Predicting global cognitive decline in the general population using the Disease State Index

Lotte G.M. Cremers*, Wyke Huizinga*, Wiro J. Niessen, Gabriel P. Krestin, M. Arfan Ikram, Jyrki Lötjönen, Stefan Klein**, Meike W.Vernooij**

Submitted

**Denotes equal contribution*

***Denotes shared last author*

ABSTRACT

Background Identifying persons at risk for cognitive decline may aid in early detection of persons at risk of dementia and to select those that would benefit most from therapeutic or preventive measures for dementia.

Objective In this study we aimed to validate whether cognitive decline in the general population can be predicted with multi-variate data using a previously proposed supervised classification method: Disease State Index (DSI).

Methods We included 2,542 participants, non-demented and without mild cognitive impairment at baseline, from the population-based Rotterdam Study (mean age 60.9 ± 9.1 years). Participants with significant global cognitive decline were de- fined as the 5% of participants with the largest cognitive decline per year. We trained DSI to predict occurrence of significant global cognitive decline using a large variety of baseline features. Prediction performance was assessed as area under the receiver operating characteristic curve (AUC), using 500 repetitions of 2-fold cross-validation experiments.

Results A mean AUC (95% confidence interval) for DSI prediction was 0.78 (0.77 - 0.79) using only age as input feature. When using all available features, a mean AUC of 0.77 (0.75 - 0.78) was obtained. Without age, and with age-corrected features and feature selection on MRI features, a mean AUC of 0.70 (0.63 - 0.76) was obtained, showing the potential of other features besides age.

Conclusion The best performance in the prediction of global cognitive decline in the general population by DSI was obtained using only age as input feature. Other features showed potential, but did not improve prediction. Future studies should evaluate whether the performance could be improved by new features, e.g., longitudinal features, and other prediction methods.

INTRODUCTION

It is well established that neuropathological brain changes related to dementia accumulate over decades, and that the disease has a long preclinical phase. This may facilitate early disease detection and prediction.[1] A large amount of body of literature on potential features and risk factors for dementia exists. However, clinicians often struggle to integrate all the data obtained from a single patient for diagnostic or prognostic purposes. Therefore, there is a need for information technologies and computer-based methods that support clinical decision making.[2] Disease State Index (DSI) is a supervised machine learning method intended to aid clinical decision making.[3] This method compares a variety of patient variables with those variables from previously diagnosed cases, and computes an index that measures the similarity of the patient to the diagnostic group studied. The DSI method has previously been tested in specific patient populations and has shown to perform reasonably well in the early prediction of progression from mild cognitive impairment (MCI) to Alzheimer's disease and has been successful in the classification of different dementia subtypes [3–6]. In a recent study DSI has been validated in a population-based setting to predict late-life dementia.[7] Identification of persons at risk for global cognitive decline may aid in early detection of persons at risk of dementia and may help to develop therapeutic or preventive measures to postpone or even prevent further cognitive decline and dementia.[8] This is especially important since previous research has shown that preventive interventions for dementia were more effective in persons at risk than in unselected populations.[9, 10] We therefore used DSI to predict global cognitive decline in the general population to select the persons at risk. The main aim of this study was to investigate whether multi-variate data can predict global cognitive decline in the general population. If a high risk group can be selected from the general population, a population screening program for this group might facilitate early detection of dementia. We evaluated the prediction performance using several sets of clinical features and features derived brain images acquired with magnetic resonance imagaging (MRI), to assess whether the prediction is dependent on the combination of the input features. DSI was chosen as a classification method because this method is able to handle datasets with missing data, which is often the case in population study datasets. Also, this method has been succesfully applied in previous studies and performed comparable to other state-of-the-art classificiers.[4, 7, 11]

MATERIALS AND METHODS

Study population

We included participants from three independent cohorts within the Rotterdam Study (RS), a prospective population-based cohort study in a suburb of Rotterdam

that investigates the determinants and occurrence of diseases in the middle-aged and elderly population.[12] Brain MRI-scanning was implemented in the study protocol since 2005.[13] The Rotterdam Study has been approved by the medical ethics committee according to the Population Study Act Rotterdam Study, executed by the Ministry of Health,Welfare and Sports of the Netherlands. Written informed consent was obtained from all participants.[14] We used data from RS cohorts I, II and III, of which each consists of multiple subcohorts. In this study a subcohort of RS cohort I, II and III were used, to which we refer as sI, sII and sIII, respectively. Baseline features of sI were collected during 2009-2011 and sII were collected during 2004-2006. The participants of the both cohorts were 55 years or older. For RS cohort III participants were 45 years or older at time of inclusion. Baseline features of sIII were collected during 2006-2008. Participants with prevalent dementia, mild cognitive impairment (MCI) and MRI defined cortical infarcts at baseline were excluded for all analyses. In total, 4328 participants with baseline information on cognition, MRI and other features were included. Baseline MRI was acquired on average 0.3± 0.45 years after collecting the non-imaging features. Furthermore, diffusion-MRI was acquired. However, for a subset of 680 participants in RS cohort II diffusion-MRI data was obtained on average 3.5±0.2 years later than the other baseline MRI features. Longitudinal data on global decline was available for 2542 out of 4328 participants. The follow-up cognitive assessment was on average 5.7±0.6 years after the baseline visit.

Disease State Index

Prediction was performed with DSI.[3] This classifier derives an index indicating the disease state of the partici- pant under investigation based on the available features of that participant. DSI has two major advantages: 1) it can cope with missing data and 2) it gives an interpretable result because DSI also provides a decision tree that can be quite well explained.

DSI classifier is composed of the components: fitness and relevance.[3] Let N be the total number of negatives, P the total number of positives, $FN(x)$ the number of false negatives, and $FP(x)$ the number of false positives, when x is used as classification cut-off. Then the fitness function is estimated for each feature i as:

$$f_i(x) = \frac{FNR_i(x)}{FNR_i(x) + \mathrm{FPR}_i(x)} - \frac{FN_i(x)}{FN_i(x) + \frac{P}{N} FP_i(x)} \quad (1)$$

where $FNR(x) = FN(x)/P$ is the false negative rate and $FPR(x) = FP(x)/N$ is the false positive rate in the training data when the feature value x is used as the classification cut-off. The fitness automatically accounts for the imbalance in class size making implicitly both classes equal in size, as the fraction P/N in the denominator scales the negative class (related to $FP(x)$) to correspond the size of the positive class. The fitness function is a classifier where the values < 0.5 imply negative class and > 0.5 positive class. The relevance of each feature is estimated by:

$$R = \max\{\mathrm{sensitivity} + \mathrm{specificity} - 1, 0\}, \quad (2)$$

which measures how good the feature is in differentiating the two classes. The lower the overlap between the distributions of positives and negatives, the higher R. Finally, DSI is computed from the equation:

$$\mathrm{DSI} = \frac{\sum_i R_i f_i}{\sum_i R_i} \quad (3)$$

DSI is a value between zero and one; somebody is classified as positive if DSI > 0.5 and as negative if DSI < 0.5. DSI is an ensemble classifier, meaning that it combines multiple independent classifiers (fitness functions) defined for each feature separately. Because of that, DSI can tolerate missing data. Features can be grouped in a hierarchical manner. The final DSI is a combination of the levels in the hierarchy. The fitness, relevance and their combination as a composite DSI are repeated recursively by grouping the data until a single DSI value is obtained.[11] Therefore, the final DSI, which is used for the classification, depends on the hierarchy structure, as a different structure leads to a different averaging of the feature combinations. The top-level part of the hierarchy defined for this study is shown in **Figure 1**.

