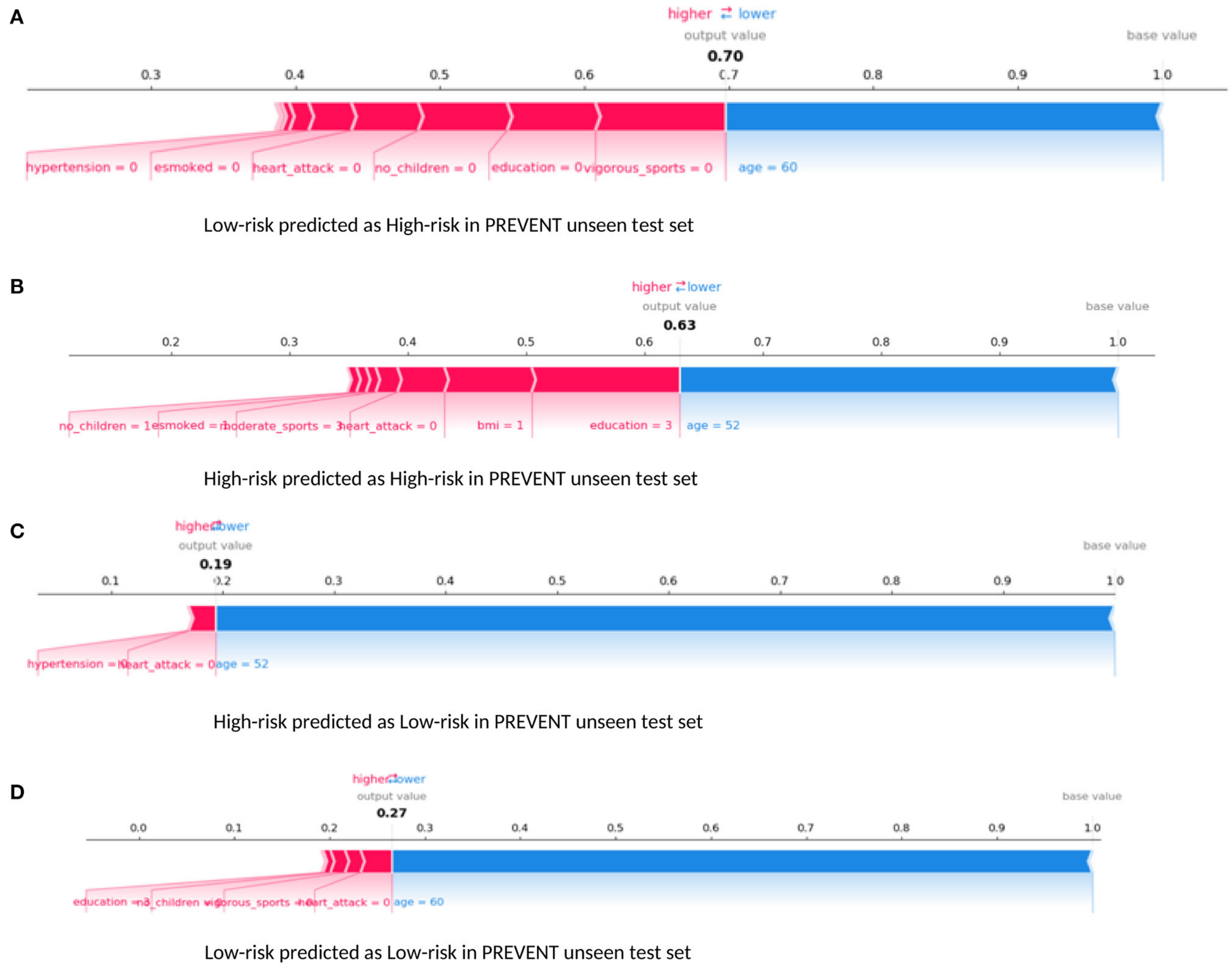
Non-AD predicted as AD in SHARE unseen test set

Non-AD predicted as Non-AD in SHARE unseen test set

FIGURE 5 | Force plot showing the effect of SHAP values on the interaction of features and the overall prediction at the individual level. This shows examples of prediction outputs with taken from SHARE as predicted by the SHARE model. Features in red show risk factors pushing up the overall probability while those in blue are protective factors pushing down the probability. **(A)** Shows SHARE participants predicted to have AD with 85% probability. **(B)** Shows SHARE participant diagnosed to have AD but has been predicted by the model to be Non-AD with 6% probability. **(C)** Shows a Non-AD participant predicted as AD with 63% probability. **(D)** Shows as Non-AD participant predicted as Non-AD with 4% probability. Feature labels are: **esmoked** (0 = never smoked); **emotional disorders** (0 = no); **hypertension** (0 = no); **osteoarthritis** (1 = yes); **high cholesterol** (0 = no); **heart attack** (0 = no); **education** (2 = lower secondary education or second stage of basic education; 3 = upper secondary education); **moderate sports** (0 = hardly ever, or never, 1 = one to three times a month); **vigorous sports** (0 = hardly ever, or never, 1 = one to three times a month); **no children** (0 = no children); **widowed** (1 = yes); **BMI** [1 = under weight (<18.5)]; **married, living with spouse** (0 = no) and **gender** (1 = male).

2017). This is based on collecting data from these Brain Health Clinics to support Real World machine learning approaches and using these algorithms to support the development of personalised prevention plans driven by early disease detection and comprehensive risk profiling.

Even though the performance of the study's prediction model demonstrates a potential clinical utility, we do acknowledge that it would benefit from further development and validation. Firstly, it would be beneficial to evaluate the effect of additional data sources derived from biological samples and neuroimaging on the overall performance of the study's model as well as the effect of the interactions of additional features at both population and individual levels. Secondly, further validation of the model using data from non-research settings is crucial. The dataset used in training the model is obtained from research settings, which is considered to be of high quality due to the strict data collection protocols that are used in these settings. Thirdly, the problem of imbalanced data and the ability to develop accurate prediction models that account for these problems are major challenges (Khalilia et al., 2011). However, RF and XGBoost have consistently been shown to have the capacity to handle imbalanced challenges due to the strategy employed in learning. For example, Facal et al. (2019) compared the performance of number learning algorithms,



