View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Plymouth Electronic Archive and Research Library

Identification of Optimum Panel of Blood-based Biomarkers for Alzheimer's Disease Diagnosis Using Machine Learning

C.S. Eke, E. Jammeh, X. Li, C. Carroll, S. Pearson, and E. Ifeachor

***Abstract*— With the increasing number of people living with Alzheimer's disease (AD), there is a need for low-cost and easy to use methods to detect AD early to facilitate access to appropriate care pathways. Neuroimaging biomarkers (such as those based on PET and MRI) and biochemical biomarkers (such as those based on CSF) are recommended by international guidelines to facilitate diagnosis. However, neuroimaging is expensive and may not be widely available and CSF testing is invasive. Blood-based biomarkers offer the potential for the development of a low-cost and more time efficient tool to detect AD to complement CSF and neuroimaging as blood is much easier to obtain.

Although no single blood biomarker is yet able to detect AD, combinations of biomarkers (also called panels) have shown good results. However, a large number of biomarkers are often needed to achieve a satisfactory detection performance. In addition, it is difficult to reproduce reported results within and across different study cohorts because of data overfitting and lack of access to the datasets used in the studies. In this study, our focus is to identify an optimum panel (in terms of the least number of blood biomarkers to meet the specified diagnostic performance of 80% sensitivity and specificity) based on a widely accessible data set, and to demonstrate a testing methodology that reinforces reproducibility of results. Realizing a panel with reduced number of markers will have significant impact on the complexity and cost of diagnosis and potential development of cost-effective point of care devices.

Index Terms– AD, Biomarkers, ADNI, Machine learning

I. Introduction

Alzheimer's disease (AD) is the leading disease of neurodegeneration in the older population. It is the cause of more than 60% of all dementia cases [1]. The disease is clinically manifest in the form of progressive cognitive and, subsequently, functional decline to a degree that the sufferer's daily living is impaired. More than 47 million people are estimated to suffer dementia worldwide, with an expected increase to 131 million by 2050 [2]. The economic burden of dementia is enormous with the annual global estimated cost above US818 billion and expected to hit a trillion dollars in 2018 [2]. There is a delay of between 3-10 years between symptoms onset and diagnosis of AD, by which time irreversible cell damage would have occurred. There is no cure for AD, but new drugs and therapies are being developed which may slow, halt or reverse the disease processes. Notwithstanding, a significant proportion of dementia patients are undiagnosed as a result of inadequate access to diagnosis. Of those that receive a dementia diagnosis, a significant number may have received it late, when extensive cell damage would have occurred and when treatments are less effective. In view of this, it is thought that providing accessible early diagnosis, may decrease the burden of dementia, facilitate access to evidence based pathway to treatment. It may also facilitate planning and timely receipt of suitable health and social care services [2].

Traditionally, AD was only diagnosed at the clinically symptomatic stage by neuropsychological examination alone. However, research in the last decades has redefined the concept of AD as a clinical-biological entity [3, 4]. This conceptual framework has led to the introduction and utilization of biomarkers in AD assessment. A biomarker is a parameter (physiological, biochemical, or anatomic) that can be objectively measured and evaluated in vivo as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [5, 6]. Recommended biomarkers of AD have been incorporated as supportive evidence in AD diagnostic criteria. These biomarkers are extracted from amyloid positron emission tomography (PET) and cerebrospinal fluid (CSF) analysis. However, CSF is obtained through lumbar puncture, which is widely considered as an invasive procedure, and neuroimaging tests are expensive.

Research studies are investigating several putative AD biomarkers, including ones found in peripheral blood. These areas are receiving closer research attention as huge efforts target identifying and validating AD biomarkers that are minimally invasive, simple to use and cost-effective, and able to reliably discriminate target population in the light of the disease [7, 8]. Blood-based biomarkers may be more cost and time efficient to assess AD. AD blood biomarkers have shown promising results and are anticipated to have the potential to meet these targets and thus foster improved accessibility to diagnosis. However, no single blood biomarker of AD is yet able to provide acceptable diagnostic performance in terms of sensitivity and specificity, only a panel of blood biomarkers may achieve such performance. Consequently, a number of research studies have investigated AD diagnostic performance of some blood biomarker panels. Ray et al. [9] identified an 18-biomarker panel that attained sensitivity and specificity values of 90% and 88% respectively, while a 30-biomarker panel was identified by O'Bryant et al. [10], which achieved sensitivity and specificity values of 94% and 84%, respectively. Daniel et al. [11] identified 5 to 15 biomarker panels that detected AD with 74% sensitivity and 85%

specificity. Using Alzheimer's Neuroimaging Initiative (ADNI) cohort, Doecke et al. [12] identified an 18-marker panel that identified AD with sensitivity and specificity values of 80%. A study by Guo et al.'s [13] obtained sensitivity and specificity values of 89.36% and 79.17%. Recently, Jammeh et al. [14] identified a panel of six blood biomarkers that was able to detect AD with sensitivity and specificity values of 85.4% and 78.6%, respectively.

