Correcting for the Absence of a Gold Standard Improves Diagnostic Accuracy of Biomarkers in Alzheimer's Disease

Peer-reviewed author version

Coart, Els; GARCIA BARRADO, Leandro; Duits, Flora H.; Scheltens, Philip; van der Flier, Wiesje M.; Teunissen, Charlotte E.; van der Vies, Saskia M. & BURZYKOWSKI, Tomasz (2015) Correcting for the Absence of a Gold Standard Improves Diagnostic Accuracy of Biomarkers in Alzheimer's Disease. In: JOURNAL OF ALZHEIMERS DISEASE, 46 (4), p. 889-899.

DOI: 10.3233/JAD-142886
Handle: http://hdl.handle.net/1942/19148

Journal of Alzheimer's Disease xx (20xx) x–xx
DOI 10.3233/JAD-142886
IOS Press

Correcting for the Absence of a Gold Standard Improves Diagnostic Accuracy of Biomarkers in Alzheimer's Disease

Els Coart[a,*], Leandro García Barrado[b], Flora H. Duits[c], Philip Scheltens[c], Wiesje M. van der Flier[c,d], Charlotte E. Teunissen[e], Saskia M. van der Vies[f], Tomasz Burzykowski[a,b] and for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]*International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium*
[b]*Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium*
[c]*Alzheimer Center & Department of Neurology, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands*
[d]*Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands*
[e]*Neurochemistry Laboratory and Biobank, Department of Clinical Chemistry, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands*
[f]*Department of Pathology, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands*

Handling Associate Editor: Henrik Zetterberg

Accepted 26 March 2015

Abstract.
Background: Studies investigating the diagnostic accuracy of biomarkers for Alzheimer's disease (AD) are typically performed using the clinical diagnosis or amyloid-β positron emission tomography as the reference test. However, neither can be considered a gold standard or a perfect reference test for AD. Not accounting for errors in the reference test is known to cause bias in the diagnostic accuracy of biomarkers.
Objective: To determine the diagnostic accuracy of AD biomarkers while taking the imperfectness of the reference test into account.
Methods: To determine the diagnostic accuracy of AD biomarkers and taking the imperfectness of the reference test into account, we have developed a Bayesian method. This method establishes the biomarkers' true value in predicting the AD-pathology status by combining the reference test and the biomarker data with available information on the reliability of the reference test. The new methodology was applied to two clinical datasets to establish the joint accuracy of three cerebrospinal fluid biomarkers (amyloid-$\beta_{1\text{-}42}$, Total tau, and P-tau_{181}) by including the clinical diagnosis as imperfect reference test into the analysis.

[1]Part of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://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: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

*Correspondence to: Els Coart, International Drug Development Institute (IDDI), Avenue Provinciale 30, 1340 Louvain-la-Neuve, Belgium. Tel.: +32 10 61 44 44; Fax: +32 10 61 88 88; E-mail: Elisabeth.Coart@iddi.com.

Results: The area under the receiver-operating-characteristics curve to discriminate between AD and controls, increases from 0.949 (with 95% credible interval [0.935,0.960]) to 0.990 ([0.985,0.995]) and from 0.870 ([0.817,0.912]) to 0.975 ([0.943,0.990]) for the cohorts, respectively.
Conclusions: Use of the Bayesian methodology enables an improved estimate of the exact diagnostic value of AD biomarkers and overcomes the lack of a gold standard for AD. Using the new method will increase the diagnostic confidence for early stages of AD.

Keywords: Alzheimer's disease, Bayesian method, biomarkers, diagnostic test, reference standard

INTRODUCTION

Biomarkers for Alzheimer's disease (AD) that are linked to the pathological process are of paramount importance for early diagnosis of AD and selection of appropriate patients for clinical trials [1, 2]. Before biomarkers can be used clinically, their diagnostic accuracy needs to be thoroughly ascertained. To this end, a reference test against which the biomarker is verified needs to be selected.

The first choice would be the definite AD diagnosis, provided by postmortem neuropathological analysis. However, autopsy confirmation suffers from considerable between-laboratory differences [3], is by definition *post hoc,* and is only rarely available. In general, the accuracy of early AD biomarkers, or any diagnostic test for AD for that matter, is typically assessed using the clinical diagnosis as the reference test. The latter is imperfect because the clinical diagnosis suffers from classification errors (misdiagnosis) [4] and the onset of the pathogenic process as reflected in biomarker changes can precede the manifestation of clinical symptoms by at least a decade [5]. Hence, a clinical non-AD diagnosis does not exclude underlying AD-pathology and the clinical diagnosis of AD does not predict underlying pathology, as was recently shown in the phase III study with Bapineuzimab [6].

