Methodological Issues in the Clinical Validation of Biomarkers for Alzheimer's Disease: The Paradigmatic Example of CSF

Marco Canevelli[1,2]*, Ilaria Bacigalupo[2], Giuseppe Gervasi[2], Eleonora Lacorte[2], Marco Massari[3], Flavia Mayer[2], Nicola Vanacore[2] and Matteo Cesari[4,5]

[1]Department of Human Neuroscience, Sapienza University, Rome, Italy, [2]National Center for Disease Prevention and Health Promotion, National Institute of Health, Rome, Italy, [3]National Center for Drug Research and Evaluation, National Institute of Health, Rome, Italy, [4]Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy, [5]Geriatric Unit, Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy

The use of biomarkers is profoundly transforming medical research and practice. Their adoption has triggered major advancements in the field of Alzheimer's disease (AD) over the past years. For instance, the analysis of the cerebrospinal fluid (CSF) and neuroimaging changes indicative of neuronal loss and amyloid deposition has led to the understanding that AD is characterized by a long preclinical phase. It is also supporting the transition towards a biology-grounded framework and definition of the disease. Nevertheless, though sufficient evidence exists about the *analytical validity* (i.e., accuracy, reliability, and reproducibility) of the candidate AD biomarkers, their *clinical validity* (i.e., how well the test measures the clinical features, and the disease or treatment outcomes) and *clinical utility* (i.e., if and how the test improves the patient's outcomes, confirms/changes the diagnosis, identifies at-risk individuals, influences therapeutic choices) have not been fully proven. In the present review, some of the methodological issues and challenges that should be addressed in order to better appreciate the potential benefits and limitations of AD biomarkers are discussed. The ultimate goal is to stimulate a constructive discussion aimed at filling the existing gaps and more precisely defining the directions of future research. Specifically, four main aspects of the clinical validation process are addressed and applied to the most relevant CSF biomarkers: (1) the definition of reference values; (2) the identification of reference standards for the disease of interest (i.e., AD); (3) the inclusion within the diagnostic process; and (4) the statistical process supporting the whole framework.

Keywords: biomarkers, Alzheimer's disease, validation, diagnostic research, epidemiology, mild cognitive impairment

INTRODUCTION

A biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention (Biomarkers Definitions Working Group, 2001). The use of biomarkers is profoundly transforming medical research and practice (the so called "biomarker revolution";

Schisterman and Albert, 2012). In fact, they may: (1) support the identification of pathophysiological processes causing or contributing to diseases; (2) define and predict the individual's health trajectories and clinical outcomes; and (3) help in selecting interventions and monitoring the response to treatments. Thus, they play a relevant role within the promise of precision medicine approaches where medical choices are driven by individually targeted genetic and biological profiles (Jameson and Longo, 2015).

Biomarkers are particularly relevant in the study of pathological conditions affecting the central nervous system (CNS), considering that brain tissue is not readily accessible for diagnostic or research purposes. Specifically, their adoption has triggered major advancements in the field of Alzheimer's disease (AD) over the past years. For instance, the analysis of the cerebrospinal fluid (CSF) and neuroimaging abnormalities indicative of neuronal loss and protein deposition has led to the understanding that AD is characterized by a long preclinical phase (Jack et al., 2013). This finding has been responsible for opening new perspectives in researching novel preventive/therapeutic strategies. It has also supported the transition towards a biology-grounded framework and definition of the disease (Jack et al., 2018). Furthermore, the use of these markers, when adopted as surrogate measures of AD in animal models, has contributed in accelerating the development of possible disease-modifying treatments (Cummings et al., 2018).

To date, although increasingly adopted in specialized clinical settings (Frisoni et al., 2017), the use of biomarkers to detect AD is still recommended only for research purposes and in selected atypical cases (McKhann et al., 2011; Dubois et al., 2014; Jack et al., 2018). Their adoption in the routine clinical practice remains controversial as confirmed by different systematic reviews and meta-analyses reaching heterogeneous results on the topic (Noel-Storr et al., 2013; Olsson et al., 2016; Ritchie et al., 2017). In particular, though sufficient (albeit inconclusive) evidence exists about the *analytical validity* (i.e., is the test accurate, reliable, and reproducible?) of the proposed AD biomarkers (Hansson et al., 2018; Lewczuk et al., 2018), their *clinical validity* (i.e., how well the test measures the clinical features, and the disease or treatment outcomes) and *clinical utility* (i.e., if and how the test improves the patient's outcomes, confirms/defines the diagnosis, identifies at-risk individuals, influences therapeutic choices) have not yet been fully proven (Frisoni et al., 2017; Kraus, 2018).