Despite the progress, some of the identified panels of blood biomarkers consist of a large number of biomarkers, or have not met acceptable performance specification. Furthermore, there are difficulties with replicating results, due to many factors such as overfitting in model development [15]. In addition, some of the studies cannot be replicated because panels were identified using datasets that are difficult to access or are based on biomarkers that are not found in accessible databases. These challenges impede continued investigation of the utility of blood biomarkers in AD diagnosis and progress in identifying blood biomarker panels with clinical utility.

The main objectives of this research are to (1) identify an optimum panel of adequately cross-validated blood biomarkers of AD that can detect Alzheimer's with acceptable diagnostic performance of at least 80% sensitivity and specificity values [8], using widely accessible and well characterized blood proteomic dataset and (2) demonstrate a methodological framework of testing the overall reliability of blood biomarker panels to facilitate replication of results. It is noteworthy that realizing a minimum number of biomarkers that provide high and reliable diagnostic performance may result in reduced complexity and cost of implementation of point of care diagnostic devices for AD.

II. METHODS

The methodology that was adopted in this study is illustrated in Figure 1.

A. Clinical data

Data used in this study were obtained from ADNI (http://adni.loni.ucla.edu). ADNI phase 1 (ADNI1) baseline quality-controlled data of the concentration of 146 blood plasma proteins/biomarkers derived from a cohort of 112 Alzheimer's disease patients and 58 healthy controls were downloaded alongside the demographic and clinical status of the subjects. The Alzheimer's disease subjects used in this study were at the dementia stage of the disease. Summary of the demographics of the subjects is shown in TABLE I. Data from four of the Alzheimer's disease subjects, including 3 that were diagnosed as possible Alzheimer's disease and 1 diagnosed with only mild level of confidence, were excluded from our study, leaving 108 AD subjects. The most cross-validated putative blood biomarkers of AD were identified from the literature, and the ones available in ADNI were used to characterize AD and healthy control (HC) clinical groups.

B. Statistical analysis

The probability distribution of each of the markers used in characterizing the AD patients and HCs was normal and their differential abundance between the two clinical groups was analyzed using Student's t-test. Only the ones with statistically significant difference ($p \leq 0.05$) were used as candidates for the identification of biomarker panels.

C. Computation of possible biomarker panels

A brute force approach was used in making the choice of biomarker combinations to consider. All the possible combinations, also called panels, (consisting of 2, 3, 4, and 5 biomarkers) of the markers selected after the statistical analysis were computed. Each of the possible panels was used for the classification procedure described in Section D. This method of panel identification is different from the usual methods seen in blood biomarker studies, where some sort of reductionist approach is often implemented. The danger with such methods is that some useful biomarker panels may be missed.

D. Classification and biomarker panel selection

Support vector machine (SVM) supervised machine learning algorithm implemented in MATLAB (R2017b) was used to identify an optimum panel of biomarkers that met the desired performance. SVM is a machine learning algorithm that creates a hyperplane between data sets to indicate the class to which they belong. It defines the decision boundary with only a subset of the training points, called the support vectors. SVM has been extensively applied in Alzheimer's research [16, 17].

The classifier algorithm was trained and tested with each of the possible panels of markers described in Section C using 10-fold cross-validation technique. This technique randomly divides the applied dataset into 10 sub-datasets, and ensures that each subset is used for both training and testing. 10-fold cross-validation, implements a mechanism to avoid model overfitting. Apolipoprotein ε4 (APOE4) genotype was used as a covariate to the biomarkers, as it has been reported as one of the major clinical AD risk factors [18].The training and testing of a model with each panel was repeated five times and the performance metrics recorded per time. Performance was measured in terms of sensitivity, specificity, accuracy and area under the operating curve (AUC). The panel that showed high consistency in performance with sensitivity and specificity values greater than 80%, was selected for reliability test.