Baseline features

Figure 1 shows the used categories of features in hierarchical manner. Please note that not all individual features are shown in this figure. The sections below describe all the used features (indicated in bold font) in detail.

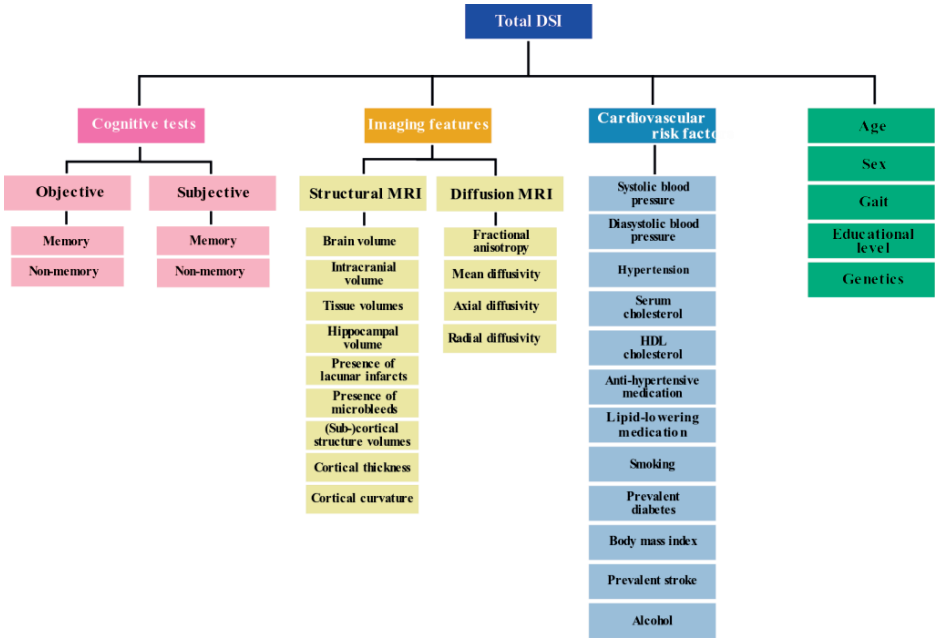

Figure 1. Feature categories shown in a hierarchy as used by the DSI. Please note that not all individual features are included in this graph

MRI features

Multi-sequence MR imaging was performed on a 1.5 Tesla MRI scanner (GE Signa Excite). The imaging protocol and sequence details were described extensively elsewhere.[13] Morphological imaging was performed with T1- weighted, proton density-weighted and fluid-attenuated inversion recovery (FLAIR) sequences. These sequences were used for an automated tissue segmentation approach to segment scans into grey matter, white matter, cere- brospinal fluid (CSF) and background tissue.[15] **Intracranial volume (ICV)** (excluding the cerebellum and surrounding CSF cerebellar) was estimated by summing total grey and white matter and CSF. Brain tissue segmentation was complemented with a **white matter lesion** segmentation based on the tissue segmentation and the FLAIR image with extraction of white matter lesion voxels by intensity thresholding.[16] We obtained **(sub)cortical struc- ture volumes**, **cortical thickness**, and **curvature** of the cortex and hippocampal volume using the publicly avail- able FreeSurfer 5.1 software [17–19]. For **cerebral blood flow** measurements, we performed a 2D phase-contrast imaging as previously described.[20] In short, blood flow velocity (mm/sec) was calculated based on regions of interest (ROI) drawn on the phase-contrast images in the carotid arteries and basilar artery at a level just under the skull base. The value of mean signal intensity in each ROI reflected the flow velocity with the cross-sectional area of the vessel. Flow was calculated by multiplying

the average velocity with the cross-sectional area of the vessel.[20] A 3D T2*-weighted gradient-recalled echo was used to image **cerebral microbleeds**. Microbleeds were defined as focal areas of very low signal intensity, smaller than 10 mm in size and were rated by one of five trained raters who were blinded to other MRI sequences and to clinical data.[21, 22] **Lacunar infarcts** were defined as focal parenchymal lesions >3 mm and < 15 mm in size with the same signal characteristics as cerebrospinal fluid on all sequences and with a hyperintense rim on the FLAIR image (supratentorially). Probabilistic tractography was used to segment 15 different white matter tracts in diffusion-weighted MR brain images, and we obtained **mean FA**, **mean MD**, **axial** and **radial diffusivity** inside each white matter tract.[23]

Cardiovascular risk factors

Cardiovascular risk factors were based on information derived from home interviews and physical examinations during the center visit. **Blood pressure** was measured twice at the right brachial artery in sitting position using a random-zero sphygmomanometer. We used the mean of two measurements in the analyses. Information on the use of **antihypertensive medication** was obtained by using questionnaires and by checking the medication cabinets of the participants. Hypertension was defined as a systolic blood pressure >140 mmHg or a diastolic blood pressure
>90 mmHg or the use of anti-hypertensive medication at baseline. **Serum total cholesterol and high-density lipoprotein (HDL) cholesterol** were measured in fasting serum, taking lipid-lowering medication into account.
Smoking was assessed by interview and coded as never, former and current. **Body-mass index (BMI)** is defined as weight kilograms divided by height in meters squared. **Diabetes mellitus** status was defined as a fasting serum glucose level (>7.0 mmol/l) or, if unavailable, non-fasting serum glucose level (>11.1 mmol/l) or the use of anti-diabetic medication.[14] **Alcohol consumption** was acquired in a questionnaire. **Prevalent stroke** was ascertained as previously described.[24] **Educational level** was assessed during a home interview and was categorized into 7 categories, ranging from primary education only to university level.[14]

APOE-ε4 allele carriership

APOE-E4 allele carriership was assessed on coded genomic DNA samples.[25] APOE-genotype was in Hardy- Weinberg equilibrium. APOE-E4 allele carriership was coded positive in case of one or two APOE-E4 alleles.

