Journal of Alzheimer's Disease 75 (2020) 717–728
DOI 10.3233/JAD-190967
IOS Press

Predicting Cognitive Impairment and Dementia: A Machine Learning Approach

Damaris Aschwanden[a,*], Stephen Aichele[b,c], Paolo Ghisletta[b,d,e], Antonio Terracciano[a], Matthias Kliegel[b,e], Angelina R. Sutin[a], Justin Brown[a] and Mathias Allemand[f,g]
[a]*Florida State University, Tallahassee, FL, USA*
[b]*Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland*
[c]*Colorado State University, Fort Collins, CO, USA*
[d]*Swiss Distance University Institute, Switzerland*
[e]*Swiss National Centre of Competence in Research LIVES – Overcoming Vulnerability: Life Course Perspectives, Universities of Lausanne and of Geneva, Switzerland*
[f]*University of Zurich, Zurich, Switzerland*
[g]*University Research Priority Program Dynamics of Healthy Aging, University of Zurich, Switzerland*

Handling Associate Editor: Kay Deckers

Accepted 13 March 2020

Abstract.
Background: Efforts to identify important risk factors for cognitive impairment and dementia have to date mostly relied on meta-analytic strategies. A comprehensive empirical evaluation of these risk factors within a single study is currently lacking.
Objective: We used a combined methodology of machine learning and semi-parametric survival analysis to estimate the relative importance of 52 predictors in forecasting cognitive impairment and dementia in a large, population-representative sample of older adults.
Methods: Participants from the Health and Retirement Study (N = 9,979; aged 50–98 years) were followed for up to 10 years ($M = 6.85$ for cognitive impairment; $M = 7.67$ for dementia). Using a split-sample methodology, we first estimated the relative importance of predictors using machine learning (random forest survival analysis), and we then used semi-parametric survival analysis (Cox proportional hazards) to estimate effect sizes for the most important variables.
Results: African Americans and individuals who scored high on emotional distress were at relatively highest risk for developing cognitive impairment and dementia. Sociodemographic (lower education, Hispanic ethnicity) and health variables (worse subjective health, increasing BMI) were comparatively strong predictors for cognitive impairment. Cardiovascular factors (e.g., smoking, physical inactivity) and polygenic scores (with and without *APOE* ε4) appeared less important than expected. *Post-hoc* sensitivity analyses underscored the robustness of these results.
Conclusions: Higher-order factors (e.g., emotional distress, subjective health), which reflect complex interactions between various aspects of an individual, were more important than narrowly defined factors (e.g., clinical and behavioral indicators) when evaluated concurrently to predict cognitive impairment and dementia.

Keywords: Aging, cognitive impairment, Cox proportional hazard survival analysis, dementia, machine learning, protective factors, random forest survival analysis, risk factors

INTRODUCTION

In 2018, it was estimated that 50 million persons were living with dementia worldwide, with an associated annual care costs of approximately 1 trillion

*Correspondence to: Damaris Aschwanden, Department of Geriatrics, College of Medicine, Florida State University, 1115 West Call Street, Tallahassee, FL 32306, USA. E-mail: damaris.aschwanden@med.fsu.edu.

USD [1]. As such, dementia poses a major global challenge for health, well-being, and social care [2, 3]. Overcoming this challenge will require confronting considerable gaps in our understanding of the multifactorial etiology of dementia and the lack of a disease-modifying treatment. Even if effective treatments are developed, prevention and risk reduction will still be fundamental strategies [2, 4, 5]. The World Health Organization (WHO) has recently released guidelines on risk reduction of cognitive decline and dementia [3] that point to multiple predictors of disease onset and progression [2, 5–12]. Such a multifactorial etiology prompts questions about the relative importance of individual predictors to be targeted for prevention.

To date only a few empirical studies have compared the relative importance of predictors, and these approaches have mostly relied on meta-analytic methodologies. Barnes and colleagues [13], for example, developed a late-life dementia risk index that identified age and cognitive performance as the most salient predictive factors. The Alzheimer's Association concluded that the strongest predictors are age, family history, apolipoprotein E (*APOE*) $\varepsilon 4$, midlife obesity, midlife hypertension, smoking, diabetes, education, and physical activity [5]. The risk ranking of a mixed-methods study included depression, midlife hypertension, physical inactivity, diabetes, midlife obesity, hyperlipidemia, and smoking [8]. The Lancet Commission on Dementia Prevention, Intervention, and Care estimated that approximately 35% of dementia is attributable to a combination of childhood education, midlife hypertension, midlife obesity, hearing loss, late-life depression, diabetes, physical inactivity, smoking, and social isolation [2].

Evidence from these prior studies shows relative consistencies among the identified risk factors (e.g., smoking, diabetes, physical inactivity, midlife hypertension) [2, 5, 8]. However, results of meta-analytic approaches are vulnerable to sources of bias related to methodological differences across the summarized research (e.g., differences in data quality and sample selection criteria, heterogeneity across instruments assumed to measure the same predictors, potential publication bias). Moreover, interactions among risk factors are notoriously difficult to test in a comprehensive way using classical parametric regression techniques. Therefore, the aggregation of results from such studies may exclude variables that act mainly through indirect pathways (e.g., via amplification of other risk factors' effects on the given outcome). A clearer picture of such associations may emerge when risk factors are concurrently evaluated using data from the same (population representative) sample. Nevertheless, we were unable to identify any such study in which multiple risk factors for cognitive impairment and dementia were comparatively tested for differences in predictive strength.

To address this limitation in the literature, we used a machine learning approach (random forest survival analysis; RFSA) to evaluate the relative importance of 52 predictors for cognitive impairment and dementia in a large, population representative sample of adults (N = 9,979; aged 50–98 years) assessed over a period of up to 10 years. RFSA is an extension of regression trees, a non-parametric statistical method in which observations are recursively partitioned to identify variables most strongly associated with the outcome of interest [14–16]. RFSA aggregates estimates of predictor-outcome strength across multiple regression trees to estimate each predictor's variable importance (VIMP) and relative importance (I_{rel}) with respect to other tested predictors. Unlike more common approaches based on regression, RFSA implicitly adjusts for all possible linear, non-linear, and higher-order interaction effects. It also provides built-in safeguards against multicollinearity (i.e., predictors so closely interrelated that their effects cannot be disentangled) and model overfit (i.e., model estimates that cannot be replicated in independent samples because they relate also to the noise present in the sample analyzed). However, RFSA was not developed within a standard probabilistic framework. Therefore, we first applied RFSA to a random half subsample ($n_1 = 4{,}990$) of the original sample, to assess the most important risk factors for cognitive impairment and dementia. Then, in the remaining random half subsample ($n_2 = 4{,}989$), we examined these most important risk predictors using Cox proportional hazards analysis (Cox PH [17]), which is better suited to interpretation of effect sizes based on a known statistical distribution. This multi-analytic (split sample) methodology allowed us to comprehensively test the relative importance of numerous, interrelated risk factors and to estimate effect sizes for the strongest predictors of cognitive impairment and dementia.