Low-risk predicted as High-risk in PREVENT unseen test set

High-risk predicted as High-risk in PREVENT unseen test set

High-risk predicted as Low-risk in PREVENT unseen test set

Low-risk predicted as Low-risk in PREVENT unseen test set

FIGURE 6 | Force plot showing the effect of SHAP values on the interaction of features and the overall prediction at the individual level. This shows examples of prediction outputs PREVENT target models. Features in red show risk factors pushing up the overall probability while those in blue are protective factors pushing down the probability. **(A)** Shows PREVENT participant assigned to the Low-risk group but has been predicted by the model to be High-risk with 70% probability. **(B)** Shows a High-risk participant predicted as High-risk with 63% probability. **(C)** Shows a High-risk participant predicted as Low-risk with 19% probability. **(D)** Shows Low-risk participant predicted as Low-risk with 27% probability. Feature labels are: **esmoked** (0 = never smoked); **emotional disorders** (0 = no); **hypertension** (0 = no); **osteroarthritis** (1 = yes); **high cholesterol** (0 = no); **heart attack** (0 = no); **education** (2 = lower secondary education or second stage of basic education; 3 = upper secondary education); **moderate sports** (0 = hardly ever, or never, 1 = one to three times a month); **vigorous sports** (0 = hardly ever, or never, 1 = one to three times a month); **no children** (0 = no children); **widowed** (1 = yes); **BMI** [1 = under weight (<18.5)]; **married, living with spouse** (0 = no) and **gender** (1 = male).

including RF and XGBoost, to predict mild cognitive impairment to dementia conversion with highly skewed class distribution, and XGBoost demonstrated superior performance over the rest of the algorithms and outperforming RF, which is consistent with the study's findings. Nevertheless, the study's model may benefit from incorporating some of the numerous imbalanced data techniques discussed by Fernández et al. (2018) in the processing pipeline as part of future work. Lastly, all missing data were removed from the training set as part of the pre-processing step, which may have led to loss of data. This approach is not ideal and sub-optimal particularly when dealing with longitudinal datasets with long follow-up periods as well as real-world datasets, which mostly have a high prevalence of missing data. Therefore, approaches to handling missing data such as those described by Buck (1960) could potentially be explored.

Even though the study's source model achieved a relatively good performance, the performance of our target model could be better. The 63% AUROC score and a transfer learning efficacy rate of 20% achieved by the study's target model could be attributed to the limited sample used to update the decision boundaries of the study's source model. This could be considered a limitation, and therefore a bigger sample size will be required to further update and evaluate the model.

CONCLUSION

Drawing on the transfer learning paradigm of artificial intelligence, we developed ensemble-based models capable of predicting Alzheimer's dementia onset in a relatively younger population up to 14 years in advance of the mean in the training set with promising results. The models not only predict dementia risk but also provide a visualisation of the interactions between risk factors to determine those driving the risk prediction at the individual level. The complex nature of dementia requires powerful machine learning models to be able to learn complex patterns from the interactions between risk factors, and the study's proposed model achieves this with reasonable accuracy. While some of the risk factors identified are well-documented, our model further identified less suspected risk factors that appear to be significant in driving the risk of AD. We believe that with further development and validation, our prediction model has the potential to support the early detection for appropriate interventions to be developed to prevent dementia.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analysed in this study. This data can be found at: DOIs: 10.6103/SHARE.w1.710, 10.6103/SHARE.w2.710, 10.6103/SHARE.w3.710, 10.6103/SHARE.w4.710, 10.6103/SHARE.w5.710, 10.6103/SHARE.w6.710, 10.6103/SHARE.w7.710. PREVENT WL210v1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by PREVENT Dementia Programme Consortium. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SD designed the study. SD and ZZ carried out the experiments and analysed the results. All authors contributed to drafting the manuscript.

FUNDING

This work was supported by The PREVENT dementia programme funded by grants from the UK Alzheimer's Society (Grant Numbers: 178 and 264), the U.S. Alzheimer's Association (Grant Number: TriBEKa-17–519007), and philanthropic donations.

ACKNOWLEDGMENTS

This paper used data from SHARE Waves 1, 2, 3, 4, 5, 6, and 7 (DOIs: 10.6103/SHARE.w1.710, 10.6103/SHARE.w2.710, 10.6103/SHARE.w3.710, 10.6103/SHARE.w4.710, and 10.6103/SHARE.w5.710, 10.6103/SHARE.w6.710, 10.6103/SHARE.w7.710) (see Börsch-Supan et al., 2013 for methodological details). The SHARE data collection has been funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646), and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782) and by DG Employment, Social Affairs & Inclusion. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Ageing (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C), and from various national funding sources is gratefully acknowledged (see www.share-project.org).

REFERENCES

Adam, H., Dyer, A. H., Murphy, C., Lawlor, B., Kennelly, S. P., Segurado, R., et al. (2020). Cognitive outcomes of long-term benzodiazepine and related drug (BDZR) use in people living with mild to moderate Alzheimer's disease: results from NILVAD. *J. Am. Med. Direct. Assoc.* 21, 194–200. doi: 10.1016/j.jamda.2019.08.006

Barnes, D. E., Covinsky, K. E., Whitmer, R. A., Kuller, L. H., Lopez, O. L., and Yaffe, K. (2009). Predicting risk of dementia in older adults: the late-life dementia risk index. *Neurology* 73, 173–179. doi: 10.1212/WNL.0b013e3181a81636

Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Machine Learn. Res.* 13, 281–305.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., et al. (2013). Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *Int. J. Epidemiol.* 42, 992–1001. doi: 10.1093/ije/dyt088

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Stat. Soci. B* 22, 302–306.

Caruana, R., Lou Y., Gehrke J., Koch P., Sturm M., and Elhadad, N. (2015). "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney), 1721–1730. doi: 10.1145/2783258.2788613

Cui, R., Liu, M., and Alzheimer's Disease Neuroimaging Initiative (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computer. Med. Imaging Graph.* 73, 1–10. doi: 10.1016/j.compmedimag.2019.01.005

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Efron, B., and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Facal, D., Valladares-Rodriguez, S., Lojo-Seoane, C., Pereiro, A. X., Anido-Rifon, L., and Juncos-Rabadán, O. (2019). Machine learning approaches to studying the role of cognitive reserve in conversion from mild cognitive impairment to dementia. *Int. J. Geriatr. Psychiatry* 34, 941–949. doi: 10.1002/gps.5090

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning From Imbalanced Data Sets*. Berlin: Springer.