The imperfectness of the clinical diagnosis is usually ignored, resulting in a biased assessment of the diagnostic accuracy of biomarkers and suboptimal biomarker thresholds for clinical applications [7]. If the biomarker and reference test do not tend to misclassify the same patients, the diagnostic accuracy of the biomarker will be underestimated. When the biomarker and the reference test are dependent, the diagnostic accuracy of the biomarker can be either underestimated or overestimated, depending on the strength of the association [8, 9]. Recently, Toledo et al. [10] demonstrated that using the clinical diagnosis as a perfect reference leads to an underestimation of cerebrospinal fluid (CSF) AD biomarker sensitivity and specificity values and shifts the cut-offs compared to using the autopsy confirmed diagnosis as reference test.

Different statistical methods have been developed to correctly estimate diagnostic accuracy when an imperfect reference test is used. Reitsma et al. [7] systematically reviewed the different solutions and provided methodological guidelines depending on the medical test under evaluation and the availability and nature of the data.

To date, these methods have not systematically been applied to estimate the diagnostic accuracy of AD biomarkers. An interesting attempt was undertaken by De Meyer et al. [11], who proposed a method to evaluate the CSF AD biomarkers while completely ignoring the clinical diagnosis.

More recently, positron emission tomography (PET) amyloid imaging was used as reference test for evaluation of the diagnostic accuracy of (mainly CSF) AD biomarkers for brain amyloid-β (Aβ) deposition [12]. Although this correctly reduces the time-lag in expected onset of changes between biomarkers and reference test, amyloid PET imaging cannot (yet) be considered a gold standard or a perfect reference test for early AD. There is no true *in vivo* gold standard for amyloid burden and there is substantial overlap between the distribution of PET measurements for presumed AD and non-AD groups [13, 14]. In addition, as for all tests, PET analysis is not free from measurement errors, and standardization of different measurement procedures is still ongoing [14].

As an alternative to search for a surrogate gold standard, it has been suggested that the complexity of dementia diagnosis would be best served by integrating multiple sources of information [3]. A Bayesian framework integrates different data sources in a natural way and is most suited for this purpose.

Bayesian methods have become increasingly popular, notably in medical research [15]. A Bayesian approach can include prior information, accommodate adaptive clinical trials (e.g., interim analyses, change to sample size, or change to randomization scheme) and can be useful for analysis of a complex model when a frequentist analysis is difficult to implement or does not exist [16].

Recent breakthroughs in computational algorithms and computing speed have made it possible to carry out calculations of the often computationally intense Bayesian analysis. Also the fact that regulatory authorities embrace the use of Bayesian statistics has boosted its application in medical research. Already in 2003, the US Food and Drug Administration (FDA) approved a drug combination (pravastatin and aspirin) based on a Bayesian analysis [17]. Likewise, the Center for Devices and Radiological Health of the FDA, that is among others responsible for clearance of diagnostic test kits, issued a guideline for the use of Bayesian statistics and now routinely accepts applications based on Bayesian trials [18].

Bayesian statistics is currently a widespread approach in oncology. Many leading medical journals have published original oncology studies using Bayesian analysis and prominent cancer centers have implemented several clinical trials, which were designed using Bayesian methods [19]. In pediatric science, care providers are accustomed with and often obliged to rely on evidence from adult studies; borrowing information from adult trials using a Bayesian approach is common practice [20]. Also in diagnostic medicine, Bayesian approaches are well-established and often help to validate diagnostics test with smaller-sized and shorter-duration pivotal trial [18, 21].

In this paper, we present a Bayesian framework which establishes the diagnostic accuracy of AD biomarkers by integrating different data sources, without the need for a gold standard or perfect reference test. We applied the new Bayesian analysis method to establish the performance of the three CSF AD biomarkers, $A\beta_{1\text{-}42}$, Total tau, and P-tau_{181p} present in two datasets, with the clinical diagnosis considered as an imperfect reference test. We hypothesized that the diagnostic performance of the CSF AD biomarkers would be higher when analyzed with the Bayesian analysis method that accounts for the imperfectness of the clinical diagnosis.