Based on the consensus of aforementioned literature, we expected health-related variables such as diabetes, physical inactivity, and smoking to be the most important risk factors for cognitive impairment and dementia risk. We also expected education [2, 5] and emotional distress [2, 8] to be implicated.

METHODS

Participants

We analyzed data from the Health and Retirement Study (HRS); a longitudinal panel study that surveys a representative sample of older adults in the United States. Data were collected by the University of Michigan and the research protocol was approved by their Institutional Review Board. Since HRS' inception, participants have completed cognitive tasks designed to measure general cognitive status every two years. Starting in 2006, HRS included a psychosocial questionnaire and an in-person interview with physical measurements. Half of the participants completed the assessment in 2006, and the other half completed it in 2008. We used the combined 2006/2008 data as baseline to predict future cognitive status by implementing an analytic design ensuring longitudinal prediction. Cognitive status from 2008 (if predictor data available in 2006), 2010, 2012, 2014, and 2016 was used as outcome. The follow-up thus ranged from 2–10 years ($M = 6.85$ years for cognitive impairment, $M = 7.67$ years for dementia).

The sample size of the current longitudinal data analysis was determined using existing HRS data. Participants who met the following criteria were included in the analyses: 1) cognitively unimpaired and aged 50+ years at baseline, 2) availability of predictor data at baseline, and 3) availability of cognitive data at least at one follow-up. Our entire sample consisted of 9,979 participants ($M_{age} = 67.01$ years, $SD = 9.18$ years; 59.8% women). Over the follow-up, 3,119 participants developed cognitive impairment (31.3%; mean age at onset = 73.86 years, $SD = 8.70$ years) and 622 participants developed dementia (6.2%; mean age at onset = 74.68 years, $SD = 8.79$ years; see below for screening procedure). Supplementary Figure 1 displays the percentage of cognitive impairment/dementia per total sample and age group across follow-ups.

Predictors

We included a comprehensive list of predictors identified from previous literature and available in HRS that included 52 variables: demographic (10), biomarkers/polygenic (7), health (26), and psychosocial (9). The assessment of predictors is described in the Supplementary Material (S1.1). Descriptive statistics are presented in Table 1.

Outcomes: Cognitive impairment and dementia

We used the Langa-Kabeto-Weir (L-K-W) algorithm [18, 19] to define cognitive status. The L-K-W algorithm applies cutoffs to derived scores summarizing cognitive data for self-respondents and both cognitive and functional data for proxy-respondents. In HRS, a proxy-respondent is sought for respondents who are unwilling or unable to answer to an interview themselves. Proxies are usually the spouse or another close family member [20]. Proxy interviews are essential to maintaining coverage of the cognitively impaired [21]. For self-respondents, the summary score is cognitive test performance on the modified Telephone Interview for Cognitive Status (TICSm) [22]. The total TICSm score is the sum of three cognitive tasks: immediate and delayed recall of 10 words (0–20 points), serial 7 subtraction (0–5 points), and backward counting (0–2 points). A total score of 27 points is possible. Based on the total score, participants were classified into "normal cognition" (12–27 points), "cognitively impaired not dementia (CIND)" (7–11 points), and "dementia" ($\leq$6 points). For proxy-respondents, the summary score consists of cognitive and functional data: a memory rating ranging from excellent to poor (score 0–4), an assessment of limitations in five instrumental activities of daily living (managing money, taking medication, preparing hot meals, using phones, and buying groceries; score 0–5), and the interviewer assessment of difficulty completing the interview because of cognitive limitation (score 0–2 indicating none or some limitation, or prevents completion). High scores are classified as affected by dementia (6–11) and medium scores (3–5) as CIND.

For the analyses, we examined predictors for cognitive impairment (i.e., CIND and dementia; TICSm $\leq$11 points for self-respondents, L-K-W score >3 for proxy-respondents) and dementia (TICSm $\leq$6 points, L-K-W score >6) to address whether predictors varied by severity of impairment.

Statistical analyses

Two survival analyses were conducted for each outcome: RFSA and Cox PH. Both analyses are described in more detail in the Supplementary Material (S1.2). We chose age-at-onset instead of time-in-study (i.e., length of follow-up) as time scale in the survival analyses, because we expected the risk of cognitive impairment/dementia to change as a function of age rather than a function of time-in-

