2020-07

A Robust Blood-based Signature of Cerebrospinal Fluid A42 Status

Eke, Chima S.

A Robust Blood-based Signature of Cerebrospinal Fluid Aβ_{42} Status

Chima S. Eke, Fatemah Sakr, E. Jammeh, Peng Zhao, E. Ifeachor

Abstract— **Early detection of AD is of vital importance in the development of disease-modifying therapies. This necessitates the use of early pathological indicators of the disease such as amyloid abnormality to identify individuals at early disease stages where intervention is likely to be most effective. Recent evidence suggests that cerebrospinal fluid (CSF) amyloid $\beta_{1\text{-}42}$ ($A\beta_{42}$) level may indicate AD risk earlier compared to amyloid positron emission tomography (PET). However, the method of collecting CSF is invasive. Blood-based biomarkers indicative of CSF $A\beta_{42}$ status may remedy this limitation as blood collection is minimally invasive and inexpensive. In this study, we show that APOE4 genotype and blood markers comprising EOT3, APOC1, CGA, and $A\beta_{42}$ robustly predict CSF $A\beta_{42}$ with high classification performance (0.84 AUC, 0.82 sensitivity, 0.62 specificity, 0.81 PPV and 0.64 NPV) using machine learning approach. Due to the method employed in the biomarker search, the identified biomarker signature maintained high performance in more than a single machine learning algorithm, indicating potential to generalise well. A minimally invasive and cost-effective solution to detecting amyloid abnormality such as proposed in this study may be used as a first step in a multi-stage diagnostic workup to facilitate enrichment of clinical trials and population-based screening.**

I. Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disease accounting for over 60% of all dementia cases [1]. It is characterized in part by the accumulation of amyloid-beta ($A\beta_{42}$) plaques in the brain – a condition known as amyloid pathology – that is present long before clinical symptoms (cognitive) are apparent [2, 3].

No cure or disease-modifying treatment for AD currently exists. There are ongoing efforts in clinical trials to combat this challenge. Current clinical trials target individuals at the earliest stages of AD, where intervention is thought to be most likely successful, following the high failure rates of previous trials [4]. Amyloid screening is used in these trials to identify individuals with amyloid pathology and may therefore be at the early stages of the disease before symptom onset. It may also be beneficial in the future for population-based screening [5, 6].

Current validated biomarkers of abnormal amyloid accumulation include Aβ positron emission tomography (PET) and Aβ42 measurement in cerebrospinal fluid (CSF) [7]. Use of these markers is internationally recommended [3, 8]. However, PET scans are expensive and available only at specialized centres while lumbar punctures required for CSF testing are invasive.

Notwithstanding the invasiveness of CSF collection, there is growing evidence that CSF $A\beta_{42}$ may be an earlier indicator of AD pathology compared to $A\beta_{42}$ PET [9-11] and thus may be a more suitable biomarker for disease detection at the earliest stages. To mitigate the limitation of invasiveness posed by CSF-based amyloid testing, there is strong interest in identifying blood-based biomarkers reflective of amyloid status as would CSF. Such biomarkers may be used as a reliable initial step in a multistage diagnostic procedure.

A few studies [12, 13] have demonstrated the potential of blood-based markers predictive of amyloid status as measured by CSF $A\beta_{42}$ with area under receiver operating curve (AUC) reaching 0.88 (in 46 samples) and 0.81 (in 358 samples), respectively. However, the novel method employed by [12] in measuring the blood-based markers remains to be established and the results from [13] are yet to be validated in independent cohorts.

In this study, we explore the utility of blood-based proteins to predict CSF $A\beta_{42}$ status using support vector machines with recursive feature elimination (SVM-RFE) that has shown effectiveness in similar research domains [14]. We also give particular consideration to the robustness of identified markers, to enhance the likelihood of reproducing results since reproducibility of results is one of the challenges in AD blood biomarker discovery domain [15].