MATERIALS AND METHODS

Data sets

We used two independent cohorts. The VUmc (VU University Medical Center) data set that consists of patients from the memory-clinic-based Amsterdam Dementia Cohort who received a diagnosis of either subjective memory complaints (SMC) or probable AD. Baseline CSF was collected between October 1999 and November 2011. All patients underwent standard dementia screening at baseline, including physical and neurological examination, EEG, MRI, and laboratory tests. Cognitive screening included a Mini-Mental State Examination (MMSE) and a comprehensive neuropsychological test battery. Diagnoses were made by consensus in a multidisciplinary team without knowledge of CSF results. The label of SMC was given when results of all clinical examinations were normal, and there was no psychiatric diagnosis. Patients with subjective complaints were considered as controls, but were only included when the diagnosis was confirmed at follow-up visits. This resulted in 251 SMC subjects. Probable AD ($n = 631$) was diagnosed according to the criteria of the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders association (NINCDS-ADRDA), and all patients met the core clinical NIA-AA criteria [22]. More details about this cohort have been provided elsewhere [23]. All subjects gave written informed consent for the use of their clinical data for research purposes. The current study was approved by the local ethical review board. CSF levels of $A\beta_{1\text{-}42}$, Total tau, and P-tau_{181p} were determined using commercially available single-parameter ELISA kits (respectively, INNOTEST® AMYLOID(1-42), INNOTEST® hTAU Ag, INNOTEST® PHOSPHOTAU(181P)) and were not used for diagnosis.

The second data set consisted of Alzheimer's Disease Neuroimaging Initiative (ADNI)-I patients. ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the FDA, private pharmaceutical companies, and non-profit organizations. ADNI-I subjects who (i) agreed to undergo a lumbar puncture, (ii) had results for all three CSF biomarkers at baseline, and (iii) belonged to either the control or AD group at baseline, were selected for the current study. This selection resulted in a dataset including 96 AD and 109 control subjects. The CSF biomarker data were obtained using the xMAP platform (Luminex Corp, Austin, Texas) and INNO-BIA AlzBio3 research-use-only reagents.

Table 1 provides baseline characteristics for the two study populations.

Statistical methodology

Measure of diagnostic accuracy

To establish the joint diagnostic accuracy of the AD biomarkers, the biomarkers were combined into a diagnostic score (see below). As a measure of the

Table 1
Baseline characteristics of the study populations (mean ± SD)

Dataset	Group	n	Age (y)	Female (%)	MMSE	Aβ42* (pg/mL)	Tau* (pg/mL)	Ptau-181* (pg/mL)
VUmc	SMC	251	64 ± 6.6	104 (41)	28 ± 1.5	874 ± 251.0	302 ± 197.7	52 ± 24.0
	AD	631	68 ± 7.5	326 (52)	21 ± 5.0	465 ± 161.6	690 ± 415.4	89 ± 39.2
ADNI	Control	109	76 ± 5.3	55 (50)	29 ± 1.0	206 ± 54.4	69 ± 30.2	25 ± 14.8
	AD	96	75 ± 8.0	40 (42)	24 ± 1.9	142 ± 4.0	122 ± 57.0	42 ± 19.8

*CSF levels of Aβ1-42, Total tau, and P-tau181p were determined using commercially available single-parameter ELISA kits (INNOTEST® AMYLOID(1-42), INNOTEST® hTAU Ag, INNOTEST® PHOSPHOTAU(181P)) and using the xMAP platform (Luminex Corp, Austin, Texas) and INNO-BIA AlzBio3 reagents at VUmc and ADNI, respectively.