Frisoni, G. B., Molinuevo, J. L., Altomare, D., Carrera, E., Barkhof, F., Berkhof, J., et al. (2020). Precision prevention of Alzheimer's and other dementias: Anticipating future needs in the control of risk factors and implementation of disease-modifying therapies. *Alzheimer's Dement*. 16, 1457–1468. doi: 10.1002/alz.12132

Gaugler, J., James, B., Johnson, T., Marin, A., and Weuve, J. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers Dementia* 15, 321–387. doi: 10.1016/j.jalz.2019.01.010

Goerdten, J., Cukić, I., Danso, S. O., Carrière, I., and Muniz-Terrera, G. (2019). Statistical methods for dementia risk prediction and recommendations for future work: a systematic review. *Alzheimer Dementia Transl. Res. Clin. Intervent*. 5, 563–569. doi: 10.1016/j.trci.2019.08.001

Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. (2015). A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* 7, 299–309. doi: 10.1126/scitranslmed.aab3719

Houssami, N., Lee, C. I., Buist, D. S., and Tao, D. (2017). Artificial intelligence for breast cancer screening: opportunity or hype? *Breast* 36, 31–33. doi: 10.1016/j.breast.2017.09.003

Johnson, D. K., Storandt, M., Morris, J. C., and Galvin, J. E. (2009). Longitudinal study of the transition from healthy aging to Alzheimer disease. *Arch. Neurol.* 66, 1254–1259. doi: 10.1001/archneurol.2009.158

Khalilia, M., Chakraborty, S., and Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Informatics Decision Making* 11:51. doi: 10.1186/1472-6947-11-51

Kim, M. J., Kang, D. K., and Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems Appl.* 42, 1074–1082. doi: 10.1016/j.eswa.2014.08.025

Lee, S., Zhou, X., Gao, Y., Vardarajan, B., Reyes-Dumeyer, D., Rajan, K. B., et al. (2018). Episodic memory performance in a multi-ethnic longitudinal study of 13,037 elderly. *PLoS ONE* 13:e0206803. doi: 10.1371/journal.pone.0206803

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., et al. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet* 396, 413–446. doi: 10.1016/S0140-6736(20)30367-6

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., et al. (2017). Dementia prevention, intervention, and care. *Lancet* 390, 2673–2734. doi: 10.1016/S0140-6736(17)31363-6

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* 2, 749–760. doi: 10.1038/s41551-018-0304-0

Lyketsos, C. G., Lopez, O., Jones, B., Fitzpatrick, A. L., Breitner, J., and DeKosky, S. (2002). Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. *JAMA* 288, 1475–1483. doi: 10.1001/jama.288.12.1475

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *J. Thoracic Oncol.* 5, 315–1316. doi: 10.1097/JTO.0b013e3181ec173d

Natekin, A., and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorob*. 7:21. doi: 10.3389/fnbot.2013.00021

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Machine Learn. Res*. 12, 2825–2830.

Pellegrini, E., Ballerini, L., Hernandez, M. D. C. V., Chappell, F. M., González-Castro, V., Anblagan, D., et al. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimer Dementia Diagnosis Assessment Dis. Monitor*. 10, 519–535. doi: 10.1016/j.dadm.2018.07.004

Pollack, I. (1970). A nonparametric procedure for evaluation of true and false positives. *Behav. Res. Methods Instrument*. 2, 155–156. doi: 10.3758/BF03209289

Prince, M., Bryce, R., and Ferri, C. (2018). *World Alzheimer Report 2011. The Benefits of Early Diagnosis and Intervention*. Alzheimer's Disease International. Available online at: www.alz.co.uk/research/WorldAlzheimerReport2011.pdf

Ritchie, C. W., and Ritchie, K. (2012). The PREVENT study: a prospective cohort study to identify mid-life biomarkers of late-onset Alzheimer's disease. *BMJ Open* 2:e001893. doi: 10.1136/bmjopen-2012-001893

Ritchie, K., Ropacki, M., Albala, B., Harrison, J., Kaye, J., Kramer, J., et al. (2017). Recommended cognitive outcomes in preclinical Alzheimer's disease: consensus statement from the European Prevention of Alzheimer's Dementia project. *Alzheimer Dementia* 13, 186–195. doi: 10.1016/j.jalz.2016.07.154

Sharp, E. S., and Gatz, M. (2011). The relationship between education and dementia an updated systematic review. *Alzheimer Dis. Assoc. Disord*. 25:289. doi: 10.1097/WAD.0b013e318211c83c

Skolariki, K., Terrera, G. M., and Danso, S. O. (2021). Predictive models for mild cognitive impairment to Alzheimer's disease conversion. *Neural Regen. Res.* 16, 1766–1767. doi: 10.4103/1673-5374.306071

Song, J., Lee, W. T., Park, K. A., and Lee, J. E. (2014). Association between risk factors for vascular dementia and adiponectin. *BioMed Res. Int*. 2014:261672. doi: 10.1155/2014/261672

Taylor, M. E., and Stone, P. (2009). Transfer learning for reinforcement learning domains: a survey. *J. Mach. Learn. Res*. 10, 1633–1685.

van Maurik, I. S., Vos, S. J., Bos, I., Bouwman, F. H., Teunissen, C. E., Scheltens, P., et al. (2019). Biomarker-based prognosis for people with mild cognitive impairment (ABIDE): a modelling study. *Lancet Neurol*. 18, 1034–1044. doi: 10.1016/S1474-4422(19)30283-2

World Health Organization (2017). *Global Action Plan on the Public Health Response to Dementia*. Geneva. 2017–2025.

Yao, Y., and Doretto, G. (2010). "Boosting for transfer learning with multiple sources," in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE), 1855–1862. doi: 10.1109/CVPR.2010.5539857

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 February 2020 | Revised: 23 April 2020 | Accepted: 11 May 2020 | Published online: 12 June 2020

DOI: 10.1002/trc2.12041

Alzheimer's & Dementia
Translational Research
& Clinical Interventions

RESEARCH ARTICLE

Comparison of Cox proportional hazards regression and generalized Cox regression models applied in dementia risk prediction

Jantje Goerdten[1] | Isabelle Carrière[2] | Graciela Muniz-Terrera[1]

[1] Edinburgh Dementia Prevention & Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

[2] INSERM, Neuropsychiatry: Epidemiological and Clinical Research, Montpellier University, Montpellier, France

Correspondence
Jantje Goerdten, Edinburgh Dementia Prevention, Kennedy Tower, Royal Edinburgh Hospital, Edinburgh, EH10 5HF, UK.
E-mail: Jantje.Goerdten@ed.ac.uk

Funding information
European Commission, Grant/Award Numbers: QLK6-CT-2001-00360, CIT5-CT-2005-028857, GA N° 261982; German Ministry of Education and Research; U.S. National Institute on Aging, Grant/Award Numbers: U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, HHSN271201300071C

Abstract

Introduction: The frequently used Cox regression applies two critical assumptions, which might not hold for all predictors. In this study, the results from a Cox regression model (CM) and a generalized Cox regression model (GCM) are compared.

Methods: Data are from the Survey of Health, Ageing and Retirement in Europe (SHARE), which includes approximately 140,000 individuals aged 50 or older followed over seven waves. CMs and GCMs are used to estimate dementia risk. The results are internally and externally validated.

Results: None of the predictors included in the analyses fulfilled the assumptions of Cox regression. Both models predict dementia moderately well (10-year risk: 0.737; 95% confidence interval [CI]: 0.699, 0.773; CM and 0.746; 95% CI: 0.710, 0.785; GCM).

Discussion: The GCM performs significantly better than the CM when comparing pseudo-R^2 and the log-likelihood. GCMs enable researcher to test the assumptions used by Cox regression independently and relax these assumptions if necessary.