Gait features

Gait was assessed by three walking tasks over a walkway: “normal walk”, “turn” and “tandem walk” (heel to toe).[26] Using a principal component analysis we obtained the

following gait factors which we used by DSI: **Rhythm**, **Variability**, **Phases**, **Pace**, **Base of Support**, **Tandem**, and **Turning**.[27]

Baseline cognitive function

We included the following **objective memory** and **non-memory** cognitive tests: **15-Word Learning Test imme- diate** and **delayed recall**,[28] **Stroop tests** (**reading**, **color-naming** and **interference**),[29, 30] **The Letter-Digit Substitution Task**,[31] **Word Fluency Test**,[32] and the **Purdue Pegboard test**.[33] **Subjective cognitive complaints** were evaluated by interview. This interview included three questions on mem- ory (difficulty remembering, forgetting what one had planned to do, and difficulty finding words), and three questions on everyday functioning (difficulty managing finances, problems using a telephone, and difficulty getting dressed).[34]

Outcome: definition of cognitive decline

A g-factor was constructed by a principal component analysis on the delayed recall score of the 15-Word Learning Test, Stroop interference Test, Letter-Digit Substitution Task, Word Fluency Test, and the Purdue Pegboard test.[34] Cognitive decline was defined by the g-factor from the follow-up visit minus the g-factor from the baseline visit resulting in a delta g-factor. Since the follow-up time was not the same for each participant, the delta g-factor was divided by the follow-up time to obtain global cognitive decline per year. Significant global cognitive decline (yes/no) was defined as belonging to the 5% of participants with the highest cognitive decline (delta g-factor) per year. In the used dataset, consisting of 2,542 participants, this resulted in 127 participants with a positive class label.

Evaluation experiments

Prediction performance evaluation

The performance of DSI in predicting occurrence of global cognitive decline was evaluated using cross-validation. The area under the receiver-operator curve (AUC) was determined using 500 repetitions of 2-fold cross-validation (CV) experiments. This means that with each repetition 50% of the study dataset was used for training and the other 50% was used for testing, and vice versa, keeping the class ratio in the training and test set equal. We report the mean AUC, and the uncertainty of the mean expressed by its 95% confidence interval, derived from the 1000 resulting AUC values. The confidence interval was determined with the corrected resampled t-test for CV estimators of the generalization error.[35] AUCs were considered significantly different if the 95% confidence interval of their difference did not contain zero.

Since global cognitive decline per year is age dependent, we expect that age is an important feature for the prediction. We therefore include **age** as feature in the model.

However, since other features might depend on age, correcting these features might improve the prediction performance.[36] We therefore also assessed the prediction performance using age-corrected features. We corrected the non-binary features for age using a linear regression model.[37] We evaluated four different models:

I age was included and no age-correction was performed on thenon-binary features
II age was excluded and no age-correction was performed on the non-binary features
III age was included and non-binary features, except age, were corrected for age
IV age was excluded and non-binary features, except age, were corrected for age.

To assess whether the performance of DSI was dependent on the combination of input features, we evaluated various feature combinations. In each cross-validation experiment the feature set was expanded with a feature or category of features. We analyzed four of such cumulative feature sets, differing in the order in which the feature set was expanded. Additionally, we analyzed MRI features separately and a set including all features but age.

Relevance analysis

To gain insight into the relevance weight that DSI assigns to each feature, we calculated the feature relevance distribution over the 500 repetitions of 2-fold CV, for the top-level feature categories of the hierarchy: age, sex, cognitive tests, cardiovascular risk factors, gait, education, genetics, and MRI features.

Feature selection on MRI features

In this study, hundreds of MRI features were extracted from images. It is likely that many of those features are not very efficient in detecting cognitive decline. Typically feature selection is applied to exclude poor features which may induce noise to the classifier. In DSI, weighting with relevance supresses the effect of such features. If the number of features is high, their cumulative effect may, however, be remarkable. Previous results have shown that when including many features with a low relevance, the performance of DSI may decrease.[7] We therefore included an experiment evaluating the effect of feature selection on MRI features using their relevance. Due to averaging, feature noise reduces in higher levels of the feature hierarchy. The relevance of top-level feature categories may therefore be higher than lower-level, individual features. Therefore, due to the selection on the individual features, the top-level features may drop out, despite their high relevance. To prevent entire top-level feature categories to drop out of the model, we chose to only apply feature selection on the MRI features, which made up 80% of all input features, before selection. The relevance of the MRI features was determined on the entire dataset, before training. MRI features were selected by thresholding the relevance. Subsequently, an AUC distribution was

determined in 10 repetitions of 2-fold CV. The following relevance thresholds were chosen: $t \in \{0.0, 0.01, \ldots, 0.09, 0.1\}$. For each threshold we assessed three feature sets in which the relevance-based feature selection on the MRI features was applied: 1) all features, 2) all features but age, and 3) MRI features only.

Sub-group analyses

As subjects close to the decision boundary (DSI ~0.5) are more likely to be misclassified, we evaluated classification performance when only accepting/providing the classification for test subjects with low (<0.2) or high (>0.8) DSI. In this way, the subjects with 0.2 < DSI < 0.8 are disregarded, which, in a clinical case, would mean that there is no diagnosis possible for these cases. We computed the AUC of this sub-group for DSI using all available features, both with age-correction and without age-correction. Furthermore, we performed a sensitivity analysis in which the diffusion-MRI of 680 participants in RS cohort II were ignored, because this data was obtained on average 3.47± 0.15 years later than the other baseline MRI features.

RESULTS

Table 1 presents the characteristics of the study population. The mean age of the participants was 60.9 ± 9.1 years and 55.6% were females.

Prediction performance

Figure 2a shows the mean AUC (95% confidence interval) for several combinations of features in predicting global cognitive decline, without correcting the non-binary features for age. Each color represents an expanding set of used input features, where the most left set is only MRI features and the most right set is all features except age. When using only MRI features, the AUC was 0.75 (0.70 - 0.80). When using only age as baseline feature, the AUC was 0.78 (0.74 - 0.83). Using additional features on top of age resulted in an equal or slightly lower AUC (differences not statistically significant). When using all available features with DSI, the AUC was 0.77 (0.72 - 0.82). The mean AUC of DSI without age as baseline predictor was 0.75 (0.70 - 0.80).

Figure 2b shows the mean AUC (95% confidence interval) for the same combinations of features as in **Figure 2**a, but here the non-binary features were corrected for age. The AUC for MRI features only was significantly lower with age-correction compared to without age correction, with an AUC of 0.55 (0.50 - 0.61). For the other feature sets, the AUC of the models where age correction was applied was not statistically significantly different, compared to not using age correction. When the effect of age was totally removed from the model, i.e. model iv, the AUC was 0.65 (0.58 - 0.73).