In the present article, we discuss some of the methodological issues and challenges that should be addressed in order to better assess the potential benefits and limitations of AD biomarkers. Without the intent of underestimating what has been done over the years in the field, the ultimate goal of the present article is to stimulate a constructive discussion aimed at filling the existing gaps and more precisely defining the directions of future research. The work is structured around four main aspects to be considered when adopting a biomarker in clinical practice: (1) the definition of reference values; (2) the identification of reference standards specific for the disease of interest (i.e., AD); (3) the proper inclusion and contextualization within the diagnostic process; and (4) the statistical process supporting the whole framework. In particular, these points will be addressed with regard to the most relevant CSF biomarkers.

DEFINITION OF REFERENCE VALUES

The validation of a candidate biomarker should follow two preliminary steps: (1) the assessment of its distribution in healthy people; and (2) the definition of the index test reference values (e.g., those included between the 2.5th and 97.5th percentile of the distribution, or within the interval of the mean ±1.96 standard deviations in case of symmetric distribution). The impact of common sociodemographic characteristics (e.g., age, sex, race/ethnicity) on the identified normal and abnormal values should also be considered (Sackett and Haynes, 2002; Haynes and You, 2009; Colli et al., 2014). It should be underlined, within this framework, how challenging (or even arbitrary) the selection of the reference group (i.e., healthy controls) to be used to define the 95% range of reference values might be.

To assess the methodology of the studies providing reference intervals for possible CSF biomarkers [i.e., amyloid peptides Aβ1–42 (Aβ42), total tau (T-tau), and 181-phospo-tau (P-tau)] in AD, we retrieved all available literature published up to May 2019. To this purpose, we performed a structured search on PubMed using the following search terms: (Aβ* OR A-β* OR A-beta* OR abeta* OR AB-42 OR *tau) AND (CSF OR liquor OR cerebrospinal OR cerebro-spinal) AND [(population* OR reference* OR normative*) AND (value* OR limit*)] AND (healthy OR normal OR normality OR average OR "general population"). The search strategy led to the identification of 155 abstracts. The full-texts of six selected studies were retrieved and assessed for inclusion based on the following predefined inclusion/exclusion criteria: being published in English; having sample size >50 subjects; defining as explicit aim the identification of reference intervals or limits for the considered biomarkers. Only two studies were included based on their pertinence with and relevance to the topic of interest (Sjögren et al., 2001; Burkhard et al., 2004). As reported in **Table 1**, the two included studies showed a high heterogeneity in how both methods and results were reported, thus limiting their hypothetical summarization. Both studies investigated the CSF dosage of Aβ42 and T-tau in hospital-based samples of subjects with a wide spectrum of age (i.e., ranging from less than 30 years to even more than 90 years). The studies adopted the 10th fractile (or percentile) to calculate the reference limit for Aβ42 and the 90th fractile (or percentile) to define the reference limit for T-tau. Important differences were observed for what concerns the age distribution and sex composition of the enrolled study samples. Although the inconsistencies in the reporting of results (e.g., different stratification for age groups) preclude the possibility of a direct comparison of the findings, a relevant discrepancy in the identified reference limits was evident in the two studies (e.g., for Aβ42: 150 ng/L vs. 500 ng/L, respectively). Finally, none of them assessed the role of individual characteristics (e.g., race and genetics) that could potentially affect results and conclusions.

TABLE 1 | Studies reporting reference limits for CSF Aβ42 and T-tau in healthy people.