Table 1
Descriptive statistics of the 52 variables in the sample

Demographics	N	M	SD	Min	Max
Age	9,979	67.01	9.18	50	98
Gender (female)	9,979	59.8%			
Education (y)	9,963	13.27	2.61	0	17
Race (African American)	9,976	9.3%			
Race (Other)	9,976	3.5%			
Ethnicity (Hispanic)	9,979	6.2%			
Income (in 1,000 USD)	9,966	66.88	61.97	0	300
Wealth (in 1,000 USD)	9,966	554.19	816.91	–769	5,060
Marital status (married)	9,978	67.3%			
Work	9,809	37.6%			
Type home (assisted)	9,811	0.6%			
Psychosocial	N	M	SD	Min	Max
Conscientiousness	9,895	3.41	0.45	1	4
Openness	9,877	2.99	0.53	1	4
Extraversion	9,908	3.26	0.53	1	4
Agreeableness	9,906	3.55	0.45	1	4
Emotional distress (z)	9,978	–0.01	0.68	–1.31	3.39
Life satisfaction (z)	9,857	0.06	0.98	–2.82	1.67
Positive affect (z)	9,877	0.08	0.95	–3.70	2.04
Purpose in life	9,821	4.70	0.90	1	6
Optimism	9,843	4.57	1.12	1	6
Social contact	7,213	3.72	0.76	1	6
Health	N	M	SD	Min	Max
Subjective health	9,972	3.35	1.02	1	5
Childhood health	6,765	4.23	0.94	1	5
Hearing	9,975	3.43	1.06	1	5
Hear aid	9,302	1.8%			
Sleep medication	9,975	19.8%			
Childhood traumas	9,901	0.34	0.60	0	3
Lifetime traumas	9,911	1.23	1.18	0	7
BMI	9,807	28.97	5.65	11.18	83.82
Highest BMI ever	9,908	30.82	6.55	11.33	49.71
BMI slope	9,979	0.00	0.99	–3.59	3.58
Waist circumference	8,781	39.34	5.60	25.00	55.25
Hypertension	9,960	55.8%			
Diabetes	9,961	18.0%			
Heart disease	9,965	22.9%			
Stroke	9,966	4.9%			
Cancer	9,976	14.8%			
Alcohol	9,965	56.3%			
Mild activity	9,963	93.7%			
Moderate activity	9,962	83.7%			
Vigorous activity	9,960	42.9%			
Total activity	9,966	96.6%			
Smoking ever	9,978	52.6%			
Functional limitations	8,920	2.14	2.43	0	10
Grip strength	7,802	32.12	10.82	0.50	60.50
Biomarker/Polygenic	N	M	SD	Min	Max
Cholesterol	7,802	203.24	41.69	89.04	405.41
High Density Lipoprotein	6,748	55.55	16.30	12.11	139.52
Cystatin C	7,820	0.00	0.13	–0.64	0.98
C Reactive Protein	7,891	0.30	0.51	–1.70	2.34
Hemoglobin A1C	8,119	5.81	0.90	4.07	14.79
Polygenic score *APOE* without $\varepsilon 4$	7,082	–0.03	1.00	–3.95	4.08
Polygenic score *APOE* with $\varepsilon 4$	7,082	–0.03	1.00	–3.95	4.07

(*Continued*)

Table 1
(*Continued*)

Cognitive Status	N (%)	M	SD	Min	Max
Impairment	3,119 (31.3%)				
Dementia	622 (6.2%)				
Time-to-detection impairment	9,979	6.85	2.76	2	10
Time-to-detection dementia	9,979	7.67	2.33	2	10
Age at onset impairment	9,979	73.86	8.70	52	101
Age at onset dementia	9,979	74.68	8.79	52	101

N, number of participants; M, mean; SD, standard deviation. Data of 52 predictors were obtained in 2006 or 2008, respectively, except polygenic scores, which were obtained in 2013. Smoking refers to whether participants have ever smoked from 1992 to 2008. We included three variables of BMI: BMI at baseline, highest BMI ever and BMI slope. Given that there was no measure for midlife obesity in the data, highest BMI ever and BMI slope seemed the best possible alternatives. BMI slope was calculated using information from the earliest available measurement occasion in HRS, e.g., 1992, until 2 years before year of detection of cognitive impairment/dementia or the last year of cognitive testing for unimpaired individuals. Emotional distress is an averaged factor consisting of standardized negative emotion variables (i.e., neuroticism, hostility, anxiety, negative affect, hopelessness, pessimism, depression, loneliness, and perceived constraints). The items of life satisfaction were rated on a scale from 1–6 in 2006 and from 1–7 in 2008. Scores were thus standardized before combining them. For positive affect, participants rated 6 (in 2006) and 12 (in 2008) items. Scores were thus standardized before combining them. Time-to-detection (in years) refers to years from baseline to year in which onset of impairment/dementia was detected. The assessment of all variables is described in the Supplementary Material.

study [23, 24]. For example, we would expect more of a difference in cognitive impairment/dementia risk between a 55 and 70-year-old participant with the same length of follow-up than between two 60-year-olds with a different length of follow-up. By using age-at-onset as the time scale, the model computes risk estimates at the age of cognitive impairment/dementia onset, given the event has not occurred at younger ages. Using age-at-onset as the time scale assumes that participants with similar risks (i.e., the 60-year-olds) belong to the same risk set and indirectly adjusts for a potential age effect [25]. Consequently, age at baseline is expected to have a negative predictive effect in our analyses: Older participants at baseline, compared to younger ones, show decreased risk for cognitive impairment, because they have already demonstrated their cognitive fitness, given that the event of interest (cognitive impairment or dementia) has not yet occurred for them (cf. Supplementary Figure 1). Participants who did not score in the range of cognitive impairment/dementia were censored at the time of their last cognitive assessment.

We randomly divided the data into two subsamples ($n_1 = 4{,}990$; $n_2 = 4{,}989$) so that each of the survival analyses could be conducted independently. In the Cox PH, we included predictors that 1) had RFSA relative importance >0.20 (max = 1.00), and 2) ranked among the strongest 15 predictors in at least four of six sensitivity analyses. In a series of sensitivity analyses (see Supplementary Material S2–S6), we explored the robustness of the VIMP rankings. Pearson correlations were performed to evaluate the consistency between the relative importance of main and sensitivity analyses (Supplementary Table 6).

Analyses were run with R software [26]. For RFSA, we used the "randomForestSRC" package [27]. We generated 1,000 trees per random forest and imputed missing data with 5 iterations. For Cox PH, the "survival" package [28] was used. Missing data were imputed using multiple imputation with the "mice" package [29] (method: predictive mean matching, number of imputations: 30). Missing data were imputed at run-time. All variables included in the analyses were used for data imputation. Significance was set to p-values < 0.01.

RESULTS

The findings of RFSA and Cox PH are shown in Table 2 (cognitive impairment) and Table 3 (dementia). Figure 1 summarizes the most important predictors as determined by the combined methodology of machine learning and semi-parametric survival analysis. Table 4 presents the VIMP ranking of all predictors. The Supplementary Material (S2–S6) contains the findings of the sensitivity analyses. Across all sensitivity analyses, the VIMP rankings were consistent for both cognitive impairment and dementia, and correlation coefficients indicated high consistency ($r = 0.67$–1.00) between the relative importance of main and sensitivity analyses ran with "randomForestSRC".