Table 1 Baseline characteristics

Feature	R_{nac}	R_{nac}	Positive R_{ac}(N=127)	Control N=2415)
Age, years	0.38	-	71.2 (10.1)	60.3 (8.7)
Sex, female	0.01	-	73 (54.5%)	1340 (55.6%)
Objective cognitive test results	0.28	0.16	-	-
Word Learning Test immediate recall	0.09	0.02	7.7 (2.2)	8.1 (2.0)
Word Learning Test delayed recall	0.05	0.04	7.9 (2.9)	8.2 (2.8)
Reading subtask of Stroop test, s	0.20	0.03	17.2 (2.7)	16.3 (2.9)
Color naming subtask of Stroop test, s	0.18	0.06	23.6 (3.6)	22.3 (4.0)
Interference subtask of Stroop test, s	0.32	0.10	53.8 (20.3)	44.0 (13.0)
Letter-Digit Substitution Task, number of correct digits	0.15	0.00	29.7 (6.7)	32.2 (6.2)
Word Fluency Test, number of animals	0.04	0.08	23.2 (5.7)	23.8 (5.7)
Purdue Pegboard test, number of pins placed	0.15	0.07	10.3 (2.1)	10.9 (1.7)
Mini-mental-state examination	0.14	0.11	27.8 (1.7)	28.4 (1.5)
Education[1]	0.07	0.07	3 (1-3)	3 (2-5)
Cardiovascular risk factors	0.34	0.27	-	-
Alcohol[1], glasses per week	0.06	0.04	3.5 (0.3-5.5)	5.5 (1.0-5.5)
Systolic blood pressure, mmHg	0.24	0.04	146.2 (20.3)	135.9 (19.6)
Diastolic blood pressure, mmHg	0.00	0.02	82.8 (9.4)	82.4 (10.6)
Blood pressure lowering medication	0.26	-	51 (38.3%)	284 (11.9%)
Body Mass Index, kg/m^2	0.07	0.07	28.2 (4.4)	27.4 (4.1)
Serum cholesterol, mmol/L	0.11	0.12	5.4 (0.9)	5.6 (1.1)
HDL-cholesterol, mmol/L	0.04	0.09	1.4 (0.4)	1.5 (0.4)
Lipid lowering medication	0.13	-	46 (34.6%)	510 (21.3%)
Smoking	0.08	0.08	-	-
Never	-	-	49 (36.6%)	746 (31.2%)
Former	-	-	54 (40.3%)	1154 (48.2%)
Current	-	-	31 (23.1%)	492 (20.6%)
Diabetes mellitus, presence	0.09	-	24 (18.2%)	220 (9.2%)
APOE-E4 allele carriership	0.02	-	39 (30.2%)	639 (28.3%)
MRI features	0.41	0.25	-	-
Intra-cranial volume, mL	0.03	0.00	1137 (119)	1144 (113)
Brain tissue volume	0.38	0.08	-	-
White matter volume, mL	0.13	0.01	390 (60)	419 (57)
Gray matter volume, mL	0.10	0.01	522 (54)	537 (52)
CSF volume, mL	0.29	0.07	223 (53)	186 (46)
Brain region volume	0.35	0.12	-	-
Hippocampus volume, mL	0.23	0.09	6.4 (0.8)	6.8 (0.7)
White matter lesion volume[1], mL	0.31	0.08	4.5 (2.5-9.4)	2.4 (1.4-4.3)
Cerebral microbleeds, presence	0.09	-	33 (24.6%)	370 (15.6%)
Lacunar infarcts, presence	0.04	-	10 (7.5%)	72 (3.0%)
Global FA	0.17	0.07	0.3 (0.02)	0.3 (0.01)

Table 1 Baseline characteristics (*continued*)

Feature			Positive	Control
	R_{nac}	R_{nac}	R_{ac}(N=127)	N=2415)
Global MD,10^{-3} mm^2/s	0.33	0.07	0.8 (0.03)	0.7 (0.03)
Global cortical thickness, mm	0.08	0.01	2.4 (0.2)	2.5 (0.1)
Gait	0.19	0.17	-	-

Baseline features of the study population and their relevances. The relevances were computed on the entire dataset. Continuous variables are presented as mean (standard deviation) and categorical variables as *n* (percentages).*Abbreviations*: N; number of participants, HDL; high-density lipoprotein, s; seconds, FA; fractional anisotropy, MD; mean diffusivity 10^{-3} mm2/s. CSF; cerebrospinal fluid. *Symbols*: R_{nac}; relevance when feature was not corrected for age, R_{ac}; relevance when feature was age-corrected. 1Education, alcohol and white matter lesion volume are presented as median (inter-quartile range).

Relevance analysis

Figure 3 shows the relevance weight per feature category when the non-categorical features were corrected for age prior to computing DSI and without age-correction. Without age-correction, the features with the best discriminating abilities according to their relevance weights were MRI features (0.42 (0.33 - 0.51)), age (0.39 (0.27 - 0.51)), cognitive tests (0.35 (0.24 - 0.45)) and cardiovascular risk factors (0.34 (0.26 - 0.43)). When correcting the non- binary features, except age, for age, the most discriminating features were age (0.39 (0.27 - 0.51)), MRI features (0.37 (0.24 - 0.51)), and cognitive tests (0.32 (0.17 - 0.47)).

Feature selection on MRI features

Feature selection for MRI features had no effect on the AUC in any of the three feature sets, when the non-binary features were not corrected for age (**Figure 4A)**. The AUC did increase after MRI feature selection when the non-binary features, except age, had been corrected for age (**Figure 4B**), with the optimal *t* being 0.07 (see **Figure 4**b). For $t = 0.07$, the AUC increased from 0.55 (0.50 - 0.61) to 0.62 (0.58 - 0.67) when only MRI features were included in the model. When using all features, the AUC increased from 0.75 (0.70 - 0.79) to 0.77 (0.73 - 0.82), and when using all features but age, the AUC increased from 0.65 (0.58 - 0.73) to 0.70 (0.63 - 0.76).

Sub-group analyses

When only taking into account the extreme cases, i.e. cases for which $0.2 < DSI < 0.8$ (~40% of the total dataset, i.e.~1000 subjects), the mean AUC increased to 0.82 (0.76 - 0.88) using age as input feature only. Again in this group, additional features did not significantly improve the performance of DSI (results not shown). Ignoring the diffusion-MRI features of 680 participants of whom this data was acquired on average 3.47 ± 0.15 years later than the assessment of the other baseline MRI features did not change AUC significantly compared to the performance in the total population (results not shown).