Author	Setting	Sample size	Age	Sex	Diagnostic criteria for healthy individuals	Statistics	Reference limits
Sjögren et al. (2001)	Norwegian University Norwegian Hospital	231 (total) 175 (age >50 years)	61.3 ± 18 (mean age ± SD) 21–93 (range)	177 F 54 M	No symptoms or signs of psychiatric or neurological disorders: • People aged ≥60 years: MMSE score = 28–30 • People aged <60 years: no criteria listed	• >0.10 fractile (i.e., 10th percentile) for Aβ42 • <0.9 fractile (i.e., 90th percentile) for tau	**Aβ42** 500 ng/L (not age dependent) **T-tau** • 300 ng/L (for age range 21–50) • 450 ng/L (for age range 51–70) • 500 ng/L (for age range ≥71)
Burkhard et al. (2004)	University Hospital of Geneve	105 (total) 82 (age >50 years)	69, 56–78 (median, interquartile range) 29–96 (range)	61 F 44 M	• Historical referral of absence of any CNS or PNS condition, or psychiatric condition, or chronic systemic illness possibly modifying the CSF proteins • Face-to-face questionnaire covering the most common neurological symptoms • Intake of medication potentially interfering with brain functions • Neuropathological findings confirming the absence of any dementia (for 10 autopsy proven cases)	• >10th percentile for Aβ42 • <90th percentile for tau	**Aβ42** 150 ng/L, 90% Confidence Interval 90–172 ng/L (for age >49 years) **T-tau** 385 ng/L, 90% Confidence Interval 332–451 ng/L (for age >49 years)

DEFINING DIAGNOSTIC REFERENCE STANDARDS FOR AD

The clinical validation of AD biomarkers is complicated by the lack of a unique diagnostic reference (Noel-Storr et al., 2013). Furthermore, the biological and clinical approaches to the diagnosis of AD have some relevant limitations. Neuropathology has traditionally been considered as the gold standard for the evaluation and judgment of clinical manifestations (McKhann et al., 1984). Nevertheless, its large-scale implementation is hampered by the difficulty of obtaining samples. However, the neuropathological characteristics of AD have a weak correlation with its phenotypic and clinical expression. In fact, it is well established that many individuals showing a high burden of AD pathology do not exhibit any clinical signs of the disease, whereas others with a limited amount of neuropathological changes had developed overt AD in life (Wallace et al., 2019). Beyond the absence of clear evidence supporting their causal role, some of the biological processes resulting in the AD neuropathological hallmarks (e.g., amyloid deposition) may have different pathogenic implications (Espay et al., 2019). They may, in fact, alternatively contribute to and accelerate neurodegeneration, represent epiphenomena, or even constitute compensatory mechanisms to molecular/cellular stress (Espay et al., 2019). Moreover, different latent factors, such as the individual's frailty status, may moderate the relationship between AD pathology and dementia (Wallace et al., 2019). Finally, most

of dementia cases (including AD dementia) are underlined by a mixed neuropathology (Boyle et al., 2018).

On the other hand, the adoption of clinical standards can be itself prevented by several obstacles. Logically, the cross-sectional validation of biomarkers against clinical criteria cannot result in an optimal diagnostic accuracy (Noel-Storr et al., 2013). Therefore, their use as prognostic markers, using longitudinal reference standards such as the conversion from MCI to AD dementia, are being increasingly considered for this purpose (Ritchie et al., 2014). However, the marked heterogeneity of these clinical outcomes may strongly confound their performance. For instance, the phenomenon of MCI conversion may occur in extremely variable times and ways, and be potentially affected by several additional, interacting factors (Grande et al., 2014). Moreover, it has been observed that a sizeable proportion of subjects with MCI shows a normalization of neuropsychological tests over time (Canevelli et al., 2016). Some subjects may follow even more complex clinical trajectories, by, for example, first reverting to normal cognition and subsequently progressing to dementia (Roberts et al., 2014). Theoretically, such a potential for multiple evolutions of MCI, shared by most of the risk conditions (Canevelli et al., 2017), implies the need to overcome the adoption of "classic" dichotomous outcomes (i.e., normal vs. pathological) preferring endpoints including at least 3 levels (i.e., improvement vs. stability vs. worsening). In other words, biomarkers could potentially support the identification not only of those subjects progressing to dementia, but also of those

TABLE 2 | The diagnostic research questions.

Phase I: Do the test results in patients with the target disorder differ from those in normal people?
This preliminary phase is important to provide novel insights on the pathophysiological mechanisms of the disease. It can be addressed by conducting cross-sectional studies confronting a convenience group of subjects known to have the disease and a group of people definitely known to not have it.

Phase II: Are patients with certain test results more likely to have the target disorder than patients with other test results?
The answer to this question can be derived by classic 2 × 2 contingency tables (or Error Matrices). The accuracy of the test (in terms of its results or cut-points) at distinguishing patients with the disease from normal controls is expressed by means of sensitivity, specificity, positive and negative predictive values and likelihood ratios

Phase III: Does the test result distinguish patients with and without the target disorder among patients in whom it is clinically reasonable to suspect that the disease is present?
Differently from the previous phase, the accuracy of the test is here explored in a "real world" scenario of routine clinical practice, that is among subjects whose clinical status is not already established (e.g., subjects referred from their general practitioners to specialist services for a clinical suspicion). Participants should, blindly, be assessed with both the test and what is considered as the diagnostic reference standard (ideally a gold standard).