II. METHODS

A. Study data preparation

Baseline data of 566 individuals from Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort were obtained (http://adni.loni.ucla.edu). The data comprised blood-based measurement of 190 proteins analyzed on a Rule-Based Medicine platform and 3 other proteins (including homocysteine, $A\beta_{40}$, and $A\beta_{42}$. The data also included apolipoprotein E ε4 (APOE4) genotype, demographic and diagnostic information as well as CSF $A\beta_{42}$ levels of the individuals measured on the Luminex Xmap platform. Forty-four (44) of the proteins were later excluded due to missingness, leaving 149 proteins. Finally, data from 358 individuals remainedafter 208 were dropped

*Research funded by the EU H2020 Marie Sklowdoska Curie Actions (MSCA) through the BBDiag project (number 721281).

C. S. Eke (corresponding author) is with School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth, Devon, PL4 8AA, UK (email: chimastnaley.eke@plymouth.ac.uk).

Fatemah Sakr is with University of Medicine Rostock and DZNE German Center for Neurodegenerative Diseases, Gehlsheimer Str. 20, 18147 Rostock, Germany (email: fatemah.sakr@med.uni-rostock.de).

E. Jammeh is with NIHR Sheffield Biomedical Research Centre, University of Sheffield, 385a Glossop Road, Sheffield, S10 2HQ (e.a.jammeh@sheffield.ac.uk).

Peng Zhao is with School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth, Devon, PL4 8AA, UK (email: peng.zhao@plymouth.ac.uk).

E. Ifeachor is with School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth, Devon, PL4 8AA, UK (email: e.ifeachor@plymouth.ac.uk).

TABLE I.
DEMOGRAPHIC CHARACTERISTICS OF STUDY SUBJECTS

	Clinical Diagnosis		
	CN	***MCI***	***ADD***
Number of participants (n)	58	198	102
Age (mean, (SD))	75.11(5.77)	74.37(7.49)	74.86(7.88)
Gender, female (n, (%))	28(48.28)	65(32.83)	43(42.16)
Years of education (mean, (SD))	15.67(2.78)	15.80(2.99)	15.16(3.30)
APOE4 carriers (n, (%))	5(8.62)	106(53.56)	71(69.61)
Low CSF $A\beta_{42}$ status (n, (%))	1(1.72)	147(74.24)	93(91.18)

CN: Normal control; MCI: Mild cognitive impairment; ADD: Alzheimer's dementia; n: Number of subjects; SD: Standard deviation.

due to missing CSF $A\beta_{42}$ levels. CSF $A\beta_{42}$ status for the remaining individuals (TABLE I) was obtained by dichotomizing their CSF $A\beta_{42}$ levels as normal (high) or abnormal (low) according to clinically recognized threshold of 192pg/ml for the Luminex platform.

B. Robust biomarker selection

The objective here was to identify potential blood biomarker signatures predictive of CSF $A\beta_{42}$ status, from which a signature can be selected based on robustness and performance. The measure of robustness was intended to be transparent and simple to evaluate. The method used is based on the approach proposed by *Abeel et al*. [16] with some modification.

Similar to [16], SVM-RFE [14] combined with ensemble technique was used to select features for signatures formation, while Kuncheva index (KI) [17] was used to evaluate robustness of signatures. SVM-RFE combines the embedded feature selection capability of linear SVM with backward feature elimination strategy of RFE. Absolute values of the weights (coefficients) the linear SVM provides is the contribution of each feature to the SVM hyperplane and may be used a means of ranking the importance of individual features. A feature with a larger weight is regarded as one of higher importance, and one with a lower weight is considered less important.