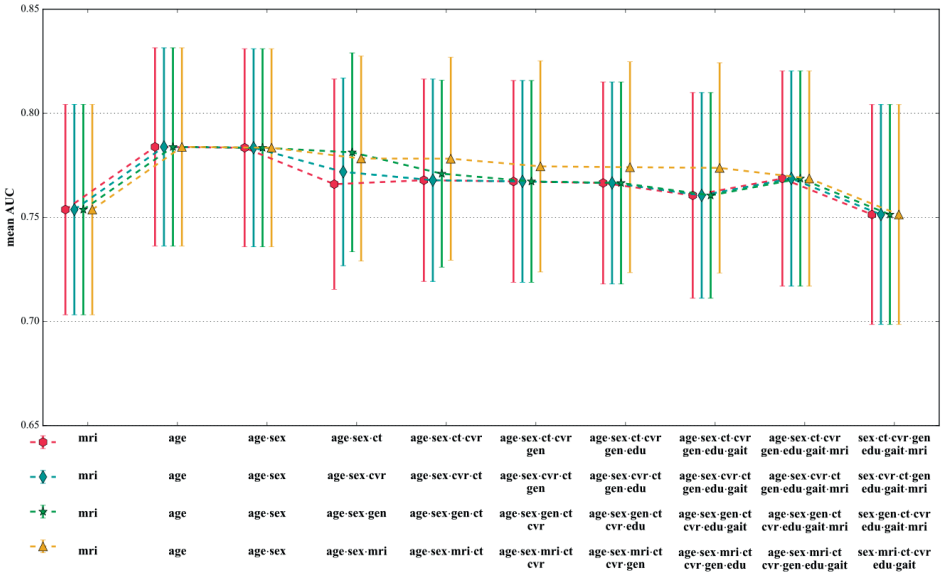

Figure 2A. Mean AUC for several combinations of features without correcting the non-binary features for age**.** Features are accumulated in four different orders, indicated by color and symbol. The bars indicate the confidence interval. Short-hand notations are used for several features: cognitive tests (ct), cardiovascular risk factors (cvr), MRI features (mri), genetics (APOE-E4 carrier-ship) (gen), and educational level (edu)

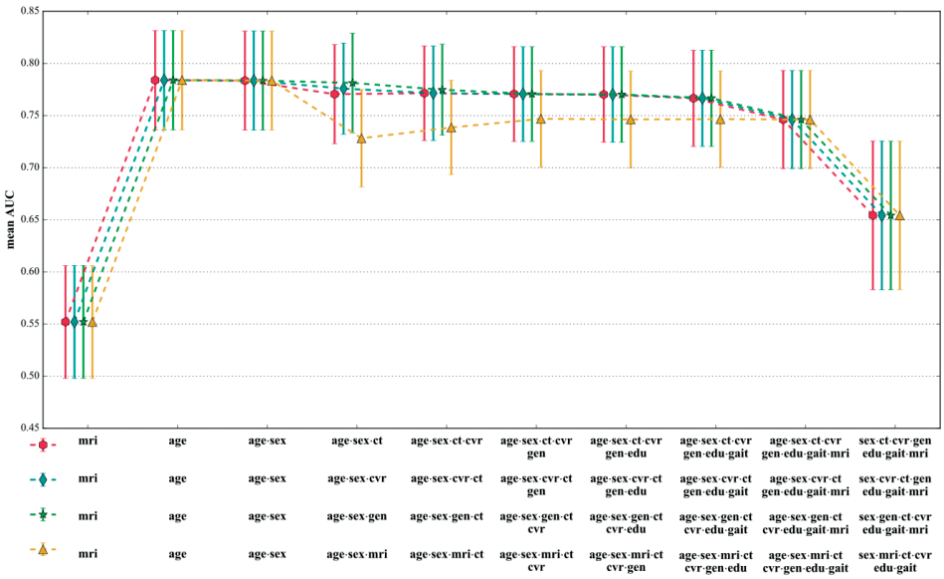

Figure 2B is with the non-binary features corrected for age. Features are accumulated in four different orders, indicated by color and symbol. The bars indicate the confidence interval. Short-hand notations are used for several features: cognitive tests (ct), cardiovascular risk factors (cvr), MRI features (mri), genetics (APOE-E4 carrier-ship) (gen), and educational level (edu).

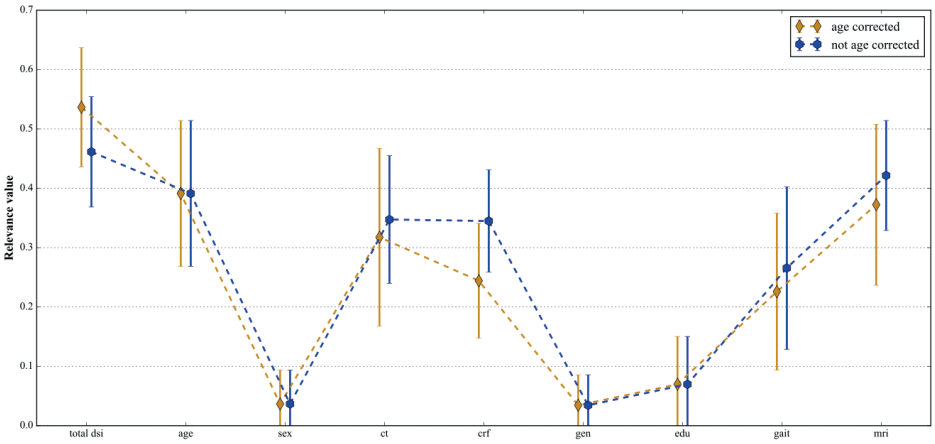

Figure 3. Mean relevance weight *R* and 95% confidence interval for the top-level features categories. The blue line shows the case where the non-binary features were non corrected for age and the golden line shows the case where the non-binary features were age-corrected.