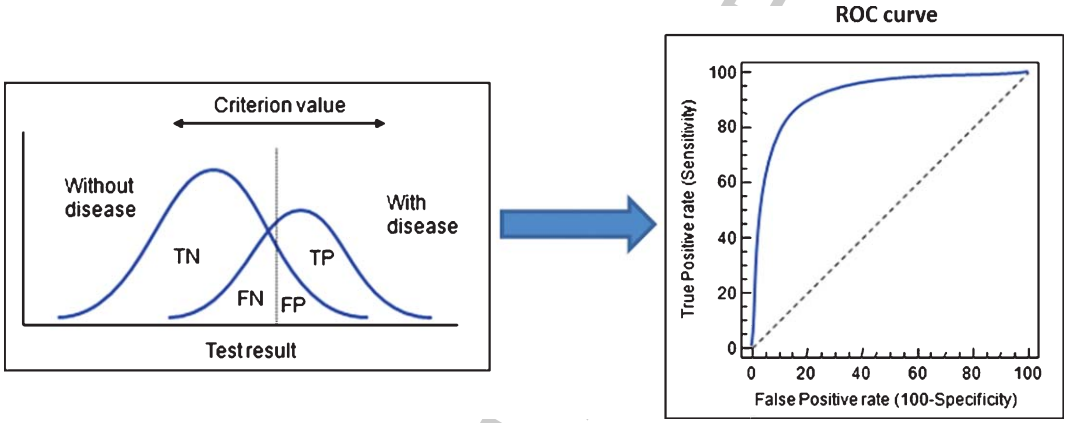

Fig. 1. Schematic summary on the construction of a receiver operating characteristic (ROC) curve and interpretation of the area under the ROC curve (AUC). The ROC curve is a plot of the sensitivity and (1-specificity) for each value of a continuous diagnostic marker. AUC can be interpreted as the probability that, for a randomly selected pair of non-AD and AD subjects, the value of the score for the AD subject will be larger than the value of the non-AD subject. For a score that perfectly separates non-AD and AD populations, the value of AUC is equal to 1, corresponding to the ROC curve passing through the (0,1) point, i.e., the point corresponding to a diagnostic test with 100% sensitivity and 100% specificity. For a score that has no discriminative ability, the value of AUC is equal to 0.5, corresponding to a ROC curve along the diagonal line. TP, true positive; FP, false positive; TN, true negative; FN, false negative.

diagnostic performance of this score, the area under the receiver-operating-characteristics (ROC) curve (AUC) was used (Fig. 1).

AD biomarker performance using a Bayesian framework that accounts for an imperfect clinical diagnosis

To account for possible errors in the clinical diagnosis, both the AD biomarkers AND the clinical diagnosis were considered as data sources carrying information about the (unknown) disease status of the subjects. Note that, in a classical analysis, the clinical diagnosis would be taken as the correct disease status, which does not reflect reality.

A Bayesian framework integrates different data sources in a natural way and is hence most suited for our purpose. At the core of the Bayesian approach lays the use of prior information [15]. The information (hereafter also termed 'prior opinion' or 'prior information') is provided in the form of probability distributions for the parameters of a model. The distribution indicates which (sets of) values of the parameters are considered to be (relatively) more likely than others. In particular, uninformative distributions (e.g., a normal distribution with a huge variance) can be used in the data analysis to imply the absence of any information, i.e., the fact that all values of a particular parameter are equally likely. If some information is available, informative prior distributions are used.

By combining the prior distribution with the data, a posterior distribution for the parameter of interest is obtained. The posterior distribution reflects the change of the opinion induced by the data, as compared to the prior opinion (see Fig. 2). When uninformative prior distributions are used, the data is used as the only source of information. In Bayesian analysis, it is best practice to perform a 'sensitivity analysis' using different priors to disentangle the effect of prior information and the analysis dataset on the reported results.

In our analyses, we made the 'conditional independence assumption', i.e., we assumed that AD biomarkers and clinical diagnosis do not misclassify