RFE implements a backward feature elimination procedure that iteratively removes the least important features in the training data samples. The algorithm starts out by fitting the training data with all the available features to a linear SVM, then ranks the features according to their weights and eliminates the least important one(s). The training data is subsequently refitted to the linear SVM but with only the retained features. This process is repeated until all features have been eliminated or a desired number of features to retain is attained. Finally, each feature in the training data is assigned an overall rank ***r*** (an integer with 1 as minimum and dimension of training data D as maximum) according to the observed feature contributions, with most significant features assigned lowest ranks.

SVM-RFE with ensemble learning is implemented to improve the robustness (stability) of feature subset selection by SVM-RFE. In this approach, ***k*** different subsamples of the original dataset (of D dimension) are generated using random sampling without replacement, each subsample containing only a slight variation (***p*** samples) of the original dataset. For each subsample (in the ***k*** subsamples), ***b*** bootstrap samples are generated. SVM-RFE provided with a specified signature size ***s*** as a stopping criterion is then applied to each bootstrap. The rank of each feature in D as well as the AUC performance ($\mathrm{AUC_{OO}}$) of the selected features on the out-of-bag samples is recorded. A candidate signature of size ***s*** is subsequently selected according to an ensemble ranking R obtained by aggregating ***r*** over all ***b*** bootstrap samples as shown in (1).

An estimate of the generalization performance of the signature is obtained by training the linear SVM on the subsample and its performance evaluated on the ***1-p*** held out samples. Ensemble method of generating signatures has shown to improve robustness and classification performance compared to simply applying SVM-RFE directly to subsamples [16]. In addition to the approach proposed in [16], we carried out a repeated stratified cross-validation of the candidate signature on the corresponding subsample as a supportive evaluation of the signature's classification performance.

$$\mathrm{R} = \sum_{i=1}^{\boldsymbol{b}} w_i r_i \tag{1}$$

The weight w_i is bootstrap-dependent. It takes either of two values depending on the chosen aggregation method. In the complete linear aggregation (CLA) method, w_i is set to 1, while $w_i = 1 - \mathrm{AUC_{OO}}$ in the complete weighted aggregation (CWA) strategy. The two methods were explored in this study albeit CWA was shown to be marginally better than CLA in [16].

To evaluate the robustness of the ***k*** candidate signatures, a stability measure defined by the Kuncheva index (KI) [17] shown in (2) was applied.

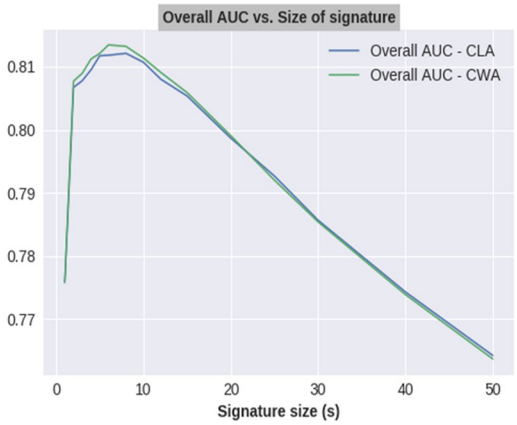
$$\mathrm{KI} = \frac{m - (s^2/k)}{s - (s^2/k)} \tag{2}$$

KI with range [-1, 1] measures the similarity between two signatures. m is the number of features common to both signatures. The greater the value of KI, the larger the number of common features. A negative index indicates that feature intersection is mostly due to chance. The overall stability KI_{tot} of a signature can be defined as the average of all pairwise similarity comparison between the signature and rest of the ***k-1*** signatures.