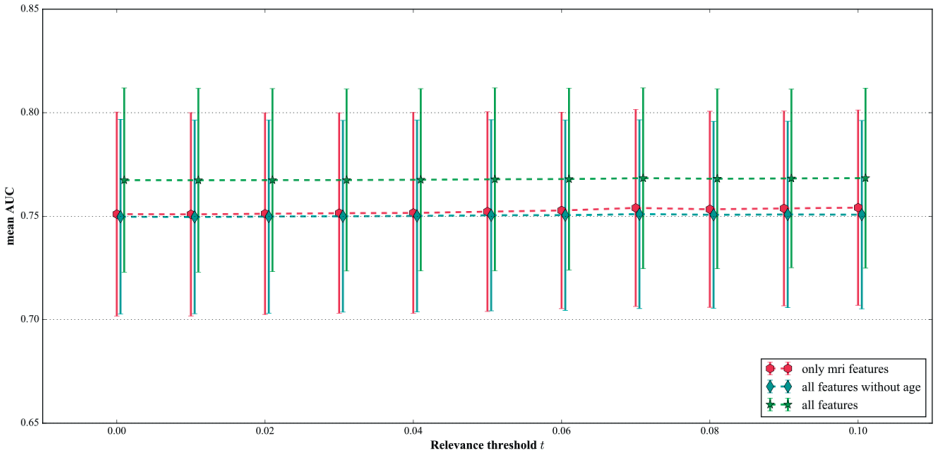

Figure. 4A. Mean AUC for several combinations of features where the MRI features were selected based on their relevance without correcting the non-binary features for age**.**

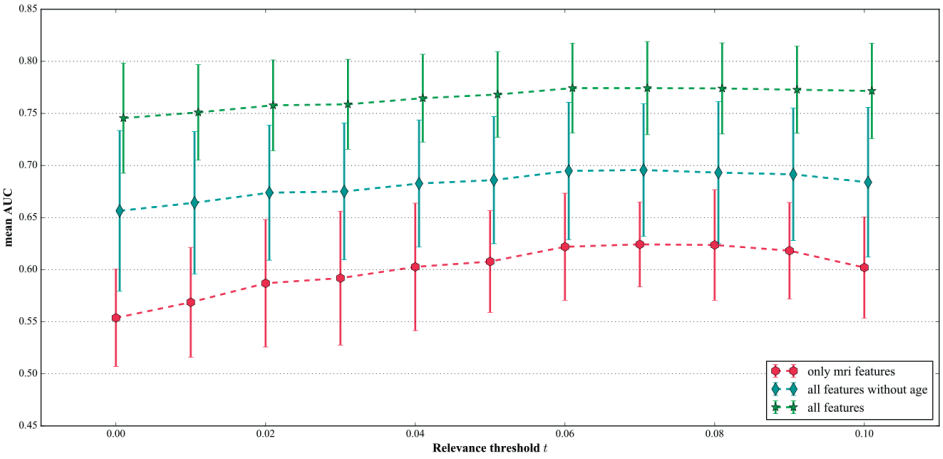

Figure 4B is with the non-binary features corrected for age. Features with $R < t$ were excluded.

DISCUSSION

The objective of this study was to assess whether global cognitive decline can be predicted using multi-variate data with the previously proposed DSI. We found the best prediction performance, evaluated with AUC, using only age as input feature. Adding more features to DSI did not improve its performance in predicting global

cognitive decline as defined in this study. Overall performance of DSI in the prediction of global cognitive decline (mean AUC 0.78) was comparable to previously reported performances of DSI for prediction of dementia[7] in the population-based CAIDE study, consisting of 2000 participants who were randomly selected from four separate, population-based samples, originally studied in midlife (1972, 1977, 1982, or 1987),[38] and to other population-based prediction models of dementia.[39] In this study we included a large number of heterogeneous features. Age was the most important feature for predicting global cognitive decline using DSI, yielding the highest AUC. This was further supported by the observation that the performance of DSI reduced when using all features except age. Our finding that age is the single strongest predictor for cognitive decline is in line with published prediction models for dementia, that invariably assign the highest weight to age.[40] We found that the relevance R, which indicates how well a feature can discriminate between persons who will develop cognitive decline and those who will not, was highest for MRI features (0.42) followed by age (0.39). DSI, however, performed worse when using only MRI features, compared to using only age. We speculate that the high relevance of the MRI features may be explained by age-specific effects that are captured in these MRI features, which is supported by our finding that MRI feature relevance (0.37) and DSI performance dropped when adjusting MRI features for age. When the non-binary features were age-corrected and age was not included in the model, the mean AUC was 0.65, still significantly better than chance (0.5), indicating that relevant information for predicting global cognitive decline could be present in the other features. In this study, however, they did not improve the predicting performance when added to age. To our surprise we found that APOE-4 allele carrier-ship had a low relevance weight and did not improve the performance of DSI, even though it is the best known genetic risk factor for AD. This is in contrast to a previous study focusing on the progression from MCI to AD, which found APOE-genotype to have high predictive value.[41] It may be that our study population was too young to show effect of APOE on prediction (mean age 60.9), since the risk progression effect of APOE-4 allele carriership has been described to peak between ages 70 and 75 years.[42] The relevance-based feature selection on the MRI features showed an increase in the AUC, but only when the nonbinary features were corrected for age. A possible explanation is that without age correction, the AUC is strongly driven by the age-factor that is present in the MRI features. In this case, less and different features were excluded compared to the age-corrected models, causing the selection to have no effect on the prediction performance. However, after removal of these age-specific effects by age correction, performance can be increased by removal of irrelevant features. When age was totally excluded from the model iv (age was excluded and age correction was applied to the non-binary features), an AUC of 0.70 was obtained, showing the potential of the other features. One limitation of

this analysis is that the relevance computation and threshold selection was done on the entire dataset, i.e. the training data was included in these computations. Therefore, AUC increase due to application of the relevance threshold might be overestimated, but can be seen as an upper limit. The overall conclusions do not change.
To our knowledge, this is the first population-based study testing the supervised machine learning DSI tool for prediction of global cognitive decline. Strengths of our study include the population-based design, large sample size and availability of an extensive set of features. However, limitations of our dataset need to be considered. We constructed a g-factor as a measure of global cognition and participants without complete cognitive data were excluded.This might have caused some selection bias towards relatively healthy subjects. Also, mortality and drop-out was not taken into account. Persons who are lost to follow-up usually have a poorer health status and are therefore more likely to develop cognitive decline or die before onset of cognitive decline. The exclusion of these ssumingly more severe cases might have lowered the performance of DSI. The result that age is the main predictor for cognitive decline indicates that the age distribution of the subjects with cognitive decline differs from the entire set of subjects. Hence age could be used to select people at risk of cognitive decline. However, when screening for significant cognitive decline, an age-dependent threshold on cognitive decline might be needed, e.g. using the 5% percentile of the cognitive decline as function of age, to detect young people at risk of developing dementia. The usage of such an age-dependent threshold will be part of future research.
Finally, it should be noted that cognitive decline is not equivalent to neurodegeneration/dementia and may result from other causes 260 as well, due to conditions affecting the participant's cognition at the time of the cognitive assessment, normal human variability and normal aging. Nevertheless, being able to predict cognitive decline would be a step forward in selecting people for therapy or prevention.