KEYWORDS
Cox proportional hazards regression, dementia risk model, dementia, prediction, splines

1 | INTRODUCTION

Dementia is one of the leading causes of dependency and disability in older individuals, with no cure yet.[1,2] However, evidence from recent studies shows the protective effects of lifestyle changes (eg, healthy diet and physical activity), regardless of genetic risk, have opened opportunities for dementia risk reduction via the implementation of behavioral interventions.[3,4] Hence, the identification of individuals at high risk of developing dementia is pivotal to apply preventive programs and to inform selection into clinical trials.

Multiple dementia risk prediction models have been developed in the last decade.[5-7] However, only a few have been recommended for clinical use, largely due to their multiple methodological weaknesses. For instance, some of the methodological limitations of the models reviewed include the overreliance on one data source and lack of internal and external validation; important concerns about the analytical techniques used were also highlighted.[6,8,9] The review of Goerdten et al.[9] summarizes the analytical techniques commonly used to derive dementia risk prediction models. Cox proportional hazards regression was one of these frequently used techniques. It belongs to the class of survival models, where the time until the event of interest, for example, death or disease diagnosis, is analyzed. With Cox regression, the influence of multiple predictors on the hazard, that is, risk of death or the disease, can be modeled. But this model relies on two critical assumptions: the proportional hazards (PH) and the log-linearity (LL) of covariates. The PH assumption supposes that the ratio of hazards between two individuals remains constant over the studied period. However, in dementia studies in which the effects of risk factors are observed over two or three decades certain individual factors may be of benefit at a time and disadvantage at another time. For instance, in a recent study Ritchie et al.[10] showed that high plasma beta amyloids were associated with an increased risk in the preclinical phase only and tended to flatten out in the approach to diagnosis while performances of cognitive tests were lowered across the 10 years before diagnosis.

Published Cox regression analyses typically impose a priori the assumption that continuous covariates have a linear effect on the logarithm of the hazard. This LL assumption implies that dementia risk changes gradually with increasing value of the prognostic factor, so that, for example, the relative risk for a 60-year-old subject compared to a 50-year-old is the same as that when comparing subjects aged 80 versus 70 years. However, if the true relationship between the continuous independent variable and the outcome does not fulfil the LL assumption, then the conventional log-linear model may result in incorrect identification of high-risk subgroups and biased prognosis.

In this article, we use generalized Cox regression models, which can incorporate non-linear and/or time-dependent effects of variables to model dementia risk.[11] To demonstrate the benefits of this modeling approach for dementia risk prediction, we compare results obtained from this methodology to results obtained from Cox regression, which is used frequently in the field.[9]

2 | METHODS

2.1 | Study population

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database with data collected on health, socio-economic status, and social and family networks. SHARE comprises approximately 140,000 participants aged 50 and older from 27 European countries and Israel. Follow-up of respondents was carried out in waves (Wave 1 to 7). SHARE was described elsewhere in more detail.[12] We use information from Wave 2 to 7,[13-18] as from Wave 2 forward the information regarding dementia diagnosis was collected from respondents aged 60 and older. Wave 3 was not included, as it focused on the childhood of respondents.[14] In SHARE participants with only baseline measures, a dementia diagnosis at baseline and/or missing information for the predictor variables were excluded, which resulted in a cohort of 11,603 participants.

HIGHLIGHTS

- The frequently used Cox regression employs two crucial assumptions, which might not hold for all predictor variables, and can lead to incorrect predictions of dementia risk.
- Generalized Cox regression can relax the assumptions made by Cox regression.
- Generalized Cox regression performs better than Cox regression in predicting dementia risk.
- Generalized Cox regression is an interesting extension of Cox regression, and should be used more frequently in dementia risk research.

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed the literature using traditional sources (PubMed) and references from previous publications.
2. Interpretation: The presented findings show the improvements made through the incorporation of splines in the model, and the relaxation of the assumptions used by Cox regression. Importantly, none of the continuous predictor variables obeyed the crucial PH assumption. Generalized Cox regression enables researchers to test the assumptions independently and relax the assumptions of Cox regression if necessary.
3. Future directions: We would like to encourage researchers to adapt the use of splines in dementia risk prediction research.

2.2 | External validation sample

The Aging, Demographics, and Memory Study (ADAMS) is a supplementary study of the Health and Retirement Study (HRS).[19] The HRS is a longitudinal panel study, looking into the changing health and economic circumstances of adults over age 50 in the United States. In ADAMS, in-person clinical assessments were conducted to gather information on the cognitive status of the participants over four waves (Wave A to D). Participants are aged 70 and older. The design and methods of ADAMS are described elsewhere in more detail.[20]

2.3 | Assessment of dementia and predictors

Dementia diagnosis was recorded by self-report in SHARE. The participants were asked if a doctor ever diagnosed them/told them they have Alzheimer's disease, dementia, or senility.[21-25]

To have a close and in-depth look at the variables selected as predictors, we chose to focus on modifiable risk factors identified by Livingston et al.[2] and age. We selected age, years of education, body mass index (BMI), hearing loss, high blood pressure, smoking status, depression, physical activity, and diabetes. The information regarding disease status and behavioral risk were collected by self-report.[21] BMI was calculated from height and weight reported by the participants. Hearing was recorded as "excellent," "very good," "good," "fair," and "poor." It was categorized into 0/1, where "excellent" to "good" was coded as 0 and "fair" to "poor" as 1. For the diagnoses of high blood pressure and diabetes the participants were asked if a doctor ever told them they have high blood pressure/hypertension or high glucose level/diabetes. For the diagnosis of depression, the participants were asked if they suffered ever/since last wave from symptoms of depression which lasted at least 2 weeks. Physical activity was recorded as "more than once a week," "once a week," "one to three times a month," and "hardly ever, or never." It was categorized into 0/1, where "more than once a week" to "one to three times a month" was coded as 0 and "hardly ever, or never" as 1.

2.4 | Generalized Cox regression

Cox proportional hazard regression is commonly used to model censored survival data. The purpose of the Cox proportional hazards regression model (CM) is to model the simultaneous effect of multiple factors on the survival.[26] The CM aims to estimate hazard ratios over time.[26] The model equation is written as follows:

$$h(t|z_1, \dots, z_p) = h_0(t)\exp\left(\sum_i \beta_i z_i\right)$$

where $(z_i)_{i=1,\dots,p}$ are the values of the covariates $Z_1, \dots, Z_p$ on which the hazard may depend and $h_0(t)$ represents the baseline hazard. The baseline hazard is defined as the value of the hazard when $z_i = 0$, for i in 1, p.

In this study, three flexible models proposed by Mahboubi et al.[27] were used, which are generalizations of the CM. With these flexible models, one or both assumptions used by Cox regression can be relaxed and tested independently. Cox regression employs the PH and LL assumption. With the generalized Cox regression model (GCM) it is possible to model time dependent hazard ratios and/ or non-linear effects of the predictor variables.