Phase IV: Do patients who undergo this diagnostic test have better health outcomes than similar patients who are not tested?
This question strongly deals with the clinical utility of the test and concerns the health outcomes following the diagnostic/therapeutic choices resulting from the test findings. Ideally, such information could be obtained by the follow-up of subjects randomized to perform the test or not to perform it.

Phase V: Does the use of the diagnostic test lead to better health outcomes at acceptable costs?
This question refers to the cost-effectiveness (the so-called "value-for-money") of the index test and can be answered by randomized controlled trials.

Adapted from Haynes and You (2009) and Sackett and Haynes (2002).

individuals showing an "inverse" trajectory towards normality. In this framework, the possibility of combining different biomarkers (or sets of biomarkers) should be considered with the objective of detecting the risk of decline as well as the possibility of restoration of a normal status.

THE ARCHITECTURE OF THE DIAGNOSTIC PROCESS

The actual validity and utility of a diagnostic test (e.g., a biomarker) can be summarized in a multistep process that should answer some crucial diagnostic questions, included in five iterative phases (**Table 2**; Sackett and Haynes, 2002; Haynes and You, 2009).

In the AD literature, a relevant number of Phase I and Phase II studies has indicated that the CSF levels of biomarkers reflecting amyloid deposition (i.e., Aβ42) and neurodegeneration (i.e., T-tau and P-tau) are significantly different between subjects diagnosed with AD and to normal controls. In this context, a recent meta-analysis of 231 studies enrolling a total of 15,699 patients with AD and 13,018 controls reported an estimate of the following AD-to-control ratios: Aβ42 (average ratio 0.56, 95% CI 0.55–0.58, $p < 0.0001$), T-tau (2.54, 2.44–2.64, $p < 0.0001$), and P-tau (1.88, 1.79–1.97, $p < 0.0001$; Olsson et al., 2016). These biomarkers could also help in distinguishing those subjects with mild cognitive impairment (MCI) that will convert to dementia from non-converting subjects (Ritchie et al., 2017). Specifically, according to a recent Cochrane systematic review (Ritchie et al., 2017), the observed accuracy ranges of CSF biomarkers in predicting the conversion from MCI to AD dementia are:

- T-tau: sensitivity: 51%–90%; specificity: 48%–88%;
- P-tau: sensitivity: 40%–100%; specificity: 22%–86%;
- P-tau/Aβ42 ratio: sensitivity: 80%–96%; specificity: 33%–95%.

Such wide variability can be attributed to relevant discrepancies in the adopted reference standards, in the source of recruitment and sampling of participants, and in the index test methodology across the retained studies. It is to be noted that most of these results were obtained in research settings, evaluating highly selected patients in whom the presence of the target disease had already been ascertained under ideal/almost utopic circumstances (e.g., by expert clinicians with the best available equipment, adopting the same reference standard for those with and without AD). These samples are unlikely to represent the overall population of patients with AD under multiple socio-demographic and clinical aspects. Therefore, it seems reasonable to expect these same biomarkers to yield different results when transferred from the research to the clinical setting (Dyer et al., 2016; Frisoni et al., 2017). To date, only few studies have provided realistic information on the validity of AD biomarkers in the "real world" (thus answering pragmatic Phase III questions). As expected, a lower accuracy in the discrimination of patients with and without AD was observed in these works (Mattsson et al., 2009; Tariciotti et al., 2018). Moreover, to our knowledge, no Phase IV and V evidence are available in this field of AD research. In other words, no study has yet robustly explored how the use of biomarkers can actually affect health outcomes (e.g., mortality, disability, response to treatment; Frisoni et al., 2017) nor their cost-effectiveness.

STATISTICAL APPROACHES ACROSS THE DIAGNOSTIC RESEARCH PROCESS

According to the previously discussed phases, different statistical approaches are required in each sequential step (Moons et al., 2012a,b; Collins et al., 2015). Phase I is exploratory by nature and is typically based on null hypothesis significance testing focused on isolating variables deemed individually relevant according to the *P*-value. The statistical methods for investigating Phase II and III questions belong to the field of prediction models (both diagnostic and prognostic) that typically focus on identifying sets of variables that can accurately predict the outcomes of interest. Considering the wide range of options and the differing

TABLE 3 | Definition and interpretation of the Likelihood Ratio (LR).