Cognitive impairment

RFSA

The predictive error rate for the random forests converged to 34% after 1,000 trees were generated. RFSA showed that African American

Table 2
Variables associated with risk of cognitive impairment

		RFSA: I_{rel}	Cox PH		
Rank	Variable	Mean	HR	95% CI	*p*
1	African American	1.00	2.52	[2.18, 2.91]	0.000
2	Wealth	0.59	0.98	[0.91, 1.06]	0.634
3	Education	0.57	0.80	[0.76, 0.84]	0.000
4	BMI Slope	0.33	1.16	[1.09, 1.22]	0.000
5	Subjective Health	0.27	0.87	[0.83, 0.92]	0.000
6	Emotional Distress	0.27	1.38	[1.24, 1.53]	0.000
7	Ethnicity (Hispanic)	0.22	1.65	[1.35, 2.01]	0.000

Random forest survival analysis (RFSA, $n = 4{,}990$) and Cox proportional hazard analysis (Cox PH, $n = 4{,}989$) were conducted in different subsets of participants. The data were randomly divided into two subsamples so that each of the survival analyses could be conducted independently. The simple splitting was based on the outcome using the function "createDataPartition" in R. Relative importance (I_{rel}) refers to the relative importance in predicting risk of cognitive impairment. The relative importance of the strongest predictor in RFSA is expected to be equal 1.00. HR, hazard ratio; 95% CI, 95% confidence intervals.

Table 3
Variables associated with risk of dementia

		RFSA: I_{rel}	Cox PH		
Rank	Variable	Mean	HR	95% CI	*p*
1	BMI Slope	0.42	0.90	[0.73, 1.11]	0.307
2	Emotional Distress	0.40	1.85	[1.41, 2.44]	0.000
3	Diabetes	0.34	1.11	[0.71, 1.74]	0.646
4	African American	0.33	2.17	[1.32, 3.55]	0.002
5	Childhood Traumas	0.23	1.20	[1.01, 1.42]	0.034
6	Hemoglobin A1C	0.20	1.53	[0.90, 2.61]	0.119

Random forest survival analysis (RFSA, $n = 4{,}990$) and Cox proportional hazard analysis (Cox PH, $n = 4{,}989$) were conducted in different subsets of participants. The data were randomly divided into two subsamples so that each of the survival analyses could be conducted independently. The simple splitting was based on the outcome using the function "createDataPartition" in R. Relative importance (I_{rel}) refers to the relative importance in predicting risk of dementia. The relative importance of the strongest predictor in RFSA is expected to be equal 1.00. HR, Hazard ratio; 95% CI, 95% confidence intervals.

was the strongest predictor, with relative importance (I_{rel}) = 1.00. Two sociodemographic variables, wealth and education, were also among the top predictors. The next top predictors were body mass index (BMI) slope, subjective health, emotional distress, and Hispanic ethnicity. The relative importance of other predictors from the literature [2, 5, 8] were smoking (ranked 10th), polygenic score with *APOE* ε4 (ranked 18th), polygenic score without *APOE* ε4 (ranked 33rd), total physical activity (ranked 34th), diabetes (ranked 37th), hypertension (ranked 47th), and BMI at baseline (ranked 48th) (Table 4). It should be noted that age at baseline had a negative predictive effect ($I_{rel} = -0.40$, rank 52), which is in line with expectations using age-at-onset as the time scale of survival metric.

Cox PH

The Cox PH model included seven predictors based on the aforementioned inclusion criteria and was run in the second subsample. Results were mostly consistent with the RFSA but one included predictor (i.e., wealth) did not have a significant hazard ratio. African American was the strongest predictor and associated with having more than twice the relative risk for developing cognitive impairment. One additional unit on education and subjective health was associated with decreased relative risk (minus 20% and 13%, respectively). In contrast, an increase of 1 SD in both BMI slope and emotional distress increased relative risk by 16% and 38%, respectively. Likewise, Hispanics were associated with 65% greater chance to develop cognitive impairment.

Dementia

RFSA

The predictive error rate for the random forests converged to 36% after 1,000 trees were generated. The three top predictors for dementia risk were BMI slope, emotional distress, and diabetes, followed by African American, childhood traumas, and hemoglobin A1C. Of further note, total physical activity ranked 15th, hypertension ranked 27th,

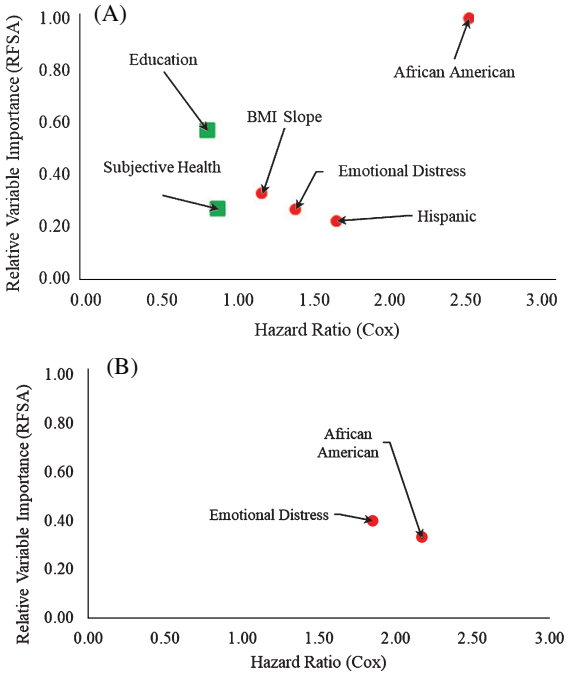

Fig. 1. The most important predictors for cognitive impairment (Part A) and dementia (Part B). The relative variable importance (as determined by random forest survival analysis) is graphed on the y-axis, ranging from 0 (lowest importance) to 1 (highest importance). The hazard ratios (as determined by the Cox PH survival analysis) are shown on the x-axis, ranging from 0 to 3. The shape of the marker indicates whether the factors were significantly related to an increased (round) or decreased (square) risk.

polygenic score without *APOE* ε4 ranked 30th, polygenic score with *APOE* ε4 ranked 41st, baseline BMI ranked 48th, and smoking ranked 50th (Table 4). Again, age at baseline had a negative predictive effect ($I_{rel} = -1.00$, rank 52).

Cox PH

Of the six included predictors, two had a significant hazard ratio. Specifically, an increase of 1 SD in emotional distress was linked to 85% increased relative risk. Furthermore, African American was also related to relative increased risk (plus 117%). BMI slope, diabetes, childhood traumas, and hemoglobin A1C were not significantly linked to incident dementia.