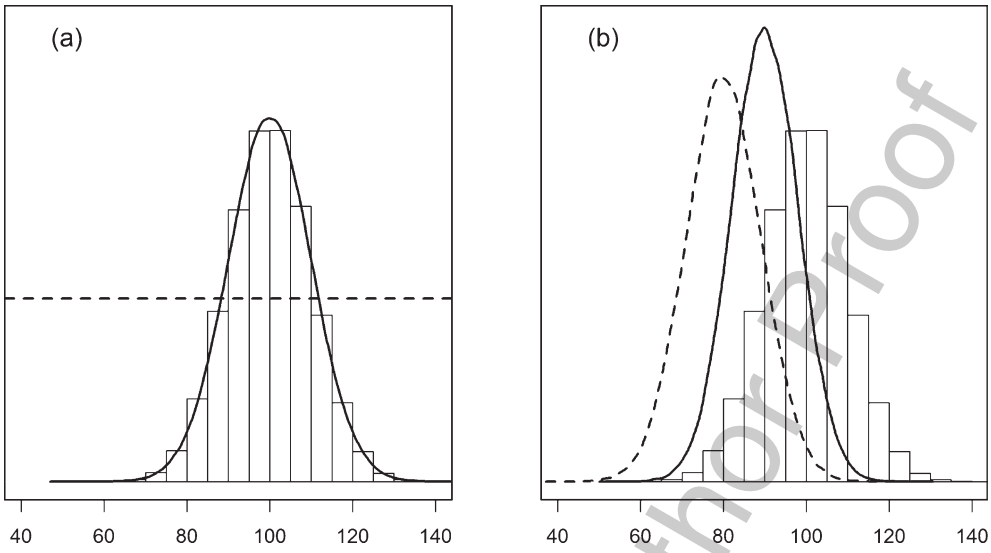

Fig. 2. Schematic illustration of the Bayesian 'Change of Opinion' approach. X-axis: Parameter of interest (e.g., average biomarker concentration in pg/μl). Y-axis: Probability of occurrence. Histogram: observed data. Dashed lines: prior opinion ('prior distribution'). Solid lines: opinion after obtaining data ('posterior distribution'), Panel (a): Application of an uninformative prior amounts to forming an opinion based solely on the observed the data. The horizontal (uninformative) prior distribution indicates that, before data collection, each value is considered equally likely to occur. As a result, the posterior distribution coincides with the observed-data histogram. Panel (b): Application of an informative prior amounts to forming an opinion based on combining the prior information and the observed the data. The bell-shaped (informative) prior distribution indicates that, before data collection, the parameter of interest lies, with 95% probability, within the range of 62 to 98. The posterior distribution combines the prior information with the observed data. As a result, the obtained posterior distribution is different from the prior distribution and from the histogram, and it indicates that the value of the parameter lies, with 95% probability, within the range of 75 to 105.

the same individuals. The diagnostic score was constructed by using a linear combination of the biomarkers that maximizes AUC for normally-distributed biomarkers [24].

We used prior distributions for the following parameters: the AUC of a combination of biomarkers, the mean value for each biomarker in the non-AD population, the variances and correlations between all biomarkers in both populations, the prevalence of diseased cases, the sensitivity of the clinical diagnosis, and the specificity of clinical diagnosis. We used uninformative prior distributions for the biomarkers' means, variances and correlations, and for the disease prevalence.

For the AUC of the linear combination of AD biomarkers, we used more informative priors based on a paper containing data from 12 publications that reported a joint AUC for CSF biomarkers [25]. The lowest reported joint AUC was equal to 0.90 (no standard error provided) [26] and the highest value was equal to 0.997 (95% CI 0.926–1) [27]. Based on those data, we formulated two prior distributions for the joint AUC (Fig. 3a). The first prior distribution implied that the probability that the AUC was larger than 0.7 and 0.9 was equal to 90% and 30%, respectively. This prior was labeled as 'optimistic' in the sense that it pointed toward a high diagnostic accuracy. The second prior distribution choice was labeled as 'skeptical' as it suggested that the AUC was around 0.75, with only 5% probability that it exceeded 0.90, the lowest value reported [25].

Also for the specificity and sensitivity of the clinical diagnosis, we used informative priors. Three studies [4, 28, 29] reported high sensitivity of the clinical AD diagnosis (ranging from 81.8% to 100%) in a mixed dementia setting; another study [30] reported much worse sensitivities ranging from 39% to 95% and specificities ranging from 33% to 100%. Based on those data, we formulated two prior distributions (Fig. 3b). The first, 'optimistic' prior in accordance with [4, 28, 29], suggested a sensitivity and specificity of about 90%, with 5% probability that sensitivity and specificity were below 80%. The second, more 'skeptical' prior, in accordance with [30], was centered at 59%, with a 95% probability that sensitivity and specificity were larger than 25%. The 'skeptical' prior assumed less information about the performance of the clinical diagnosis and allowed more flexibility for the biomarkers to 'overrule' the clinical diagnosis, as compared to the 'optimistic' prior distribution.