LR values	Change from pre-test to post-test probability	Results
>10 or <0.1	Large	Conclusive
5–10 or 0.1–0.2	Moderate	Moderately important
2–5 or 0.2–0.5	Small	Sometimes important
1–2 or 0.5–1	Small/minimum	Rarely important
1	None	Not important/useless

Adapted from Jaeschke et al. (1994).

perspectives of researchers, clinicians and public health decision makers, it is crucial to be aware about the trade-off between *model transparency* (allowing for easy interpretability and transparent scientific understanding) and *model complexity* (maximizing the predictive power through very sophisticated predictions that may often appear as an Opaque Black Box; Bzdok and Ioannidis, 2019). To this purpose, simple univariable classifications where Error Matrices (i.e., 2 × 2 contingency tables that report the number of *false positives*, *false negatives*, *true positives*, and *true negatives*) are derived by predefined cut-off values of single biomarkers as well as long-trusted multivariable statistical methods (e.g., Logistic and Cox Regression models) still remain the most suitable tools in the box. Regarding the Error Matrix and its derived measures (Akobeng, 2007), the Positive Predictive Value (PPV) and the Likelihood Ratio (LR) should always be preferred in prediction studies. In fact, Sensitivity and Specificity are indicative of the accuracy of a test (i.e., the biomarker), thus they are mostly useful for comparing the performance of different ones (with the possibility of combining two single tests in "OR"/"AND" modality to enhance the overall sensitivity/specificity; Sackett et al., 1985). The PPV and LR are, instead, informative about the single, specific individual. The PPV measures the individual probability to develop (or to have) the disease if the test is positive. The LR expresses the probability that the test is positive (or negative) in people with the disease compared to the probability that it is positive (or negative) in healthy people. It thus allows to simply update the pre-test probability of having the disease (based on the individual's characteristics and clinical history) to the post-test probability (given the test results) according to its direction and magnitude (**Table 3**; Jaeschke et al., 1994; Kent and Hancock, 2016). Candidate CSF biomarkers for AD have so far shown small to minimum LR values (i.e., LR+ 2.72, LR− 0.32 at the median specificity of 72% for T-tau; LR+ 1.55, LR− 0.39 at the median specificity of 47.5% for P-tau; Ritchie et al., 2017).

The predictive performance of a model is usually measured using discrimination measures (such as c-index that is equal to the area under the Receiver Operating Curve) and calibration plots. These measures can be inflated in the data sample from which they are derived when compared to new but comparable data samples (overfitting). K-fold cross-validation and bootstrap are the preferred internal validation techniques to evaluate a potential overfitting. However, external validation is still necessary to guarantee the generalizability of the model in the real word setting (Phase III). Finally, the appropriate reporting, communication and use of the resulting model are crucial. Therefore, the output of the predictive model (in terms of

coefficient estimates, standard error and confidence intervals) can be combined to graphic tools, such as nomograms, thus easily allowing to obtain the final outcome probability for a new patient based on his/her profile of predictive variables. This graphical approach, although not widely used in the field of AD (Jang et al., 2017), may have important practical implications in the clinical and regulatory setting (e.g., patient's counseling, risk stratification, elaboration of guidelines, drug reimbursement). Phase IV studies, while sharing inferential testing tools that are similar to those used in Phase I, are usually framed within an evidence-based decision-making context where the statistical methods are derived from the domain of well-controlled experimental study design (typically a randomized clinical trial). Phase V studies, instead, focus on the evaluation of the most effective or cost-effective diagnostic strategies through specific cost-effectiveness analysis.