DISCUSSION

This study used a combined methodology of machine learning and semi-parametric survival analyses to estimate the relative importance of 52 variables in predicting cognitive impairment and dementia in a large sample of older adults. Results showed that African Americans and individuals with high scores of emotional distress were at relatively highest risk for developing cognitive impairment and dementia. Additionally, health variables (worse subjective health, increasing change in BMI) and sociodemographic variables (lower education, Hispanic ethnicity) were comparatively strong risk factors for cognitive impairment. More commonly studied and identified risk factors from the literature, such as cardiovascular variables and polygenic scores, appeared to be of lesser importance when evaluated concurrently with the above predictors. Multiple *post-hoc* sensitivity analyses pointed to the robustness of these results.

The finding that African American was among the most important risk factors for both cognitive impairment and dementia accords with past studies [30, 31] that reported a greater dementia risk for African Americans compared to Whites and other ethnic groups. Specific factors contributing to this outcome are difficult to pinpoint, but they are likely related to underlying sociodemographic conditions that affect access to health care [32, 33]. Several of these risk factors (e.g., education, wealth) showed a similarly high relative importance in the RFSA; however, African American was overall among the most important predictors—probably because for predictive purposes it spanned multiple other risk dimensions. Of note, lower wealth was the second strongest risk factor for cognitive impairment in RFSA but was non-significant in Cox PH. It could be that African American had a direct effect, whereas wealth had an indirect effect on cognitive impairment (e.g., via complicated interactions not included in Cox PH but considered in RFSA). Or, it seems also possible that the predictive overlap of wealth and African American did not affect estimates of relative importance in RFSA (because RFSA is robust against multicollinearity), whereas in Cox PH, predictive power shared between wealth and African American was largely subsumed by the latter.

A *post-hoc* analysis revealed that, indeed, African Americans reported significantly lower wealth ($M = \$208{,}000$; $SD = \$391{,}000$) compared to Whites ($M = \$590{,}000$; $SD = \$847{,}000$), $t(973) = -17.21$, $p < 0.001$, and this supports the idea of economic disparity as a key factor contributing to risk differences between African Americans and Whites. Lower financial status could limit access to health care and may also negatively impact variables beneficial for building cognitive reserve, such as access to educational resources and environmental settings conducive to mental wellness [34]. Future studies may

Table 4
Variable Importance (VIMP) ranking for cognitive impairment (Part A) and dementia (Part B)

Rank	A: Cognitive Impairment	I_{rel}	B: Dementia	I_{rel}
1	African American	1.00	BMI Slope	0.42
2	Wealth	0.59	Emotional Distress	0.40
3	Education	0.57	Diabetes	0.34
4	BMI Slope	0.33	African American	0.33
5	Subjective Health	0.27	Childhood Traumas	0.23
6	Emotional Distress	0.27	Hemoglobin A1C	0.20
7	Ethnicity (Hispanic)	0.22	Education	0.19
8	Grip Strength	0.17	Life Satisfaction	0.17
9	Childhood Traumas	0.16	Ethnicity (Hispanic)	0.15
10	Smoking	0.15	Childhood Health	0.15
11	Marital Status	0.13	Funct. Limitations	0.13
12	Social Contact	0.11	Subjective Health	0.10
13	Cystatin C	0.10	Stroke	0.09
14	Income	0.08	Social Contact	0.08
15	Cholesterol	0.07	Activity Total	0.07
16	Work	0.07	Openness	0.07
17	Childhood Health	0.06	Hearing	0.07
18	Polygenic Score with *APOE* ε4	0.05	Cystatin C	0.06
19	Openness	0.05	Purpose in Life	0.06
20	Functional Limitations	0.05	Alcohol	0.06
21	Hemoglobin A1C	0.04	Income	0.05
22	Race (Other)	0.03	Positive Affect	0.05
23	Conscientiousness	0.03	Lifetime Traumas	0.04
24	Agreeableness	0.02	Work	0.03
25	Cancer	0.02	Marital Status	0.03
26	Alcohol	0.02	Agreeableness	0.03
27	Gender	0.02	Hypertension	0.02
28	Lifetime Traumas	0.02	Moderate Activity	0.02
29	Hearing	0.01	Wealth	0.01
30	Vigorous Activity	0.01	Polygenic Score with *APOE* ε4	0.01
31	Optimism	0.01	Mild Activity	0.01
32	Mild Activity	0.01	Conscientiousness	0.01
33	Polygenic Score without *APOE* ε4	0.01	Optimism	0.01
34	Activity Total	0.01	Highest BMI	0.01
35	Extraversion	0.00	Hear Aid	0.00
36	Purpose in Life	0.00	Type Home	0.00
37	Diabetes	0.00	Extraversion	–0.01
38	Life Satisfaction	0.00	Vigorous Activity	–0.01
39	High Density Lipoprotein	0.00	Race (Other)	–0.01
40	Type Home	0.00	Gender	–0.01
41	Hear Aid	0.00	Polygenic Score without *APOE* ε4	–0.01
42	Stroke	–0.01	High Density Lipoprotein	–0.02
43	Heart Disease	–0.01	Sleep Medication	–0.02
44	Highest BMI	–0.01	Grip Strength	–0.02
45	Moderate Activity	–0.01	Heart Disease	–0.02
46	Sleep Medication	–0.01	Cancer	–0.04
47	Hypertension	–0.01	Cholesterol	–0.04
48	BMI	–0.01	BMI	–0.05
49	Waist Circumference	–0.02	C Reactive Protein	–0.07
50	Positive Affect	–0.02	Smoking	–0.07
51	C Reactive Protein	–0.04	Waist Circumference	–0.07
52	Age at Baseline	–0.40	Age at Baseline	–1.00

Random forest survival analysis (RFSA) was conducted in a subsample ($n = 4{,}990$) separately for cognitive impairment and dementia. Relative importance (I_{rel}) refers to the relative importance in predicting risk of cognitive impairment and dementia, respectively.

examine whether contextual factors (e.g., neighborhood as indicator for socioeconomic status) provide additional predictive information concerning African Americans' elevated risk for cognitive impairment.