The first flexible model relaxes the proportional hazards assumption (NPH):

$$h(t|z_1, \dots, z_p) = h_0(t)\exp\left(\sum_i \beta_i(t)\, z_i\right)$$

The second flexible model relaxes the log-linearity assumption (NLL):

$$h(t|z_1, \dots, z_p) = h_0(t)\exp\left(\sum_i r_i(z_i)\right)$$

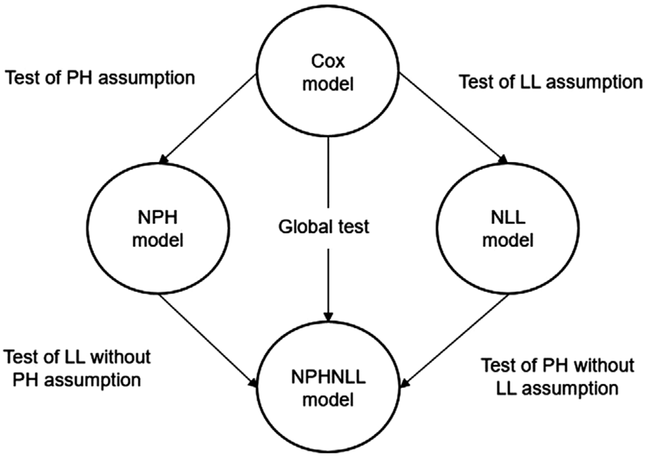


FIGURE 1 Testing of assumptions and finding best model. Arrows represent likelihood ratio test. Comparing models by likelihood ratio tests the assumptions of proportional hazards (PH) and log-linearity (LL), and the best fitting model for the predictor is identified. This figure is adapted from Mahboubi et al[27]

Last, the third flexible model relaxes both assumptions simultaneously (NPHNLL):

$$h(t|z_1, \dots, z_p) = h_0(t)\exp\left(\sum_i \beta_i(t)\, r_i(z_i)\right)$$

The function r_i is a spline function of z_i modeling the non log-linear effect of z_i and $\beta i(t)$ is a spline function of t modeling the time dependent effect of z_i. Estimations of these functions are based on the full likelihood.

The flexible models use B-splines, which are piecewise polynomials, where the pieces are joint by knots. Here, the splines are allowed to have one or two knots. The knot selection has to follow one criterion: there must be roughly the same number of events in the subintervals defined by the selected knots. The decision if one or two knots are used is based on a goodness of fit test. For example, models with one and two knots are computed and compared in terms of the Akaike information criterion (AIC). The model that produces the smallest AIC is selected. It can be tested if a variable obeys the assumptions by comparing the models described before using likelihood ratio tests (see Figure 1)[27] and by this deciding which of the four models (CM, NPH, NLL, NPHNLL) models the variable best.

2.5 | Statistical analyses

A CM and a GCM were fitted to data from SHARE to predict dementia risk. Study time was used as time scale for all analyses. Study time was calculated from study entry (Wave 2, 2007) until study exit—wave of dementia diagnosis, wave in which participant died, wave in which participant was lost to follow-up, or the end of the study (Wave 7, 2017), whichever came first. In the survival time analyses, dementia diagnosis was treated as the failure event.

To compute the full GCM, first, we tested if each predictor variable complied with the PH assumption and/or the LL assumption. To test these assumptions each predictor variable was modeled in a CM, an NPH model, an NLL model, and in an NPHNLL model. The computed models were compared by likelihood ratio test, and the best fitting model for each predictor variable was selected. All predictor variables were entered into the full model, while modeling each predictor with the best-identified knot and spline combination. Last, after fitting the model with all identified splines and knots, spline coefficients were eliminated systematically. We reduced spline coefficients if more than one coefficient was non-significant for a predictor, while comparing the smaller model with the previous one by likelihood ratio test—until the best fitting model was found. For the full CM, all predictor variables were entered into the model. To determine which model fits the data better, the model derived from Cox regression or generalized Cox regression, likelihood ratio tests were performed and the computed pseudo-R^2 proposed by Nagelkerke and Cragg and Uhler were compared.[26] C-statistics adapted for survival analyses were calculated to assess predictive ability.[26] The C-statistic is a discrimination measure for binary outcomes, and it ranges from below 0.5 (indicating very poor model discrimination) to 1 (indicating perfect model discrimination). Bootstrapping with 1000 repetitions was performed to compute 95% confidence intervals (CI) for the C-statistics and the pseudo-R^2.

SHARE was used as the development sample and ADAMS as the external validation sample.

All analyses were performed in R Studio (Version 3.5.1)[28] and the packages flexrsurv,[29] survival,[30,31] Hmisc,[32] and ggplot2[33] were used.

3 | RESULTS

Among the 11,603 SHARE participants, 757 (6.5%) reported that they had received a diagnosis of dementia during 10 years of follow-up. The mean age of diagnosis was 75.4 (7.2 standard deviation [SD]). Baseline characteristics for SHARE and ADAMS are presented in Table 1.

TABLE 1 Baseline characteristics of SHARE and ADAMS

	SHARE N = 11,603	ADAMS N = 410
Dementia (%)	757 (6.5)	102 (24.9)
Age mean (SD[a])	69.7 (7.2)	79.1 (6.1)
Years of education (SD)	10.2 (4.4)	10.71 (4.3)
Body mass index (SD)	26.7 (4.2)	26.9 (4.9)
Sex (%)		
Female	6283 (54.1)	210 (51.2)
Male	5320 (45.9)	200 (48.8)
Depression (%)	1866 (16.1)	107 (26.1)
Diabetes (%)	1354 (11.7)	86 (20.98)
High blood pressure (%)	4647 (40.1)	257 (62.7)
Poor hearing (%)	2464 (21.2)	122 (29.8)
Ever smoker (%)	1532 (13.2)	117 (28.5)
No physical activity (%)	5424 (46.8)	257 (62.7)

[a] Standard deviation (SD).
Abbreviations: ADAMS, Aging, Demographics, and Memory Study; SHARE, Survey of Health, Ageing and Retirement in Europe.

In SHARE none of the variables obeyed the PH assumption, when modeled alone (crude model). Two (years of education and BMI) of three continuous variables additionally did not obey the LL assumption. Comparisons of the estimated log hazards of dementia risk for age, years of education, and BMI in SHARE from the crude CMs (Figure 2 parts A, C, E) and GCMs(Figure 2 parts B, D, F) are presented in Figure 2.