E. Test of reliability

In the reliability test phase, the 10-fold cross-validation was iterated one thousand times (with the training and testing subsets internally randomized each time) and the average performance recorded. The reason was to thoroughly investigate the robustness of the selected panel in detecting AD across the entire dataset. In addition, the percentage of the times that the panel achieved sensitivity and specificity of not less than 80% (i.e., success rate) was calculated. The panel with high success rate was selected as the final panel. Earlier studies failed to account for success rate. However, it is of a critical importance so as to demonstrate the reliability of the panel's reported performance.

III. RESULTS

From the review of literature, 173 blood-based biomarkers of AD (excluding groups of microRNAs) were identified from 54 studies, from which 40 were cross-

validated. However, only 31 of these markers were available in ADNI database.

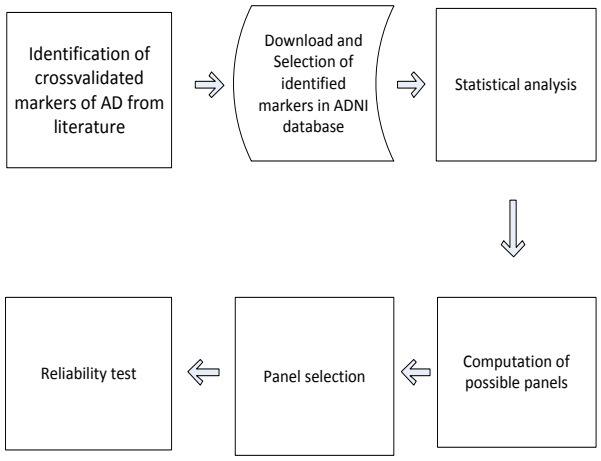

Figure 1. Description of methodology

Of the 31 proteins, only 14 that are listed in TABLE II showed significant difference between the AD and HC subjects. There were 3,458 possible candidate biomarker panels from the 14 shown in TABLE II. Each of the possible panels went through the learning procedure described in Section D. The machine learning algorithm identified a panel of five markers which includes Alpha-1 microglobulin (A1M), Alpha-2 macroglobulin (A2M), Complement C3 (C3), Immunoglobulin M (IgM), and Tenascin C (TNC). This panel detected AD with an average performance of 86.5%, 82.1%, 85% and 0.89 sensitivity, specificity, accuracy, and AUC, respectively. It achieved a success rate of 77.8% in the reliability test. This demonstrates that it is not sufficient to report the average or maximum performance of a panel in blood biomarker research but in addition, there is need to test and report its reliability across the entire dataset used in the study, as we have demonstrated in this paper.

TABLE I. SUMMARY OF THE BASELINE DEMOGRAPHICS OF AD AND HC SUBJECTS

	AD	**HC**
No. of subjects	108	58
Avg. Age (SD)	74.6(8)	75.1(5.8)
% Female	42.6	48.3
% APOE4+	68.5	8.6

AD indicates Alzheimer disease;
HC, healthy controls;
SD, standard deviation;
APOE4+, Apolipoprotein ε4 positive genotype

TABLE II. LIST OF EXAMINED MARKERS AND IDENTIFIED PANEL

Candidate markers	**Identified 5-marker panel**
Alpha-1 microglobulin	Alpha-1 microglobulin
Alpha-2 macroglobulin	Alpha-2 macroglobulin
Alpha-1 antitrypsin	Complement C3
Apolipoprotein E	Immunoglobulin M
Beta-2 microglobulin	Tenascin C
Brain natriuretic peptide	
Complement C3	
Eotaxin-3	
Immunoglobulin M	
Interleukin-3	
Macrophage inflammatory protein-1 alpha	
Pancreatic polypeptide	
Tenascin C	
Vascular cell adhesion molecule-1	

The method will aid reveal the true strength of the identified panel, improve the chances of replicating the reported performance and facilitate further refinements of the existing panels.

IV. DISCUSSION

In this study, we identified 5-biomarker panel (A1M, A2M, C3, IgM and TNC) for the diagnosis of AD in ADNI cohort, using APOE genotype as an additional feature, and thoroughly examined the robustness of the panel. The five biomarkers are well cross-validated candidate markers of AD across different cohorts. A1M is a protein involved in inflammatory response [19] that has also been identified as a plasma marker of brain atrophy in Alzheimer's disease [20]. The role of A2M in AD has been extensively researched; Bauer et al. [21] showed that A2M was present in amyloid plaques. Since then, it has further been linked to blood–brain barrier damage [22], hippocampal metabolism in early Alzheimer's disease [23] and neuronal injury [24]. Complement C3 has been identified as a marker of brain atrophy in AD [20] and cerebral amyloid in non-demented elderly [25]. IGM has been identified as blood protein marker of neocortical Amyloid-beta burden [26, 27].