Conclusion and future work

Based on our results we can conclude that age is the most important predictor for cognitive decline in the general population using DSI. Other features showed having potential, but did not improve prediction performance. A next step could be to use longitudinal features in DSI, as these might improve its prediction performance. To validate whether our findings are not due to limitations of DSI, also other methods need to be evaluated in this prediction challenge. Finally, to be able to detect younger people at risk of significant global cognitive decline in future studies, thresholds for cognitive decline should be carefully chosen depending on the population, for example be age-adjusted.

CHAPTER REFERENCES

1. C.R. Jack, D.S. Knopman, W.J. Jagust, R.C. Petersen, M.W. Weiner, P.S. Aisen, L.M. Shaw, P. Vemuri, H.J. Wiste, S.D. Weigand, T.G. Lesnick, V.S. Pankratz, M.C. Donohue and J.Q. Trojanowski, Tracking pathophysiological processes in Alzheimer´s disease: an up300 dated hypothetical model of dynamic biomarkers, Lancet Neurology 12(2) (2013), 207–216.
2. S. Kloppel, C.M. Stonnington, J. Barnes, F. Chen, C. Chu, C.D. Good, I. Mader, L.A. Mitchell, A.C. Patel, C.C. Roberts, N.C. Fox, C.R. Jack, J. Ashburner and R.S. Frackowiak, Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method, Brain 131(11) (2008), 2969–2974.
3. J. Mattila, J. Koikkalainen, A. Virkki, A. Simonsen, M. van Gils, G. Waldemar, H. Soininen, J. Lötjönen and the Alzheimer's Disease Neu305 roimaging Initiative., A disease state fingerprint for evaluation of Alzheimer's disease, Journal of Alzheimer's Disease 27(1) (2011), 163–176.
4. J. Mattila, H. Soininen, J. Koikkalainen, D. Rueckert, R.Wolz, G.Waldemar and J. Lötjönen, Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects, Journal of Alzheimer's Disease 32(4) (2012), 969–979.
5. A. Hall, J. Mattila, J. Koikkalainen, J. Lötjönen, R. Wolz, P.H. Scheltens, G.B. Frisoni, M. Tsolaki, F. Nobili, Y. Freund-Levi, L. Minthon, L. Frölich, H.J. Hampel, P.T. Visser and H. Soininen, Predicting progression from cognitive impairment to Alzheimer's disease with the Disease State Index, Current Alzheimer Research 12(1) (2015), 69–79.
6. M.A.M. noz-Ruiz, A. Hall, J. Mattila, J. Koikkalainen, S.K. Herukka, R. Vanninen, Y. Liu, J. Lötjönen, H.S. H and the Alzheimer's Disease Neuroimaging Initiative., Comparing predictors of conversion to Alzheimer's disease using the disease state index, Neurodegenerative Diseases 13(2–3) (2014), 200–202. 7 T. Pekkala, A. Hall, J. Lötjönen, J. Mattila, H. Soininen, T. Ngandu, T. Laatikainen, M. Kivipelto and A. Solomon, Development of a Late-Life Dementia Prediction Index with Supervised Machine Learning in the Population-Based CAIDE Study, Journal of Alzheimer's Disease 55(3) (2017), 1055–1067.
8. J.A. Blumenthal, P.J. Smith, S. Mabe, A. Hinderliter, K. Welsh-Bohmer, J.N. Browndyke, P.H. Lin, W. Kraus, P.M. Doraiswamy, J. Burke and A. Sherwood, Lifestyle and Neurocognition in Older Adults with Cardiovascular Risk Factors and Cognitive Impairment, Psychoso-matic Medicine 79(6) (2017), 719–727.
9. E.P.M. van Charante, E. Richard, L.S. Eurelings, J.W. van Dalen, S.A. Ligthart, E.F. van Bussel, M.P. Hoevenaar-Blom, M. Vermeulen and W.A. van Gool, Effectiveness of a 6-year multidomain vascular care intervention to prevent dementia (preDIVA): a cluster-randomised controlled trial, Lancet 388(10046) (2016), 797–805.
10. T. Ngandu, J. Lehtisalo, A. Solomon, E. Levälahti, S. Ahtiluoto, R. Antikainen, L. Bäckman, T. Hänninen, A. Jula, T. Laatikainen, J. Lind325 ström, F. Mangialasche, T. Paajanen, S. Pajala, M. Peltonen, R. Rauramaa, A. Stigsdotter-Neely, T. Strandberg, J. Tuomilehto, H. Soininen and M. Kivipelto, 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 385(9984) (2015), 2255–2263.
11. J. Mattila, J. Koikkalainen, A. Virkki, M. van Gils, J. Lötjönen and for the Alzheimer's Disease Neuroimaging Initiative, Design and Application of a Generic Clinical Decision Support System for Multiscale Data, IEEE Transactions on Biomedical Engineering 59(1) (2012), 234–240.
12. M.A. Ikram, G.G. Brusselle, S.D. Murad, C.M. van Duijn, O.H. Franco, A. Goedegebure, C.C.W. Klaver, T.E.C. Nijsten, R.P. Peeters, B.H. Stricker, H. Tiemeier, A.G. Uitterlinden, M.W. Vernooij

and A. Hofman, The Rotterdam Study: 2018 update on objectives, design and main results, European Journal of Epidemiology 32(9) (2017), 807–850.