CONCLUSION

Overall, various methodological issues remain to be addressed in order to perform an adequate and complete clinical validation of candidate CSF biomarkers for AD. First, studies reporting the distribution of biomarkers in normal/healthy subjects and their variability according to major sociodemographic and clinical attributes are still lacking. In this regard, significant sex and race disparities for Aβ42 and tau levels have recently been reported both in healthy subjects and in patients with AD (Koran et al., 2017; Morris et al., 2019). Second, there is no conclusive agreement on the most appropriate reference standard for AD (e.g., clinical vs. biological) to be adopted to test the performance of new biomarkers. Third, no biomarker has yet consistently gone through all the phases that compose the architecture of diagnostic research. In particular, their actual impact on "hard" health outcomes and their cost-effectiveness has to be clarified. Similar conclusions have been reached by Mattsson et al. (2017) who have adopted an alternative model for developing the framework concerning AD biomarkers. Their approach, borrowed from oncology and structured around the natural history of the disease, should be regarded as complimentary to that adopted in the present work, essentially based on the methodological validation of biomarkers from the lens of clinical epidemiology. It is also crucial that, in each phase, the scientific contributions meet the highest quality standards. To this end, the widespread application of the checklist on reporting standards in dementia and cognitive impairment (STARDdem; Noel-Storr et al., 2014) can be a useful tool to improve consistency and transparency, and the application of the QUADAS 2 checklist (Whiting et al., 2011) can allow the identification of potential methodological biases, thus enabling a more effective assessment of candidate diagnostic tests. Moreover, multivariate statistical methodologies, possibly resulting in clinically-oriented tools such as nomograms, should be increasingly used to capture the complexity of the disease, both from a pathophysiological and phenotypic perspective, and to understand the actual clinical relevance of potential new biomarkers. It should be emphasized how these considerations, here paradigmatically referred to CSF, can be extended to all the candidate biomarkers for

AD, regardless of their origin and nature (e.g., plasma, serum, urine, neuroimaging).

In conclusion, despite the enormous progress made in the field, there is still insufficient evidence to promote the use of candidate CSF biomarkers for AD in the routine clinical practice, As already pointed out by previous works on this topic, leaving the discussed methodological issues unaddressed raises the risk to provide clinicians with tools and tests whose answers are difficult to interpret and translate into concrete decisions. This might ultimately result in potential harm to patients, families, and healthcare systems.

AUTHOR CONTRIBUTIONS

MCa: study conception and writing of the manuscript. IB, GG, EL, MM and FM: literature search and drafting of the manuscript. NV and MCe: study conception and revision of the manuscript for important intellectual content.

REFERENCES

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr.* 96, 338–341. doi: 10.1111/j.1651-2227.2006.00180.x

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* 69, 89–95. doi: 10.1067/mcp.2001.113989

Boyle, P. A., Yu, L., Wilson, R. S., Leurgans, S. E., Schneider, J. A., and Bennett, D. A. (2018). Person-specific contribution of neuropathologies to cognitive loss in old age. *Ann. Neurol.* 83, 74–83. doi: 10.1002/ana.25123

Burkhard, P. R., Fournier, R., Mermillod, B., Krause, K.-H., Bouras, C., and Irminger, I. (2004). Cerebrospinal fluid tau and Aβ42 concentrations in healthy subjects: delineation of reference intervals and their limitations. *Clin. Chem. Lab. Med.* 42, 396–407. doi: 10.1515/cclm.2004.071

Bzdok, D., and Ioannidis, J. P. A. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci.* 42, 251–262. doi: 10.1016/j.tins.2019.02.001

Canevelli, M., Bruno, G., Remiddi, F., Vico, C., Lacorte, E., Vanacore, N., et al. (2017). Spontaneous reversion of clinical conditions measuring the risk profile of the individual: from frailty to mild cognitive impairment. *Front. Med. Lausanne.* 4:184. doi: 10.3389/fmed.2017.00184

Canevelli, M., Grande, G., Lacorte, E., Quarchioni, E., Cesari, M., Mariani, C., et al. (2016). Spontaneous reversion of mild cognitive impairment to normal cognition: a systematic review of literature and meta-analysis. *J. Am. Med. Dir. Assoc.* 17, 943–948. doi: 10.1016/j.jamda.2016.06.020

Colli, A., Fraquelli, M., Casazza, G., Conte, D., Nikolova, D., Duca, P., et al. (2014). The architecture of diagnostic research: from bench to bedside—research guidelines using liver stiffness as an example. *Hepatology* 60, 408–418. doi: 10.1002/hep.26948

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* 162, 55–63. doi: 10.7326/l15-0078-4

Cummings, J., Lee, G., Ritter, A., and Zhong, K. (2018). Alzheimer's disease drug development pipeline: 2018. *Alzheimers Dement.* 4, 195–214. doi: 10.1016/j.trci.2018.03.009