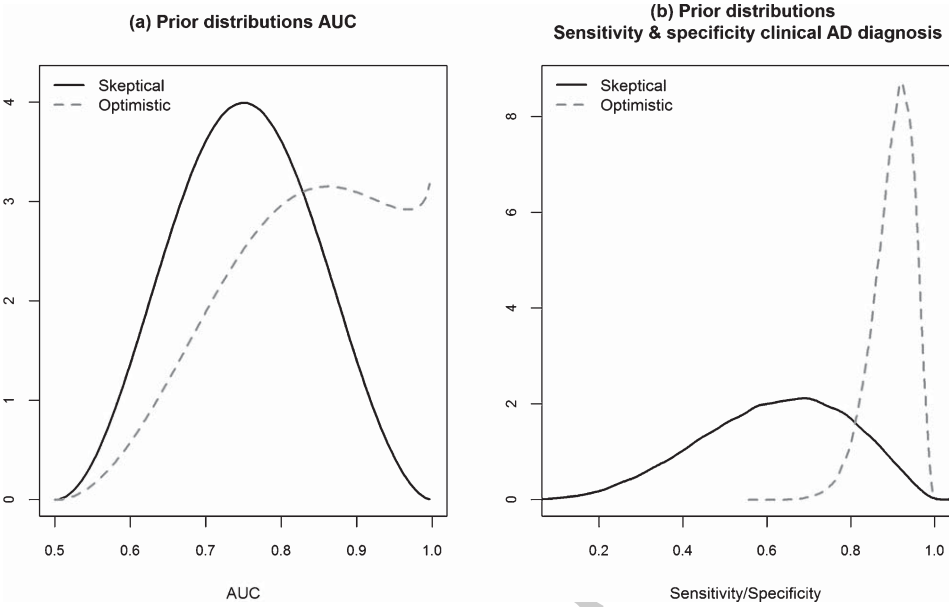

Fig. 3. Prior distributions for the AUC (a) and sensitivity/specificity of the clinical diagnosis (b).

If we treat the clinical diagnosis as an imperfect reference test, the true disease status of the subjects is unknown. It is hence not possible to use a binary classifier to establish a ROC curve. Informally speaking, the model we use predicts the disease status of the individuals that best fits the biomarker and clinical data. At the same time, the parameters of a multivariate normal distribution for the biomarkers are estimated for each group, defined by the predicted disease status of the individuals. Based on the estimated distributional parameters, a 'bi-normal ROC-curve' [9] is obtained, providing estimates of sensitivity and specificity. More details on the Bayesian methodology can be found in the Supplementary Material.

AD biomarker performance assuming that the clinical diagnosis is a perfect reference test

To evaluate the impact of allowing for errors in the clinical diagnosis, we also performed two analyses that assumed that the clinical diagnosis indicates the correct disease status.

First, the data were analyzed using logistic regression, a methodology that is often applied to evaluate AD biomarkers' performance [31, 32]. A diagnostic score was calculated with the regression parameters and the diagnostic performance of this score was evaluated against the clinical diagnosis.

Second, we analyzed the AD biomarkers' performance with the new Bayesian method (see above), assuming that the clinical diagnosis is a perfect reference test. Toward this end, sensitivity and specificity of clinical diagnosis in the Bayesian model were set to 1 (i.e., 'extremely' informative priors were used) and the prevalence of AD was estimated by the proportion of clinical AD subjects in the datasets.

By comparison of the results obtained for the latter two analyses the effect of the methodology (Bayesian method versus classical logistic regression) could be evaluated. In addition, the comparison of the results of the two Bayesian analyses allowed the evaluation of the effect of handling the clinical diagnosis data (perfect versus imperfect reference test) on the assessment of the diagnostic performance of the AD biomarkers.

Model fitting

The proposed Bayesian method assumed that all biomarkers display a normal distribution. To conform to this assumption, Total tau and Ptau-181p values were log transformed for all analyses. The analyses were performed using R [33], version 3.0.1 and OpenBUGS [34]. More information on model fitting is provided in the Supplementary Material.

After fitting the models, the median AUC was obtained from the posterior distribution, together with a 95% credible interval (CrI), the Bayesian counterpart of the 'classical' confidence interval (CI). CrI provides the range of values that are expected with 95% probability according to the (posterior) distribution.

RESULTS

Figure 4 shows the ROC curves for different analyses of the VUmc data (grey) and ADNI data (black). In particular, it shows the curves for the analysis using the logistic regression (dotted), for the Bayesian model obtained by assuming a perfect reference test (dashed), and by assuming an imperfect reference test (solid). Note that the latter were obtained by using the 'skeptical' AUC prior and 'optimistic' priors for sensitivity and specificity of the clinical diagnosis.