Emotional distress was related to increased risk of cognitive impairment and dementia. However, at present, the corresponding literature is somewhat fragmented by the use of numerous negative emotion-

ality measures [5, 8, 35, 36]. In this study, we used factor analytical techniques to create a composite score of emotional distress (i.e., neuroticism, hostility, anxiety, negative affect, hopelessness, pessimism, depression, loneliness, and perceived constraints) that captures a core latent construct underlying these highly interrelated measures. Emotional distress may lead to social and cognitive disengagement, which in turn may exacerbate cognitive decline [2, 5, 8, 9]. Additionally, higher emotional distress is associated with higher cortisol levels [37], which over time may contribute to brain atrophy, loss of cognitive function, and ultimately dementia [38].

Additionally, two health variables (poor subjective health, increasing change in BMI) and two sociodemographic variables (lower education, Hispanic ethnicity) were comparatively strong predictors for cognitive impairment. Subjective health has been shown to be a reliable and robust predictor of mortality risk [14], but its prognostic capacity for cognitive impairment has been rarely examined [39]. Subjective health may reflect complex interrelations between mental, social, functional, and biological aspects of an individual [40]. This higher-order integration might, compared with more narrowly defined risk factors, be especially important for predicting cognitive impairment in older adults [14].

As for BMI, its relation with cognitive impairment/dementia appears to vary with age: A comparatively higher BMI in midlife may increase risk, whereas a lower BMI in later life may be a marker of dementia [41–43]. Our results show that gain in BMI (positive slope) in later life predicts increased risk of cognitive impairment. This suggests that future research may benefit by considering not only stable between-person differences in BMI at a given time or age but also within-person changes in BMI as predictive of increased cognitive risk.

Education had a protective effect on cognitive health, as previously shown in other studies [2, 44, 45]. Higher education may impact cognitive health through multiple and complex pathways [46]. Specifically, higher education is associated with a healthy lifestyle, secure income, supportive relationships, and better general health [47], any of which may confer increased protection against cognitive decline. Considering its potential modifiability, prioritizing education in future research and public policy related to cognitive decline appears very much worthwhile. Along these lines, improved access to education [45] and evaluating whether education after secondary school provides additional protection against cognitive decline [4] are important goals. In addition to education, Hispanic ethnicity was among the most important sociodemographic risk factor for cognitive impairment. The risk for Hispanics was found to be lower than for African Americans but higher than for Whites [30, 48, 49].

To summarize, African American and emotional distress appeared as key risk factors for both cognitive impairment and dementia across both analyses (Fig. 1). Risk factors identified by previous ranking approaches [2, 5, 8, 13], such as smoking, physical inactivity, or hypertension, as well as polygenic scores, were of less importance. It seems likely that these more narrowly defined risk factors were here subsumed by higher-order factors, such as emotional distress or subjective health, as the latter reflect complex interrelations between various aspects of an individual [40].

Limitations

This study has several limitations. First, predictive associations were based on correlational analyses which are inherently bidirectional. It could be that the preclinical phase preceding the onset of cognitive impairment/dementia leads to changes in the predictors (reverse causation). Considering possible reverse causality, we conducted a sensitivity analysis by excluding participants whose cognitive impairment/dementia had developed within the first 2 years of follow-up (Supplementary Material). For cognitive impairment, the VIMP ranking was robust for 13/15 predictors (Supplementary Table 1). For dementia, the VIMP ranking was robust for 12/15 predictors (Supplementary Table 2). For example, emotional distress decreased from rank 2nd ($I_{rel} = 0.40$) to 6th ($I_{rel} = 0.12$) when excluding the individuals whose dementia had developed within the first 2 years of follow-up. This may indicate that prodromal dementia leads to higher emotional distress. Although we only included cognitively healthy individuals at baseline and adopted an analytical framework that was inherently longitudinal (i.e., we used earlier assessments to predict upcoming cognitive impairment or dementia), we cannot disentangle completely risk factors from reverse causation. Future research could aim to do so by extending our analytical framework to examine longer follow-ups and time-ordered predictor-outcome associations during the transition period from the preclinical to the overt clinical phase.

Second, cognitive outcome measures were based on a performance measure for self-respondents and a rating measure for proxy-respondents rather than on clinical diagnosis. It would be helpful to validate our findings against clinical diagnoses of cognitive impairment/dementia and with respect to the different etiologies of dementia (e.g., Alzheimer's disease, frontotemporal dementia). Additionally, we did not consider baseline cognitive performance as a predictor of cognitive impairment or dementia risk because 1) these cognitive variables were based on the same measures as our outcomes (leading to inflated estimates of predictive power), 2) doing so would have discarded data from individuals who were present at a single follow-up measurement occasion only, thereby greatly reducing generalizability of our findings because of selectivity, and 3) we considered it important to examine risk factors for cognitive impairment and dementia independently of knowledge of prior cognitive history (e.g., consistent with initial presentation in a clinical setting, where patients do not show up with prior cognitive assessments).

Third, the follow-up was relatively short, especially for cardiovascular risk factors. It is possible that variables such as hypertension and BMI (i.e., obesity) appear of greater importance when assessed in midlife rather than late life. For instance, a review [50] concluded that dementia risk was larger in studies that measured hypertension, obesity, and dyslipidemia in midlife (compared to late life) and had a longer follow-up. Thus, the timing of when during the life course different risk factors are important is crucial and requires further investigation.

Lastly, the relative importance scores (I_{rel}) are not weighted by population prevalence, so although their relevance is appropriately estimated from a statistical perspective within our sample, it may have varying implications at the population level. Yet, given the overall representativeness of our sample, we believe that a weighting procedure applied to the estimated relative importance scores would not have substantially modified the ranking of predictors we present here. Surely, from a public health perspective, it would be invaluable to know the proportion of individuals in the total population that are affected by the various risk factors (population attributable risk), but this analysis is beyond the scope of the current work.