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., et al. (2014). Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurol.* 13, 614–629. doi: 10.1016/S1474-4422(14)70090-0

Dyer, S. M., Flicker, L., Laver, K., Whitehead, C., and Cumming, R. (2016). The clinical value of fluid biomarkers for dementia diagnosis. *Lancet Neurol.* 15:1204. doi: 10.1016/S1474-4422(16)30238-1

Espay, A. J., Vizcarra, J. A., Marsili, L., Lang, A. E., Simon, D. K., Merola, A., et al. (2019). Revisiting protein aggregation as pathogenic in sporadic Parkinson and Alzheimer diseases. *Neurology* 92, 329–337. doi: 10.1212/WNL.0000000000006926

Frisoni, G. B., Boccardi, M., Barkhof, F., Blennow, K., Cappa, S., Chiotis, K., et al. (2017). Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *Lancet Neurol.* 16, 661–676. doi: 10.1016/S1474-4422(17)30159-X

Grande, G., Vanacore, N., Maggiore, L., Cucumo, V., Ghiretti, R., Galimberti, D., et al. (2014). Physical activity reduces the risk of dementia in mild cognitive impairment subjects: a cohort study. *J. Alzheimers Dis.* 39, 833–839. doi: 10.3233/jad-131808

Hansson, O., Mikulskis, A., Fagan, A. M., Teunissen, C., Zetterberg, H., Vanderstichele, H., et al. (2018). The impact of preanalytical variables on measuring cerebrospinal fluid biomarkers for Alzheimer's disease diagnosis: a review. *Alzheimers Dement.* 14, 1313–1333. doi: 10.1016/j.jalz.2018.05.008

Haynes, R. B., and You, J. J. (2009). "The architecture of diagnostic research," in *The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research, 2nd Edn*, eds J. A. Knottnerus and F. Buntinx (Blackwell Publishing Ltd), 20–41.

Jack, C. R. Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* 14, 535–562. doi: 10.1016/j.jalz.2018.02.018

Jack, C. R. Jr., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* 12, 207–216. doi: 10.1016/S1474-4422(12)70291-0

Jaeschke, R., Guyatt, G. H., and Sackett, D. L. (1994). Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 271, 703–707. doi: 10.1001/jama.271.9.703

Jameson, J. L., and Longo, D. L. (2015). Precision medicine—personalized, problematic, and promising. *N. Engl. J. Med.* 372, 2229–2234. doi: 10.1056/NEJMsb1503104

Jang, H., Ye, B. S., Woo, S., Kim, S. W., Chin, J., Choi, S. H., et al. (2017). Prediction model of conversion to dementia risk in subjects with amnestic mild cognitive impairment: a longitudinal, multi-center clinic-based study. *J. Alzheimers Dis.* 60, 1579–1587. doi: 10.3233/JAD-170507

Kent, P., and Hancock, M. J. (2016). Interpretation of dichotomous outcomes: sensitivity, specificity, likelihood ratios, and pre-test and post-test probability. *J. Physiother.* 62, 231–233. doi: 10.1016/j.jphys.2016.08.008

Koran, M. E. I., Wagener, M., and Hohman, T. J. (2017). Sex differences in the association between AD biomarkers and cognitive decline. *Brain Imaging Behav.* 11, 205–213. doi: 10.1007/s11682-016-9523-8

Kraus, V. B. (2018). Biomarkers as drug development tools: discovery, validation, qualification and use. *Nat. Rev. Rheumatol.* 14, 354–362. doi: 10.1038/s41584-018-0005-9

Lewczuk, P., Gaignaux, A., Kofanova, O., Ermann, N., Betsou, F., Brandner, S., et al. (2018). Interlaboratory proficiency processing scheme in CSF aliquoting: implementation and assessment based on biomarkers of Alzheimer's disease. *Alzheimers Res. Ther.* 10:87. doi: 10.1186/s13195-018-0418-3

Mattsson, N., Lönneborg, A., Boccardi, M., Blennow, K., Hansson, O., and Geneva Task Force for the Roadmap of Alzheimer's Biomarkers. (2017). Clinical validity of cerebrospinal fluid Aβ42, tau, and phospho-tau as biomarkers for Alzheimer's disease in the context of a structured 5-phase development framework. *Neurobiol. Aging* 52, 196–213. doi: 10.1016/j.neurobiolaging.2016.02.034