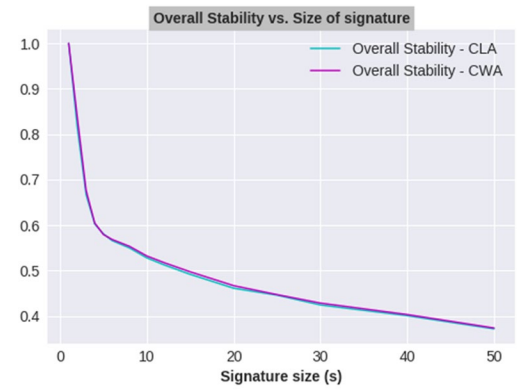
$$\mathrm{KI_{tot}} = \frac{\sum_{i=1}^{\boldsymbol{k-1}} \mathrm{KI}_i}{(k-1)} \tag{3}$$

C. Implementation

The robust biomarker selection task was implemented in python programming language. The machine learning subtasks were conducted with the scikit-learn package. Codes are available at https://github.com/chimastan/robust-blood-based-signature-of-csf-abeta42-status. The values of ***k***, ***b***, and ***p*** used were 500, 50, and 0.8, respectively, considering the recommendations by [16]. Cross-validation fold used was 10-fold with 10 repetitions with samples stratified according to



(a)



(b)

Figure 1. Comparison of (a) classification and (b) stability performance of CLA and CWA-based ensemble methods. The overall AUC and stability are the average AUC and KI_{tot} over the k (500) subsamples.

the target label distribution. The C parameter for the linear SVM was set to default (C=1). In the RFE, number of features to eliminate per run was set to 20% of the total available features to improve speed of processing.

III. RESULTS

A. Potential robust signatures

We realized several potential signatures with different levels of classification and stability performance for prediction of CSF $A\beta_{42}$ status. Fig. 1 illustrates the variation between signature size ***s*** and the average cross-validated AUC as well as average KI_{tot} over the 500 subsamples. It can be seen that the average AUC gradually increased with increasing ***s*** up to a point (***s***≈8) and then declined, while stability steadily dropped with increasing ***s***. The results of CWA and CLA ensemble methods were largely equivalent as shown in Fig.1. Thus all further reports are based on results of the simpler CLA method. Consideration of potential signatures was also limited to ones consisting of 5 biomarkers, being that stability remained moderate at ***s***=5 while the increase in average AUC beyond that point was minimal.

A total of 229 unique candidate signatures were obtained from the 500 subsamples. We then identified the top 10 signatures with best values of stability KI_{tot} (ranging between 0.67 and 0.61) and subsequently carried out further analysis to aid making a final selection.

B. Final selection of signature

We conducted additional analyses with similar approach as in II(B) but with ***s*** limited to 5 and random forests (RF) used as the machine learning algorithm. Therefore in this case, RF-RFE was applied instead of SVM-RFE. The number of trees per forest was set to 2000, each forest containing a maximum of $D^{3/4}$ features as recommended in [18]. The purpose was to obtain candidate signatures with best KI_{tot} values and compare them to the top 10 realized earlier with SVM-RFE. This would allow identifying signatures whose classification and stability performance may be agnostic to type of machine learning algorithm and thus likely to generalize better. With the RF-RFE, we realized 169 unique potential signatures and identified the top 10 with best stability values. A comparison of the signatures with ones obtained with SVM-RFE implicated one signature as common. The signature consists of APOE4 genotype, eotaxin-3 (EOT3), apolipoprotein-C1 (APOC1) and chromogranin-A (CGA), and $A\beta_{42}$. The signature achieved 0.64 stability (KI_{tot}) value. Average AUC, sensitivity, specificity, negative predictive value (PPV) and negative predictive value (NPV) for the repeated 10-fold cross-validation were 0.85, 0.84, 0.63, 0.83 and 0.67, respectively. The average values on the unseen held-out samples were 0.84 AUC, 0.82 sensitivity, 0.62 specificity, 0.81 PPV, and 0.64 NPV, respectively. Contribution of individual biomarkers to the classification performance of the signature is as shown in Fig. 2 with APOE4 unsurprisingly making the most contribution.

IV. DISCUSSION

In this study, we investigated the utility of blood-based signature predictive of CSF $A\beta_{42}$ status with a robust performance. We showed that APOE4 genotype and levels of four proteins predicted CSF $A\beta_{42}$ status with high AUC. This is the first study to demonstrate a signature with a stable performance beyond a single machine learning algorithm. It is a positive indicator of the signature's potential to generalize.