TNC is an extracellular glycoprotein that has been linked to different biological processes, including inflammation and angiogenesis, which have association with AD [28]. Both IGM and TNC have been linked to apolipoprotein E genotype as well [29].

V. CONCLUSION

In summary, this piece of work further validates the utility of blood-based markers in diagnosing Alzheimer's disease and that reliable diagnosis with only a few number of biomarkers may be feasible and thus, widely accessible and routine diagnosis of Alzheimer's may be possible.

ACKNOWLEDGMENT

This research is funded by H2020 MSCA-TN-ETN-BBDiag Network. The authors are grateful to ADNI for granting them access to the data used in this study.

REFERENCES

[1] Alzheimer's Association, "2017 Alzheimer's disease facts and figures," *Alzheimer's & Dementia,* vol. 13, no. 4, pp. 325-373, 2017.

[2] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[3] B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. DeKosky, S. Gauthier, D. Selkoe, and R. Bateman, "Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria," *The Lancet Neurology,* vol. 13, no. 6, pp. 614-629, 2014.

[4] B. Dubois, H. H. Feldman, C. Jacova, S. T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, and G. Jicha, "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria," *The Lancet Neurology,* vol. 6, no. 8, pp. 734-746, 2007.

[5] J. A. AJ, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, and B. A. Spilker, "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clinical pharmacology and therapeutics,* vol. 69, no. 3, pp. 89-95, 2001.

[6] C. R. Jack, M. S. Albert, D. S. Knopman, G. M. McKhann, R. A. Sperling, M. C. Carrillo, B. Thies, and C. H. Phelps, "Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia: the journal of the Alzheimer's Association,* vol. 7, no. 3, pp. 257-262, 2011.

[7] P. Schneider, H. Hampel, and K. Buerger, "Biological marker candidates of Alzheimer's disease in blood, plasma, and serum," *CNS neuroscience & therapeutics,* vol. 15, no. 4, pp. 358-374, 2009.

[8] P. Davies, J. Resnick, B. Resnick, S. Gilman, J. H. Growdon, Z. S. Khachaturian, T. S. Radebaugh, A. D. Roses, D. J. Selkoe, and J. Q. Trojanowski, "Consensus report of the working group on:'Molecular and biochemical markers of Alzheimer's disease'," *Neurobiology of Aging,* vol. 19, no. 2, pp. 109-116, 1998.

[9] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. F. Friedman, D. R. Galasko, M. Jutel, and A. Karydas, "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins," *Nature medicine,* vol. 13, no. 11, pp. 1359, 2007.

[10] S. E. O'Bryant, G. Xiao, R. Barber, R. Huebinger, K. Wilhelmsen, M. Edwards, N. Graff-Radford, R. Doody, R. Diaz-Arrastia, T. A. s. Research, and C. Consortium, "A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI," *PloS one,* vol. 6, no. 12, pp. e28092, 2011.

[11] D. A. Llano, V. Devanarayan, A. J. Simon, and A. s. D. N. Initiative, "Evaluation of plasma proteomic data for Alzheimer disease state classification and for the prediction of progression from mild cognitive impairment to Alzheimer disease," *Alzheimer Disease & Associated Disorders,* vol. 27, no. 3, pp. 233-243, 2013.

[12] J. D. Doecke, S. M. Laws, N. G. Faux, W. Wilson, S. C. Burnham, C.-P. Lam, A. Mondal, J. Bedo, A. I. Bush, and B. Brown, "Blood-based protein biomarkers for diagnosis of Alzheimer disease," *Archives of neurology,* vol. 69, no. 10, pp. 1318-1325, 2012.

[13] L.H. Guo, P. Alexopoulos, S. Wagenpfeil, A. Kurz, R. Perneczky, and A. s. D. N. Initiative, "Plasma proteomics for the identification of Alzheimer's disease," *Alzheimer disease and associated disorders,* vol. 27, no. 4, 2013.

[14] E. Jammeh, P. Zhao, C. Carroll, S. Pearson, and E. Ifeachor, "Identification of blood biomarkers for use in point of care diagnosis tool for Alzheimer's disease" *in Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the,* 2016, pp. 2415-2418: IEEE.

[15] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," *Frontiers in Aging Neuroscience,* vol. 9, pp. 329, 2017.