The following section discusses the full prediction model derived from Cox regression and generalized Cox regression; both include the same predictor variables (age, years of education, BMI, depression, diabetes, high blood pressure, hearing, smoking status, and physical activity). In the full GCM, age, years of education, and BMI were modeled non-proportional with time (NPH). When comparing the CM and GCM in terms of the log-likelihood, the test results in a *P*-value of <.001. The pseudo-R^2 for the CM is 0.06 (95% CI: 0.048, 0.062) and for the GCM 0.493 (95% CI: 0.460, 0.506). The C-statistic for the predicted 10-year dementia risk is 0.737 (95% CI: 0.699, 0.773; CM) and 0.746 (95% CI: 0.710, 0.785; GCM). The C-statistic for the predicted 4-year dementia risk is 0.711 (95% CI: 0.678, 0.74; CM) and 0.709 (95% CI: 0.673, 0.74; GCM). Within ADAMS the two models generate a C-statistic for the predicted 6-year dementia risk of 0.743 (95% CI: 0.58, 0.924; CM) and 0.764 (95% CI: 0.607, 0.952; GCM). All computed C-statistics for the time points from the models are presented in Table 2.

The regression coefficients computed by CM and GCM from SHARE are presented in Appendix A in supporting information. The computed overall C-statistics for the CM and GCM in SHARE and ADAMS are presented in Appendix B in supporting information.

4 | DISCUSSION

In this study, we compared dementia risk prediction models derived from generalized Cox regression and Cox regression. Our results show that the model derived from the generalized Cox regression fits the data significantly better than the model derived from Cox regression. The predictive ability of the CM and GCM range from moderate to good.

4.1 | Cox regression versus generalized Cox regression

The GCM performs in the development sample and in the validation sample better than the CM. Both GCM and CM reach moderate to good predictive ability, which is in line with previous dementia risk prediction models.[7]

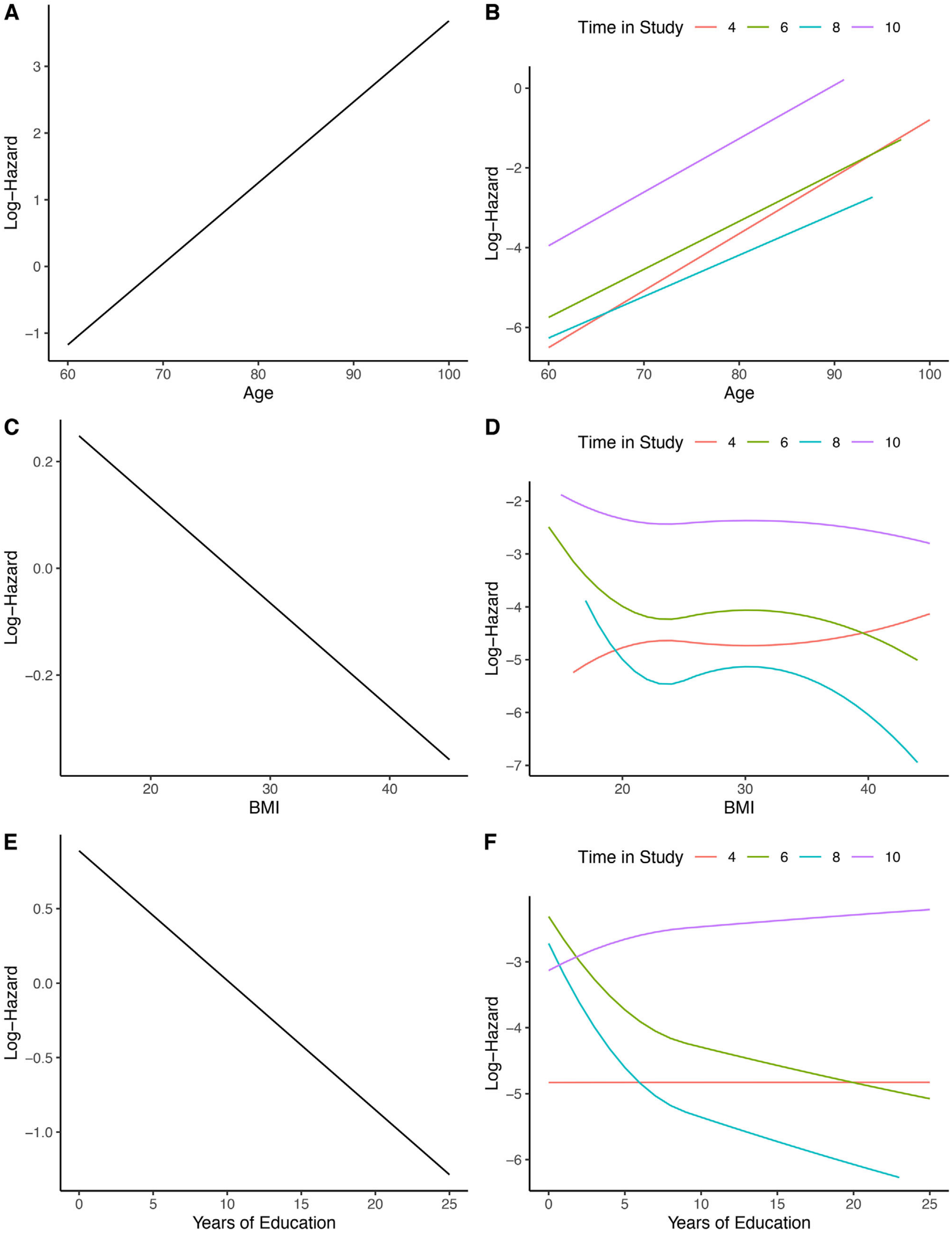


FIGURE 2 Estimated log-hazards from crude Cox models (CMs) and generalized Cox models. Graphs A, C, and E show estimated log-hazards for age, body mass index (BMI), and years of education from crude CMs; graphs B, D, and F show estimated log-hazards, for each follow-up time point, for age from a crude non proportional hazards model (NPH), BMI, and years of education from crude non proportional hazards and non log linear models (NPHNLL)

TABLE 2 C-statistics for SHARE models

	Number of cases	Cox regressionC-statistic (95% CI)	Generalized Cox regressionC-statistic (95% CI)
In SHARE			
10 years	177	0.737 (0.699, 0.773)	0.746 (0.710, 0.785)
8 years	173	0.658 (0.616, 0.699)	0.659 (0.616, 0.698)
6 years	150	0.735 (0.693, 0.773)	0.736 (0.695, 0.775)
4 years	257	0.711 (0.678, 0.74)	0.709 (0.673, 0.74)
In ADAMS			
≥6 years	13	0.747 (0.601, 0.917)	0.805 (0.695, 0.942)
6 years	10	0.743 (0.58, 0.924)	0.764 (0.607, 0.952)
5 years	23	0.51 (0.367, 0.652)	0.517 (0.368, 0.659)
4 years	23	0.592 (0.436, 0.768)	0.589 (0.430, 0.775)
3 years	4	1.0	1.0
2 years	16	0.708 (0.558, 0.869)	0.708 (0.555, 0.865)
1 years	23	0.602 (0.45, 0.765)	0.62 (0.468, 0.795)

Abbreviations: ADAMS, Aging, Demographics, and Memory Study; CI, confidence interval; SHARE, Survey of Health, Ageing and Retirement in Europe.