Compared to existing studies, four out of the five predictors (APOE4, CGA, $A\beta_{42}$ and EOT3) in the signature were implicated in a multi-marker panel from a recent study [13] as predictive of CSF $A\beta_{42}$ status with RF. A number of studies have shown evidence of association between some of the identified markers and AD. In line with our observed prominent contribution of APOE4 in the identified signature, it is the strongest and most prevalent genetic risk factor for late-onset AD and considered as a possible therapeutic target [19]. Serum and CSF but not plasma levels of EOT3 have been shown to be dysregulated in individuals with AD [20]. APOC1 genes, in combination with APOE4, are suggested to play an important risk factor role in AD [21, 22]. However, association between plasma levels of APOC1 and AD has not been evidenced. CGA has an amount of co-localisation with brain amyloid plaques [23].

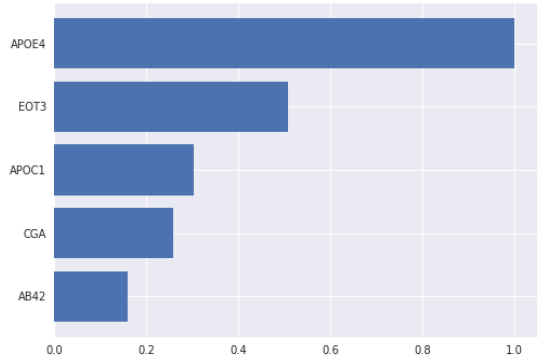


Figure 2. Contribution of individual marker to classification performance of the selected signature. The contribution was determined from the feature weights of linear SVM, normalized by the largest weight during training.

Notwithstanding, CSF and blood levels of CGA have not been reported to be correlated. Interestingly, plasma and CSF Aβ_{42} have shown to be correlated in individuals with AD [24, 25].

This study has several limitations. All analyses were conducted with the ADNI cohort with its peculiarity such as age and level of education of participants. Distribution of individuals with abnormal CSF Aβ_{42} levels across the clinical groups (CN, MCI, and ADD) was biased, with nearly all samples belonging to the MCI or ADD group. This might have impacted our analyses as the individuals are likely to have developed other confounding conditions.

V. Conclusion

Early detection of AD is crucial to the future success of disease modifying therapies which are thought to be most effective at the earliest disease stages. This necessitates the use of early pathological indicators of the disease such as amyloid abnormality. A minimally invasive and cost-effective solution to detecting amyloid abnormality such as proposed in this study may serve as a first step in a multi-stage diagnostic workup to facilitate enrichment of clinical trials and population-based screening.

Acknowledgment

This research was funded by the EU H2020 Marie Skłodowska-Curie Actions (MSCA) through the BBDiag consortium project.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

[1] A. s. Association, "2019 Alzheimer's disease facts and figures," *Alzheimer's & Dementia,* vol. 15, no. 3, pp. 321-387, 2019.

[2] V. L. Villemagne *et al.*, "Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study," *The Lancet Neurology,* vol. 12, no. 4, pp. 357-367, 2013.

[3] B. Dubois *et al.*, "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria," *Alzheimer's & Dementia,* vol. 12, no. 3, pp. 292-323, 2016.

[4] J. Godyń, J. Jończyk, D. Panek, and B. Malawska, "Therapeutic strategies for Alzheimer's disease in clinical trials," *Pharmacological Reports,* vol. 68, no. 1, pp. 127-138, 2016.

[5] H. Zetterberg *et al.*, "Plasma tau levels in Alzheimer's disease," *Alzheimer's research & therapy,* vol. 5, no. 2, p. 9, 2013.

[6] H. Zetterberg and S. C. Burnham, "Blood-based molecular biomarkers for Alzheimer's disease," *Molecular brain,* vol. 12, no. 1, p. 26, 2019.