Conclusion

A combined methodology of machine learning and semi-parametric survival analysis allowed for robust, simultaneous estimation of the relative importance of 52 multi-domain risk factors for cognitive impairment and dementia within a large, representative sample of adults from the US population. The present findings suggest that higher-order factors (e.g., emotional distress, subjective health) are more important than narrowly defined factors (e.g., clinical and behavioral indicators) when evaluated concurrently. Higher-order factors are likely to reflect complex interactions between functional, social, mental and biological aspects of an individual, accumulated over their lifespan. Identifying these interactions seems a considerable challenge, but one that affords the potential to better understand the multifactorial etiology of dementia. Multi-interdisciplinary collaborations that integrate multidimensional layers of data (e.g., health, lifespan, environmental, social, genetic) and apply multi-methodological approaches (e.g., machine learning, survival analysis, multilevel modeling) are now necessary to move the field forward and address one of the most urgent global health challenges of our time.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Aging of the National Institutes of Health (Grant Numbers R21AG057917 and R01AG053297), by a Joint Seed Grant of the universities Geneva and Zurich (Project "*Bridging Personality and Cognition: Conceptual and Methodological Challenges in the Age of Digital Transformation*"), and by the European Commission, Horizon2020, Grant agreement number: 732592–Lifebrain–H2020-SC1-2016-2017/H2020-SC1-2016-RTD. The Health and Retirement Study (HRS) is supported by the National Institute on Aging (NIAU01AG009740).

P. Ghisletta, M. Allemand, and D. Aschwanden developed the concept for the broader research project from which the current work stems (Joint Seed Project: "*Bridging Personality and Cognition: Conceptual and Methodological Challenges in the Age of Digital Transformation*").

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/19-0967r2).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-190967.

REFERENCES

[1] Alzheimer's Disease International (2018) *World Alzheimer's Report 2018. The state of the art of dementia research: New frontiers*. Alzheimer's Disease International, London, UK.

[2] Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, Ballard C, Banerjee S, Burns A, Cohen-Mansfield J, Cooper C, Fox N, Gitlin LN, Howard R, Kales HC, Larson EB, Ritchie K, Rockwood K, Sampson EL, Samus Q, Schneider LS, Selbæk G, Teri L, Mukadam N (2017) Dementia prevention, intervention, and care. *Lancet* **390**, 2673-2734.

[3] World Health Organization (2019) *Risk reduction of cognitive decline and dementia: WHO guidelines*. World Health Organization, Geneva, Switzerland.

[4] Alzheimer's Disease International (2019) *From plan to impact II: The urgent need for action.* Alzheimer's Disease International, London, UK.

[5] Baumgart M, Snyder HM, Carrillo MC, Fazio S, Kim H, Johns H (2015) Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective. *Alzheimers Dement* **11**, 718-726.

[6] Anstey KJ, Lipnicki DM, Low L-F (2008) Cholesterol as a risk factor for dementia and cognitive decline: A systematic review of prospective studies with meta-analysis. *Am J Geriatr Psychiatry* **16**, 343-354.

[7] Bellou V, Belbasis L, Tzoulaki I, Middleton LT, Ioannidis JPA, Evangelou E (2017) Systematic evaluation of the associations between environmental risk factors and dementia: An umbrella review of systematic reviews and meta-analyses. *Alzheimers Dement* **13**, 406-418.

[8] Deckers K, van Boxtel MPJ, Schiepers OJG, de Vugt M, Muñoz Sánchez JL, Anstey KJ, Brayne C, Dartigues J-F, Engedal K, Kivipelto M, Ritchie K, Starr JM, Yaffe K, Irving K, Verhey FRJ, Köhler S (2015) Target risk factors for dementia prevention: A systematic review and Delphi consensus study on the evidence from observational studies. *Int J Geriatr Psychiatry* **30**, 234-246.

[9] Fratiglioni L, Paillard-Borg S, Winblad B (2004) An active and socially integrated lifestyle in late life might protect against dementia. *Lancet Neurol* **3**, 343-353.

[10] Terracciano A, Stephan Y, Luchetti M, Albanese E, Sutin AR (2017) Personality traits and risk of cognitive impairment and dementia. *J Psychiatr Res* **89**, 22-27.

[11] Terracciano A, Sutin AR (2019) Personality and Alzheimer's disease: An integrative review. *Personal Disord Theory Res Treat* **10**, 4-12.

[12] Chapman BP, Huang A, Peters K, Horner E, Manly J, Bennett DA, Lapham S (2019) Association between high school personality phenotype and dementia 54 years later in results from a national us sample. *JAMA Psychiatry* **77**, 148-154.

[13] Barnes DE, Covinsky KE, Whitmer RA, Kuller LH, Lopez OL, Yaffe K (2009) Predicting risk of dementia in older adults: The late-life dementia risk index. *Neurology* **73**, 173-179.

[14] Aichele S, Rabbitt P, Ghisletta P (2016) Think fast, feel fine, live long: A 29-year study of cognition, health, and survival in middle-aged and older adults. *Psychol Sci* **27**, 518-529.

[15] BTreiman L (2001) Random forests. *Mach Learn* **45**, 5-32.

[16] Strobl C, Malley J, Tutz G (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* **14**, 323-348.

[17] Cox DR (1972) Regression models and life-tables. *J R Stat Soc Ser B Methodol* **34**, 187-220.

[18] Gianattasio KZ, Wu Q, Glymour MM, Power MC (2019) Comparison of methods for algorithmic classification of dementia status in the health and retirement study. *Epidemiology* **30**, 291-302.

[19] Crimmins EM, Kim JK, Langa KM, Weir DR (2011) Assessment of cognition using surveys and neuropsychological assessment: the Health and Retirement Study and the Aging, Demographics, and Memory Study. *J Gerontol B Psychol Sci Soc Sci* **66B**, i162-i171.

[20] Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR (2014) Cohort profile: The Health and Retirement Study (HRS). *Int J Epidemiol* **43**, 576-585.

[21] Weir D, Faul J, Langa K (2011) Proxy interviews and bias in the distribution of cognitive abilities due to non-response in longitudinal studies: A comparison of HRS and ELSA. *Longitud Life Course Stud* **2**, 170-184.

[22] van den Berg E, Ruis C, Biessels GJ, Kappelle LJ, van Zandvoort MJE (2012) The Telephone Interview for Cognitive Status (modified): Relation with a comprehensive neuropsychological assessment. *J Clin Exp Neuropsychol* **34**, 598-605.

[23] Thiébaut ACM, Bénichou J (2004) Choice of time-scale in Cox's model analysis of epidemiologic cohort data: A simulation study. *Stat Med* **23**, 3803-3820.

[24] Kom EL, Graubard BI, Midthune D (1997) Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *Am J Epidemiol* **145**, 72-80.