The overall estimated C statistic for SHARE and ADAMS from the GCM shows an interesting problem: the C-statistic is lower than 0.5, which would mean the model performs worse than chance (see Appendix B). However, this is not the case when looking at the estimated C-statistics for the follow-up time points. The C-statistic is a rank correlation test; a high C-statistic translates to a model which is able to estimate higher risks for individuals experiencing the outcome than individuals who did not during follow-up.[26] In this case—in which we relaxed the PH assumption for all three continuous predictors—the C-statistic test is not able to rank the estimated risks correctly, because the GCM estimates time dependent risks. The overall risk of individuals who had a follow-up of 10 years is higher than for example of individuals who had a follow-up of 4 years, regardless of dementia risk (Appendix C in supporting information). Hence, the test ranks all individuals who had a follow-up of 10 years over individuals with a follow-up of 4 years, which results in an incorrect low C-statistic. It might be useful to evaluate in which time frame a dementia risk prediction model derived from a GCM performs best.

Additionally, it needs to be mentioned that the C-statistic or area under the receiver operating characteristic curve (AUROC) is not recommended to compare models, as it is a low power procedure.[34] They should only be used to describe the predictive ability of a model. Instead, a high-power test should be carried out to asses which model fits the data better, for example, a likelihood ratio test and/or comparing R^2. In this study the likelihood ratio test suggests the GCM fits the data significantly better than the CM. The pseudo R^2 suggests that the GCM improves greater upon the null model than the CM and hence is better able to predict the outcome than the CM. When looking at the results from the likelihood ratio test and the pseudo R^2 we can conclude that the GCM performs better than the CM in modeling dementia risk in SHARE.

4.2 | Improvements by generalization

As summarized by Goerdten et al.,[9] most published dementia risk prediction studies overlook the fulfilment of the assumptions of the analytical technique used for the estimation of risk. Consistent testing of these assumptions is crucial, as their violation can lead to biased results.[35] This is especially important for continuous variables (eg, age) as shown in our work. This problem might lead researchers to categorize continuous variables, a practice that in turn leads to information loss and residual confounding.[36] Instead, Moons et al.[37] recommend the incorporation of splines, if there are any uncertainties about whether a variable complies with the linearity assumption, as the incorporation of splines makes the categorization of continuous variables unnecessary.

In this study we incorporated splines to test and relax the two strong assumptions used by Cox regression: (1) assumption of LL, that is, a linear relationship between the independent variable and the log-hazard of dementia and (2) assumption of PH, that is, the effect of a variable is constant over time. There are other (simpler) options to assess the PH assumption of Cox regression: an interaction term with time can be added to the model or stratification by time can be performed. But using simpler testing methods implies assuming the LL assumption while testing the PH assumption and assuming the PH assumption while testing the LL assumption. The GCM allows us to test both assumptions of Cox regression independently from each other.

None of the predictor variables included in our analyses fulfilled the PH assumption. Furthermore, two of the continuous variables did not fulfil the LL assumption either. Comparing the estimated log-hazards for the three continuous variables (age, BMI, and years of education) from crude CMs and GCMs, the difference between the models becomes evident. While the CM computes linear declining or increasing log-hazards for the continuous variables, GCM computes a great variety of curves (see Figure 2). For age the PH assumption was relaxed, hence the effect of this variable on dementia risk is not constant with time and the different lines for 4 to 10 years can be seen. For BMI and years of education additionally the LL assumption was relaxed, hence the effects of the variables are not constant with time and there are non-linear relationships between the variables and the log-hazard of dementia, and the different lines with curves for 4 to 10 years can be seen.

Comparing the presented methodology with for example the CAIDE score[38]—a well-known dementia risk prediction score, computed by logistic regression, that ignores the dependence on time of the event being modeled—the applied approach could in theory model more accurately dementia risk. The CAIDE score translates to risk percentages ranging from 1% (low risk) to 16.4% (high risk). The difference to a risk model derived from GCM would be that the prediction model could inform if this risk changes over time, as the effects of some or all predictors on dementia risk change with time. The generalized Cox regression is more flexible and able to pick up changes in the effect of a predictor variable on dementia risk over time.

4.3 | Strengths and limitations

This study has several limitations. First, due to the design of the used datasets, interval censoring is present. This means that the exact date of diagnosis is not known and occurred at some point during the interval between the waves. This might have resulted in biased results, likely an overestimation of the predictor coefficients.[39] Second, censoring due to death, which is a competing event, was not taken into account. There are existing methods to incorporate competing risks in survival analyses[40] as well as generalizations of these models.[41] However, the information on death in SHARE are recorded by proxy questionnaire and the use of these information might have hampered the results even further.[21] Third, the quality of the data about dementia diagnosis in SHARE is not optimal, as it is only recorded by self-report and no further testing of the diagnosis is made. A similar limitation of the data is that the predictor variables were also self-reported. However, for the purpose of this paper, these limitations are not critical given the aims of our work.

This study has several strengths. SHARE offers a large sample size, which covers a wide range of European countries and Israel, making it representative of the European population.[12] In ADAMS the diagnosis of dementia was made by professionals. Every predictor variable was tested for the assumptions used by Cox regression. Importantly, following recommended practice, the developed models were validated internally and externally.

5 | CONCLUSION

With the generalized Cox regression, the assumptions of Cox regression can be tested thoroughly and independently, and relaxed if needed. However, while the generalized Cox regression offers advantages, such as avoiding categorization, the disadvantages need to be mentioned too: the flexible models can require long computation times and a bigger sample is needed than for a Cox regression. Additionally, the interpretation of the coefficients computed by GCMs are not straightforward and it is only possible to examine the effect of a variable visually. Taking all this into account the generalized Cox regression is an interesting option to extend a Cox regression. The possibility to add splines and herewith relax the assumptions is especially appealing when including continuous variables. We would like to encourage researchers to adapt the use of splines in dementia research, to increase the understanding of the relationship between potential predictors and dementia risk.

ACKNOWLEDGMENTS

SHARE: This paper uses data from SHARE Waves 2, 4, 5, 6, and 7 (10.6103/SHARE.w2.700 https://doi.org/10.6103/SHARE.w3.700https://doi.org/10.6103/SHARE.w4.700https://doi.org/10.6103/SHARE.w5.700https://doi.org/10.6103/SHARE.w6.700https://doi.org/10.6103/SHARE.w7.700), see Börsch-Supan et al.[12] for methodological details.

The SHARE data collection has been funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982) and Horizon 2020 (SHARE-DEV3: GA N°676536, SERISS: GA N°654221) and by DG Employment, Social Affairs & Inclusion. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C), and from various national funding sources is gratefully acknowledged (see www.share-project.org).

HRS: The Health and Retirement Study is a longitudinal project sponsored by the National Institute on Aging (NIA U01AG009740) and the Social Security Administration. The study director is Dr. David R. Weir of the Survey Research Center at the University of Michigan's Institute for Social Research.