[16] L. K. Ferreira, J. M. Rondina, R. Kubo, C. R. Ono, C. C. Leite, J. Smid, C. Bottino, R. Nitrini, G. F. Busatto, and F. L. Duran, "Support vector machine-based classification of neuroimages in Alzheimer's disease: direct comparison of FDG-PET, rCBF-SPECT and MRI data acquired from the same individuals," *Revista Brasileira de Psiquiatria*, no. AHEAD, pp. 0-0, 2017.

[17] J. Ramírez, J. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río, "Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features," *Information Sciences,* vol. 237, pp. 59-72, 2013.

[18] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, and T. J. Montine, "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia: the journal of the Alzheimer's Association,* vol. 7, no. 3, pp. 280-292, 2011.

[19] R. Zhang, L. Barker, D. Pinchev, J. Marshall, M. Rasamoelisolo, C. Smith, P. Kupchak, I. Kireeva, L. Ingratta, and G. Jackowski, "Mining biomarkers in human sera using proteomic tools," *Proteomics,* vol. 4, no. 1, pp. 244-256, 2004.

[20] M. Thambisetty, A. Simmons, A. Hye, J. Campbell, E. Westman, Y. Zhang, L.-O. Wahlund, A. Kinsey, M. Causevic, and R. Killick, "Plasma biomarkers of brain atrophy in Alzheimer's disease," *PloS one,* vol. 6, no. 12, pp. e28527, 2011.

[21] J. Bauer, S. Strauss, U. Schreiter-Gasser, U. Ganter, P. Schlegel, I. Witt, B. Yolk, and M. Berger, "Interleukin-6 and α-2-macroglobulin indicate an acute-phase state in Alzheimer's disease cortices," *FEBS letters,* vol. 285, no. 1, pp. 111-114, 1991.

[22] L. Cucullo, N. Marchi, M. Marroni, V. Fazio, S. Namura, and D. Janigro, "Blood-brain barrier damage induces release of α2-macroglobulin," *Molecular & Cellular Proteomics,* vol. 2, no. 4, pp. 234-241, 2003.

[23] M. Thambisetty, A. Hye, C. Foy, E. Daly, A. Glover, A. Cooper, A. Simmons, D. Murphy, and S. Lovestone, "Proteome-based identification of plasma proteins associated with hippocampal metabolism in early Alzheimer's disease," *Journal of neurology,* vol. 255, no. 11, pp. 1712-1720, 2008.

[24] V. R. Varma, S. Varma, Y. An, T. J. Hohman, S. Seddighi, R. Casanova, A. Beri, E. B. Dammer, N. T. Seyfried, and O. Pletnikova, "Alpha-2 macroglobulin in Alzheimer's disease: a marker of neuronal injury through the RCAN1 pathway," *Molecular psychiatry,* vol. 22, no. 1, pp. 13, 2017.

[25] X. Zhao, S. Lejnine, J. Spond, C. Zhang, T. Ramaraj, D. J. Holder, H. Dai, R. Weiner, and O. F. Laterza, "A candidate plasma protein classifier to identify Alzheimer's disease," *Journal of Alzheimer's Disease,* vol. 43, no. 2, pp. 549-563, 2015.

[26] S. J. Kiddle, M. Thambisetty, A. Simmons, J. Riddoch-Contreras, A. Hye, E. Westman, I. Pike, M. Ward, C. Johnston, and M. K. Lupton, "Plasma based markers of [11C] PiB-PET brain amyloid burden," *PloS one,* vol. 7, no. 9, pp. e44260, 2012.

[27] N. Voyle, D. Baker, S. C. Burnham, A. Covin, Z. Zhang, D. P. Sangurdekar, C. A. Tan Hehir, C. Bazenet, S. Lovestone, and S. Kiddle, "Blood protein markers of neocortical amyloid-β burden: a candidate study Using SOMAscan technology," *Journal of Alzheimer's Disease,* vol. 46, no. 4, pp. 947-961, 2015.

[28] K. S. Midwood, T. Hussenet, B. Langlois, and G. Orend, "Advances in tenascin-C biology," *Cellular and molecular life sciences,* vol. 68, no. 19, pp. 3175, 2011.

[29] H. D. Soares, W. Z. Potter, E. Pickering, M. Kuhn, F. W. Immermann, D. M. Shera, M. Ferm, R. A. Dean, A. J. Simon, and F. Swenson, "Plasma biomarkers associated with the apolipoprotein E genotype and Alzheimer disease," *Archives of neurology,* vol. 69, no. 10, pp. 1310-1317, 2012.