[7] A. D. Cohen, S. M. Landau, B. E. Snitz, W. E. Klunk, K. Blennow, and H. Zetterberg, "Fluid and PET biomarkers for amyloid pathology in Alzheimer's disease," *Molecular and Cellular Neuroscience,* vol. 97, pp. 3-17, 2019.

[8] C. R. Jack Jr *et al.*, "NIA-AA research framework: toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia,* vol. 14, no. 4, pp. 535-562, 2018.

[9] K. Blennow, N. Mattsson, M. Schöll, O. Hansson, and H. Zetterberg, "Amyloid biomarkers in Alzheimer's disease," *Trends in pharmacological sciences,* vol. 36, no. 5, pp. 297-309, 2015.

[10] S. Palmqvist, N. Mattsson, O. Hansson, and A. s. D. N. Initiative, "Cerebrospinal fluid analysis detects cerebral amyloid-β accumulation earlier than positron emission tomography," *Brain,* vol. 139, no. 4, pp. 1226-1236, 2016.

[11] G. D. Rabinovici, "Amyloid biomarkers: pushing the limits of early detection," *Brain,* vol. 139, no. 4, pp. 1008-1010, 2016.

[12] A. Nakamura *et al.*, "High performance plasma amyloid-β biomarkers for Alzheimer's disease," *Nature,* vol. 554, no. 7691, p. 249, 2018.

[13] B. Goudey, B. J. Fung, C. Schieber, and N. G. Faux, "A blood-based signature of cerebrospinal fluid A β 1–42 status," *Scientific reports,* vol. 9, no. 1, pp. 1-12, 2019.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning,* vol. 46, no. 1-3, pp. 389-422, 2002.

[15] L. Shi *et al.*, "A decade of blood biomarkers for Alzheimer's disease research: an evolving field, improving study designs, and the challenge of replication," *Journal of Alzheimer's Disease,* no. Preprint, pp. 1-18, 2018.

[16] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics,* vol. 26, no. 3, pp. 392-398, 2010.

[17] L. I. Kuncheva, "A stability index for feature selection," in *Artificial intelligence and applications*, 2007, pp. 421-427.

[18] H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn, "Random survival forests for high-dimensional data," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 4, no. 1, pp. 115-132, 2011.

[19] M. Safieh, A. D. Korczyn, and D. M. Michaelson, "ApoE4: an emerging therapeutic target for Alzheimer's disease," *BMC medicine,* vol. 17, no. 1, pp. 1-17, 2019.

[20] A. K. Huber, D. A. Giles, B. M. Segal, and D. N. Irani, "An emerging role for eotaxins in neurodegenerative disease," *Clinical Immunology,* vol. 189, pp. 29-33, 2018.

[21] Q. Zhou *et al.*, "Association between APOC1 polymorphism and Alzheimer's disease: a case-control study and meta-analysis," *PloS one,* vol. 9, no. 1, 2014.

[22] M. Prendecki *et al.*, "Biothiols and oxidative stress markers and polymorphisms of TOMM40 and APOC1 genes in Alzheimer's disease patients," *Oncotarget,* vol. 9, no. 81, p. 35207, 2018.

[23] T. Lechner *et al.*, "Chromogranin peptides in Alzheimer's disease," *Experimental gerontology,* vol. 39, no. 1, pp. 101-113, 2004.

[24] C. E. Teunissen *et al.*, "Plasma Amyloid-β (Aβ 42) Correlates with Cerebrospinal Fluid Aβ 42 in Alzheimer's Disease," *Journal of Alzheimer's Disease,* vol. 62, no. 4, pp. 1857-1863, 2018.

[25] O. Hanon *et al.*, "Plasma amyloid levels within the Alzheimer's process and correlations with central biomarkers," *Alzheimer's & Dementia,* vol. 14, no. 7, pp. 858-868, 2018.