CONFLICTS OF INTEREST

The authors report no conflicts of interest.

REFERENCES

1. Robinson L, Tang E, Taylor J-P. Dementia: timely diagnosis and early intervention. *BMJ*. 2015;350:h3029.
2. Livingston G, Sommerlad A, Orgeta V, et al. Dementia prevention, intervention, and care. *Lancet*. 2017;390:2673-2734.
3. Lourida I, Hannon E, Littlejohns TJ, et al. Association of lifestyle and genetic risk with incidence of dementia. *JAMA*. 2019.
4. Ngandu T, Lehtisalo J, Solomon A, et al. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. *Lancet North Am Ed*. 2015;385:2255-2263.
5. Stephan BC, Kurth T, Matthews FE, Brayne C, Dufouil C. Dementia risk prediction in the population: are screening models accurate. *Nat Rev Neurol*. 2010;6:318-326.
6. Tang EY, Harrison SL, Errington L, et al. Current developments in dementia risk prediction modelling: an updated systematic review. *PLoS One*. 2015;10:e0136181.
7. Hou XH, Feng L, Zhang C, Cao XP, Tan L, Yu JT. Models for predicting risk of dementia: a systematic review. *J Neurol Neurosurg Psychiatry*. 2019;90(4):373-379.
8. Pellegrini E, Ballerini L, Hernandez M, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimers Dement (Amst)*. 2018;10:519-535.
9. Goerdten J, Čukić I, Danso SO, Carrière I, Muniz-Terrera G. Statistical methods for dementia risk prediction and recommendations for future work: a systematic review. *Alzheimers Dement (N Y)*. 2019;5:563-569.
10. Ritchie K, Carrière I, Berr C, et al. The clinical picture of Alzheimer's disease in the decade before diagnosis: clinical and biomarker trajectories. *J Clin Psychiatry*. 2016;77:e305-11.
11. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med*. 2007;26:392-408.
12. Börsch-Supan A, Brandt M, Hunkler C, et al, Team obotSCC. Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *Int J Epidemiol*. 2013;42:992-1001.

13. Börsch-Supan A, Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2. Release version: 7.0.0. SHARE-ERIC. Data set. 2019. https://doi.org/10.6103/SHARE.w2.700.
14. Börsch-Supan A, Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 3—SHARELIFE. Release version: 7.0.0. SHARE-ERIC. Data set. 2019. https://doi.org/10.6103/SHARE.w3.700.
15. Börsch-Supan A, Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 7.0.0. SHARE-ERIC. Data set. 2019. https://doi.org/10.6103/SHARE.w4.700.
16. Börsch-Supan A, Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release version: 7.0.0. SHARE-ERIC. Data set. 2019. https://doi.org/10.6103/SHARE.w5.700.
17. Börsch-Supan A, Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6. Release version: 7.0.0. SHARE-ERIC. Data set. 2019. https://doi.org/10.6103/SHARE.w6.700.
18. Börsch-Supan A, Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7. Release version: 7.0.0. SHARE-ERIC. Data set. 2019. https://doi.org/10.6103/SHARE.w7.700.
19. Amanda S, Jessica DF, Mary Beth O, Kenneth ML, John WRP, David RW. Cohort profile: the Health and Retirement Study (HRS). *Int J Epidemiol*. 2014;43(2):576-585.
20. Heeringa, SG, Fisher, GG, Hurd, M, Langa, KM, Ofstedal, MB, Plassman, BL, Rodgers, WL, & Weir, DR , , etal. *Aging, Demographics and Memory Study (ADAMS): Sample Design, Weighting and Analysis for ADAMS*. Ann Arbor, MI: Institute for Social Research, University of Michigan; 2009.
21. Börsch-Supan A, Brugiavini A, Jürges H, et al. *First Results from the Survey of Health, Ageing, and Retirement in Europe (2004-2007): Starting the Longitudinal Dimension*. Mannheim: Mannheim Research Institute for the Economics of Aging; 2008.
22. Malter F, Börsch-Supan A. *SHARE Wave 4: Innovations & Methodology*. Munich: Mannheim Research Institute for the Economics of Aging, Max Planck Institute for Social Law and Social Policy; 2013.
23. Malter F, Börsch-Supan A. *SHARE Wave 5: Innovations & Methodology*. Munich: Mannheim Research Institute for the Economics of Aging, Max Planck Institute for Social Law and Social Policy; 2015.
24. Malter F, Börsch-Supan A. *SHARE Wave 6: Panel innovations and collecting Dried Blood Spots*. Munich: Mannheim Research Institute for the Economics of Aging, Max Planck Institute for Social Law and Social Policy; 2017.
25. Bergmann M, Scherpenzeel A, Börsch-Supan A. *SHARE Wave 7 Methodology: Panel Innovations and Life Histories*. Munich: Mannheim Research Institute for the Economics of Aging, Max Planck Institute for Social Law and Social Policy; 2019.
26. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: New York: Springer; 2001.
27. Mahboubi A, Abrahamowicz M, Giorgi R, Binquet C, Bonithon-Kopp C, Quantin C. Flexible modeling of the effects of continuous prognostic factors in relative survival. *Stat Med*. 2011;30:1351-1365.
28. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017. https://www.R-project.org/.
29. Clerc-Urmès I, Grzebyk, Hédelin M, G group Cws. flexrsurv: An R package for relative survival analysis. 2017.
30. Therneau MT, Grambsch MP. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
31. Therneau MT, A Package for Survival Analysis in S. 2015. https://CRAN.R-project.org/package=survival.
32. Harrell FE, Jr. Hmisc: Harrell Miscellaneous. 2018. R package version 4.2-0.2019. https://CRAN.R-project.org/package=Hmisc.
33. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag; 2016.
34. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115:928-935.
35. Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. *Am J Epidemiol*. 1997;145:714-729.
36. Shepherd BE, Rebeiro PF, Caribbean C. South America network for HIVe. Brief Report: assessing and interpreting the association between continuous covariates and outcomes in observational studies of HIV using splines. *J Acquir Immune Defic Syndr*. 2017;74: e60-e3.
37. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98:683.
38. 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;5:735-741.
39. Radke BR. A demonstration of interval-censored survival analysis. *Prev Vet Med*. 2003;59:241-256.
40. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol*. 2009;170:244-256.
41. Boracchi P, Biganzoli E, Marubini E. Joint modelling of cause-specific hazard functions with cubic splines: an application to a large series of breast cancer patients. *Comput Stat Data Anal*. 2003;42:243-262.

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: Goerdten J, Carrière I, Muniz-Terrera G. Comparison of Cox proportional hazards regression and generalized Cox regression models applied in dementia risk prediction. *Alzheimer's Dement*. 2020;6:1–8. https://doi.org/10.1002/trc2.12041