Mattsson, N., Zetterberg, H., Hansson, O., Andreasen, N., Parnetti, L., Jonsson, M., et al. (2009). CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA* 302, 385–393. doi: 10.1001/jama.2009.1064

McKhann, G. M., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of department of health

and human services task force on Alzheimer's disease. *Neurology* 34, 939–944. doi: 10.1212/wnl.34.7.939

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 263–269. doi: 10.1016/j.jalz.2011.03.005

Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., et al. (2012a). Risk prediction models: II. External validation, model, updating and impact assessment. *Heart* 98, 691–698. doi: 10.1136/heartjnl-2011-301247

Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., et al. (2012b). Risk prediction models: I. Development, internal validation and assessing the incremental value of a new (bio)marker. *Heart* 98, 683–690. doi: 10.1136/heartjnl-2011-301246

Morris, J. C., Schindler, S. E., McCue, L. M., Moulder, K. L., Benzinger, T. L. S., Cruchaga, C., et al. (2019). Assessment of racial disparities in biomarkers for Alzheimer disease. *JAMA Neurol.* 76, 264–273. doi: 10.1001/jamaneurol. 2018.4249

Noel-Storr, A. H., Flicker, L., Ritchie, C. W., Nguyen, G. H., Gupta, T., Wood, P., et al. (2013). Systematic review of the body of evidence for the use of biomarkers in the diagnosis of dementia. *Alzheimers Dement.* 9, e96–e105. doi: 10.1016/j. jalz.2012.01.014

Noel-Storr, A. H., McCleery, J. M., Richard, E., Ritchie, C. W., Flicker, L., Cullum, S. J., et al. (2014). Reporting standards for studies of diagnostic test accuracy in dementia. *Neurology* 83, 364–373. doi: 10.1212/WNL. 0000000000000621

Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., et al. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *Lancet Neurol.* 15, 673–684. doi: 10.1016/S1474-4422(16)00070-3

Ritchie, C., Smailagic, N., Noel-Storr, A. H., Takwoingi, Y., Flicker, L., Mason, S. E., et al. (2014). Plasma and cerebrospinal fluid amyloid-β for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst. Rev.* 6:CD008782. doi: 10.1002/14651858.CD008782.pub4

Ritchie, C., Smailagic, N., Noel-Storr, A. H., Ukoumunne, O., Ladds, E. C., and Martin, S. (2017). CSF tau and the CSF tau/Aβ ratio for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst. Rev.* 3:CD010803. doi: 10.1002/14651858.CD010803.pub2

Roberts, R. O., Knopman, D. S., Mielke, M. M., Cha, R. H., Pankratz, V. S., Christianson, T. J. H., et al. (2014). Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal. *Neurology* 82, 317–325. doi: 10.1212/wnl.0000000000000055

Sackett, D. L., and Haynes, R. B. (2002). The architecture of diagnostic research. *BMJ* 324, 539–541. doi: 10.1136/bmj.324.7336.539

Sackett, D. L., Haynes, R. B., and Tugwell, P. (1985). *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Boston: Little, Brown & Company.

Schisterman, E. F., and Albert, P. S. (2012). The biomarker revolution. *Stat. Med.* 31, 2513–2515. doi: 10.1002/sim.5499

Sjögren, M., Vanderstichele, H., Agren, H., Zachrisson, O., Edsbagge, M., Wikkelsø, C., et al. (2001). Tau and Aβ42 in cerebrospinal fluid from healthy adults 21–93 years of age: establishment of reference values. *Clin. Chem.* 47, 1776–1781.

Tariciotti, L., Casadei, M., Honig, L. S., Teich, A. F., McKhann Ii, G. M., Tosto, G., et al. (2018). Clinical experience with cerebrospinal fluid Aβ42, total and phosphorylated tau in the evaluation of 1,016 individuals for suspected dementia. *J. Alzheimers Dis.* 65, 1417–1425. doi: 10.3233/JAD-180548

Wallace, L. M. K., Theou, O., Godin, J., Andrew, M. K., Bennett, D. A., and Rockwood, K. (2019). Investigation of frailty as a moderator of the relationship between neuropathology and dementia in Alzheimer's disease: a cross-sectional analysis of data from the Rush Memory and Aging Project. *Lancet Neurol.* 18, 177–184. doi: 10.1016/s1474-4422(18) 30371-5

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., et al. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* 155, 529–536. doi: 10.7326/0003-4819-155-8-201110180-00009