[25] Cheung YB, Gao F, Khoo KS (2003) Age at diagnosis and the choice of survival analysis methods in cancer epidemiology. *J Clin Epidemiol* **56**, 38-43.

[26] R Development Core Team (2018) *R: A language and environment for statistical computing*.

[27] Ishwaran H, Kogalur UB (2019) *Random forests for survival, regression, and classification (RF-SRC)*.

[28] Therneau TM (2019) *Package "survival."*

[29] Buuren S van, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. *J Stat Softw* **45**, doi: 10.18637/jss.v045.i03

[30] Mayeda ER, Glymour MM, Quesenberry CP, Whitmer RA (2016) Inequalities in dementia incidence between six racial and ethnic groups over 14 years. *Alzheimers Dement* **12**, 216-224.

[31] Potter GG, Plassman BL, Burke JR, Kabeto MU, Langa KM, Llewellyn DJ, Rogers MAM, Steffens DC (2009) Cognitive performance and informant reports in the diagnosis of cognitive impairment and dementia in African Americans and whites. *Alzheimers Dement* **5**, 445-453.

[32] Chen C, Zissimopoulos JM (2018) Racial and ethnic differences in trends in dementia prevalence and risk factors in the United States. *Alzheimers Dement (N Y)* **4**, 510-520.

[33] Hill CV, Pérez-Stable EJ, Anderson NA, Bernard MA (2015) The National Institute on Aging Health Disparities Research Framework. *Ethn Dis* **25**, 245-254.

[34] Cadar D, Lassale C, Davies H, Llewellyn DJ, Batty GD, Steptoe A (2018) Individual and area-based socioeconomic factors associated with dementia incidence in England: Evidence from a 12-year follow-up in the English Longitudinal Study of Ageing. *JAMA Psychiatry* **75**, 723-732.

[35] Sutin AR, Stephan Y, Terracciano A (2018) Psychological distress, self-beliefs, and risk of cognitive impairment and dementia. *J Alzheimers Dis* **65**, 1041-1050.

[36] Gallacher J, Bayer A, Fish M, Pickering J, Pedro S, Dunstan F, Ebrahim S, Ben-Shlomo Y (2009) Does anxiety affect

risk of dementia? Findings from the Caerphilly Prospective Study. *Psychosom Med* **71**, 659-666.

[37] Ennis GE, An Y, Resnick SM, Ferrucci L, O'Brien RJ, Moffat SD (2017) Long-term cortisol measures predict Alzheimer disease risk. *Neurology* **88**, 371-378.

[38] Popp J, Wolfsgruber S, Heuser I, Peters O, Hüll M, Schröder J, Möller H-J, Lewczuk P, Schneider A, Jahn H, Luckhaus C, Perneczky R, Frölich L, Wagner M, Maier W, Wiltfang J, Kornhuber J, Jessen F (2015) Cerebrospinal fluid cortisol and clinical disease progression in MCI and dementia of Alzheimer's type. *Neurobiol Aging* **36**, 601-607.

[39] Bond J, Dickinson HO, Matthews F, Jagger C, Brayne C, MRC CFAS (2006) Self-rated health status as a predictor of death, functional and cognitive impairment: A longitudinal cohort study. *Eur J Ageing* **3**, 193-206.

[40] Ocampo JM (2010) Self-rated health: Importance of use in elderly adults. *Colomb Méd* **41**, 275-289.

[41] Gustafson DR, Luchsinger JA (2013) High adiposity: Risk factor for dementia and Alzheimer's disease? *Alzheimers Res Ther* **5**, 57.

[42] Singh-Manoux A, Dugravot A, Shipley M, Brunner EJ, Elbaz A, Sabia S, Kivimaki M (2018) Obesity trajectories and risk of dementia: 28 years of follow-up in the Whitehall II Study. *Alzheimers Dement* **14**, 178-186.

[43] Kivimäki M, Luukkonen R, Batty GD, Ferrie JE, Pentti J, Nyberg ST, Shipley MJ, Alfredsson L, Fransson EI, Goldberg M, Knutsson A, Koskenvuo M, Kuosma E, Nordin M, Suominen SB, Theorell T, Vuoksimaa E, Westerholm P, Westerlund H, Zins M, Kivipelto M, Vahtera J, Kaprio J, Singh-Manoux A, Jokela M (2018) Body mass index and risk of dementia: Analysis of individual-level data from 1.3 million individuals. *Alzheimers Dement* **14**, 601-609.

[44] Evans DA, Hebert LE, Beckett LA, Scherr PA, Albert MS, Chown MJ, Pilgrim DM, Taylor JO (1997) Education and other measures of socioeconomic status and risk of incident Alzheimer Disease in a defined population of older persons. *Arch Neurol* **54**, 1399-1405.

[45] Norton S, Matthews FE, Barnes DE, Yaffe K, Brayne C (2014) Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. *Lancet Neurol* **13**, 788-794.

[46] Langa KM, Larson EB, Karlawish JH, Cutler DM, Kabeto MU, Kim SY, Rosen AB (2008) Trends in the prevalence and mortality of cognitive impairment in the United States: Is there evidence of a compression of cognitive morbidity? *Alzheimers Dement* **4**, 134-144.

[47] Mirowsky J, Ross CE (2005) Education, cumulative advantage, and health. *Ageing Int* **30**, 27-62.

[48] González HM, Tarraf W, Gouskova N, Gallo LC, Penedo FJ, Davis SM, Lipton RB, Argüelles W, Choca JP, Catellier DJ, Mosley TH (2015) Neurocognitive function among middle-aged and older Hispanic/Latinos: Results from the Hispanic community health study/study of Latinos. *Arch Clin Neuropsychol* **30**, 68-77.

[49] Filshtein TJ, Dugger BN, Jin L-W, Olichney JM, Farias ST, Carvajal-Carmona L, Lott P, Mungas D, Reed B, Beckett LA, DeCarli C (2019) Neuropathological diagnoses of demented Hispanic, Black, And Non-Hispanic White Decedents seen at an Alzheimer's Disease Center. *J Alzheimers Dis* **68**, 145-158.

[50] Kloppenborg RP, van den Berg E, Kappelle LJ, Biessels GJ (2008) Diabetes and other vascular risk factors for dementia: Which factor matters most? A systematic review. *Eur J Pharmacol* **585**, 97-108.