13. M.A. Ikram, A. van der Lugt, W.J. Niessen, P.J. Koudstaal, G.P. Krestin, A. Hofman, D. Bos and M.W. Vernooij, The Rotterdam Scan Study: design update 2016 and main findings, European Journal of Epidemiology 30(12) (2015), 1299–1315.
14. A. Hofman, G.G. Brusselle, S.D. Murad, C.M. van Duijn, O.H. Franco, A. Goedegebure, M.A. Ikram, C.C.K.T.E.C. Nijsten, R.P. Peeters, B.H. Stricker, H.W. Tiemeier, A.G. Uitterlinden and M.W. Vernooij, The Rotterdam Study: 2016 objectives and design update, European Journal of Epidemiology 30(8) (2015), 661–708.
15. H.A. Vrooman, C.A. Cocosco, F. van der Lijn, R. Stokking, M.A. Ikram, M.W. Vernooij, M.M. Breteler and W.J. Niessen, Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification, NeuroImage 37(1) (2007), 71–81.
16. R. de Boer, H.A. Vrooman, F. van der Lijn, M.W. Vernooij, M.A. Ikram, A. van der Lugt, M.M. Breteler and W.J. Niessen, White matter lesion extension to automatic brain tissue segmentation on MRI, NeuroImage 45(4) (2009), 1151–1161.
17. A.M. Dale, B. Fischl and M.I. Sereno, Cortical surface-based analysis. I. Segmentation and surface reconstruction, NeuroImage 9(2) (1999), 179–194.
18. R.S. Desikan, F. Ségonne, B. Fischl, B.T. Quinn, B.C. Dickerson, D. Blacker, R.L. Buckner, A.M. Dale, R.P. Maguire, B.T. Hyman, M.S. Albert and R.J. Killiany, An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, NeuroImage 31(3) (2006), 968–980.
19. B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo,N. Makris, B. Rosen and A.M. Dale, Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain, Neuron 33(3) (2002), 341–355.
20. M.W. Vernooij, A. van der Lugt, M.A. Ikram, P.A.Wielopolski, H.A. Vrooman, A. Hofman, G.P. Krestin and M.M. Breteler, Total cerebral blood flow and total brain perfusion in the general population: the Rotterdam Scan Study, Journal of Cerebral Blood Flow & Metabolism 28(2) (2008), 412–419.
21. G. Roob, R. Schmidt, P. Kapeller, A. Lechner, H.P. Hartung and F. Fazekas, MRI evidence of past cerebral microbleeds in a healthy elderly population, Neurology 52(2) (1999), 991–994.
22. M.W. Vernooij, A. van der Lugt, M.A. Ikram, P.A. Wielopolski, W.J. Niessen, A. Hofman, G.P. Krestin and M.M. Breteler, Prevalence andrisk factors of cerebral microbleeds: the Rotterdam Scan Study, Neurology 70(14) (2008), 1208–1214.
23. M. de Groot, M.A. Ikram, S. Akoudad, G.P. Krestin, A. Hofman, A. van der Lugt, W.J. Niessen and M.W. Vernooij, Tract-specific white matter degeneration in aging: the Rotterdam Study, Alzheimer's & Dementia 11(3) (2015), 321–330.
24. S. Akoudad, M.L. Portegies, P.J. Koudstaal, A. Hofman, A. van der Lugt, M.A. Ikram and M.W. Vernooij, Cerebral Microbleeds Are Associated With an Increased Risk of Stroke: The Rotterdam Study, Circulation 132(6) (2015), 509–516.
25. P.R. Wenham, W.H. Price and G. Blandell, Apolipoprotein E genotyping by one-stage PCR, Lancet 337(8750) (1991), 1158–1159.
26. L. Lahousse, V.J. Verlinden, J.N. van der Geest, G.F. Joos, A. Hofman, B.H. Stricker, G.G. Brusselle and M.A. Ikram, Gait patterns in COPD: the Rotterdam Study, European Respiratory Journal 46(1) (2015), 88–95.

27. V.J. Verlinden, J.N. van der Geest, Y.Y. Hoogendam, A. Hofman, M.M. Breteler and M.A. Ikram, Gait patterns in a community-dwelling population aged 50 years and older, Gait Posture 37(4) (2013), 500–505.
28. M.L. Bleecker, K. Bolla-Wilson, J. Agnew and D.A. Meyers, Age-related sex differences in verbal memory, Journal of Clinical Psychology 44(3) (1988), 403–411.
29. I. Goethals, K. Audenaert, F. Jacobs, E. Lannoo, C.V. de Wiele, H. Ham, A. Otte, K. Oostra and R. Dierckx, Cognitive neuroactivation using SPECT and the Stroop Colored Word Test in patients with diffuse brain injury, Journal of Neurotrauma 21(8) (2004), 1059–1069.
30. C.J. Golden, Identification of brain disorders by the Stroop Color and Word Test, Journal of Clinical Psychology 32(3) (1976), 654–658.
31. M.D. Lezak, Neuropsychological assessment in behavioral toxicology–developing techniques and interpretative issues, Scandinavian Journal of Work, Environment & Health 10(Suppl 1) (1984), 25–29.
32. K.A. Welsh, N. Butters, R.C. Mohs, D. Beekly, S. Edland, G. Fillenbaum and A. Heyman, The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part V. A normative study of the neuropsychological battery, Neurology 44(4) (1994), 609–614.
33. J. Desrosiers, R. Hébert, G. Bravo and E. Dutil, The Purdue Pegboard Test: normative data for people aged 60 and over, Disability and Rehabilitation 17(5) (1995), 217–224.
34. Y.Y. Hoogendam, A. Hofman, J.N. van der Geest, A. van der Lugt and M.A. Ikram, Patterns of cognitive function in aging: the Rotterdam Study, European Journal of Epidemiology 29(2) (2014), 133–140.
35. C. Nadeau and Y. Bengio, Inference for the Generalization Error, Machine Learning 52(3) (2003), 239–281.
36. F. Falahati, D. Ferreira, H. Soininen, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, S. Lovestone, M. Eriksdotter, L.-O. Wahlund, A. Simmons, E. Westman, for the AddNeuroMed consortium and the Alzheimer´s Disease Neuroimaging Initiative, The effect of age correction on multivariate classification in Alzheimer´s disease, with a focus on the characteristics of incorrectly and correctly classified subjects, Brain Topography 29(2) (2016), 296–307.
37. J. Koikkalainen, H. Pölönen, J. Mattila, M. van Gils, H. Soininen, J. Lötjönen and the Alzheimer's Disease Neuroimaging Initiative, Improved Classification of Alzheimer's Disease Data via Removal of Nuisance Variability, PLoS One 7(2) (2012), 31112.
38. M. Rusanen, M. Kivipelto, E. Levälahti, T. Laatikainen, J. Tuomilehto, H. Soininen and T. Ngandu, Heart diseases and long-term risk of dementia and Alzheimer's disease: a population-based CAIDE study., Journal of Alzheimer's disease 42(1) (2014), 183–191.
39. M. Kivipelto, T. Ngandu, T. Laatikainen, B. Winblad, H. Soininen and J. Tuomilehto, Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study, Lancet Neurology 5(9) (2006), 735–741.
40. B.C. Stephan, C. Tzourio, S. Auriacombe, H. Amieva, C. Dufouil, A. Alpérovitch and T. Kurth, Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study, The British Medical Journal 350 (2015), 2863.
41. L.C. Löwe, C. Gaser, K. Franke and for the Alzheimer's Disease Neuroimaging Initiative, The Effect of the APOE Genotype on Individual BrainAGE in Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease, PLoS One 11(7) (2016), 0157514.
42. L.W. Bonham, E.G. Geier, C.C. Fan, J.K. Leong, L. Besser, W.A. Kukull, J. Kornak, O.A. Andreassen, G.D. Schellenberg, H.J. Rosen, W.P. Dillon, C.P. Hess, B.L. Miller, A.M. Dale, R.S. Desikan

and J.S. Yokoyama, Age-dependent effects of APOE_4 in preclinical Alzheimer's disease, Annals of Clinical and Translational Neurology 3(9) (2016), 668–677.