The ROC curves for the logistic regression are close to the curves corresponding to the Bayesian model that also assumed that the clinical diagnosis is a perfect reference test. These results show that the Bayesian method in principle yields the same results as the 'classical' logistic regression, proving confidence in our approach. Consequently, we have further focused on the Bayesian methodology.

When assuming that the clinical diagnosis is an imperfect reference test, the ROC curves are higher compared to the corresponding curves obtained when assuming that the reference test is perfect. This shows that, by assuming that the clinical diagnosis flawlessly indicates the pathophysiological AD status, one underestimates the joint diagnostic performance of the biomarkers.

In particular, for the VUmc dataset, the median AUC was equal to 0.949 with 95% CrI [0.935,0.960] when the diagnosis was treated as a perfect reference test and 0·990 with 95% CrI [0.985,0.995] when treated as an imperfect reference test. For the ADNI data, the corresponding values were equal to 0.870 (95% CrI: [0.817,0.912]) and 0·975 (95% CrI: [0.943,0.990]), respectively.

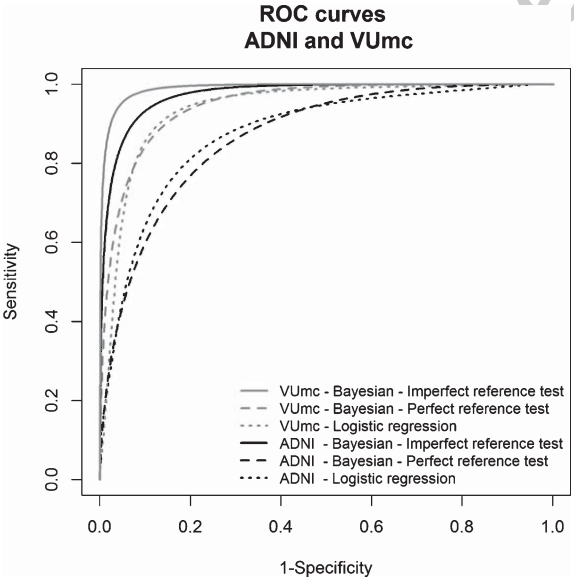

Fig. 4. ROC curves for different analyses for VUmc (grey) and ADNI (black) dataset.

Figure 5 shows the results of analyses with different prior distributions. The difference between the ROC curves (and hence, AUC) obtained with different combinations of the 'optimistic' and 'skeptical' prior distributions for the AUC and sensitivity and specificity of the clinical diagnosis was minimal (Fig. 5).

DISCUSSION

By applying the newly developed Bayesian method to the two datasets, we were able to show that the AUC to discriminate between subjects with AD pathology and controls, increases from 0.949 (with 95% credible interval [0.935,0.960]) to 0.990 ([0.985,0.995]) and from 0.870 ([0.817,0.912]) to 0·975 ([0.943,0.990]) for the VUmc and ADNI cohorts, respectively.

This effect can be intuitively explained as follows. With an imperfect clinical diagnosis, some individuals will be diagnosed as non-AD, while their AD biomarkers may be indicative of existing AD pathophysiology, as biomarker abnormalities can occur decades before clinical symptoms become apparent [35]. For these individuals, the AD biomarkers will be considered as 'incorrect' if the clinical diagnosis is regarded as the perfect reference test. Consequently, the performance of the biomarkers will be underestimated. It is in this complex situation that our proposed approach is most useful [7, 36], enabling an estimation of the biomarkers' performance by objectively examining the strength of statistical relationships among variables.

We applied a Bayesian approach because this allowed integrating different sources of information, while taking into account the absence of a perfect reference test. In Bayesian inference, the specification of prior distributions for the model parameters is needed. It is good practice to perform a sensitivity analysis to check the influence of the choice of the prior distributions on the results and to disentangle the effect of the prior distributions and of the data on the reported results. Toward this end, 'skeptical' and 'optimistic priors' for the biomarkers' AUC and sensitivity and specificity of the clinical diagnosis were used in our analysis. The 'skeptical' priors were only weakly informative (containing little prior information) while the 'optimistic' priors contained more information that pointed to a better diagnostic

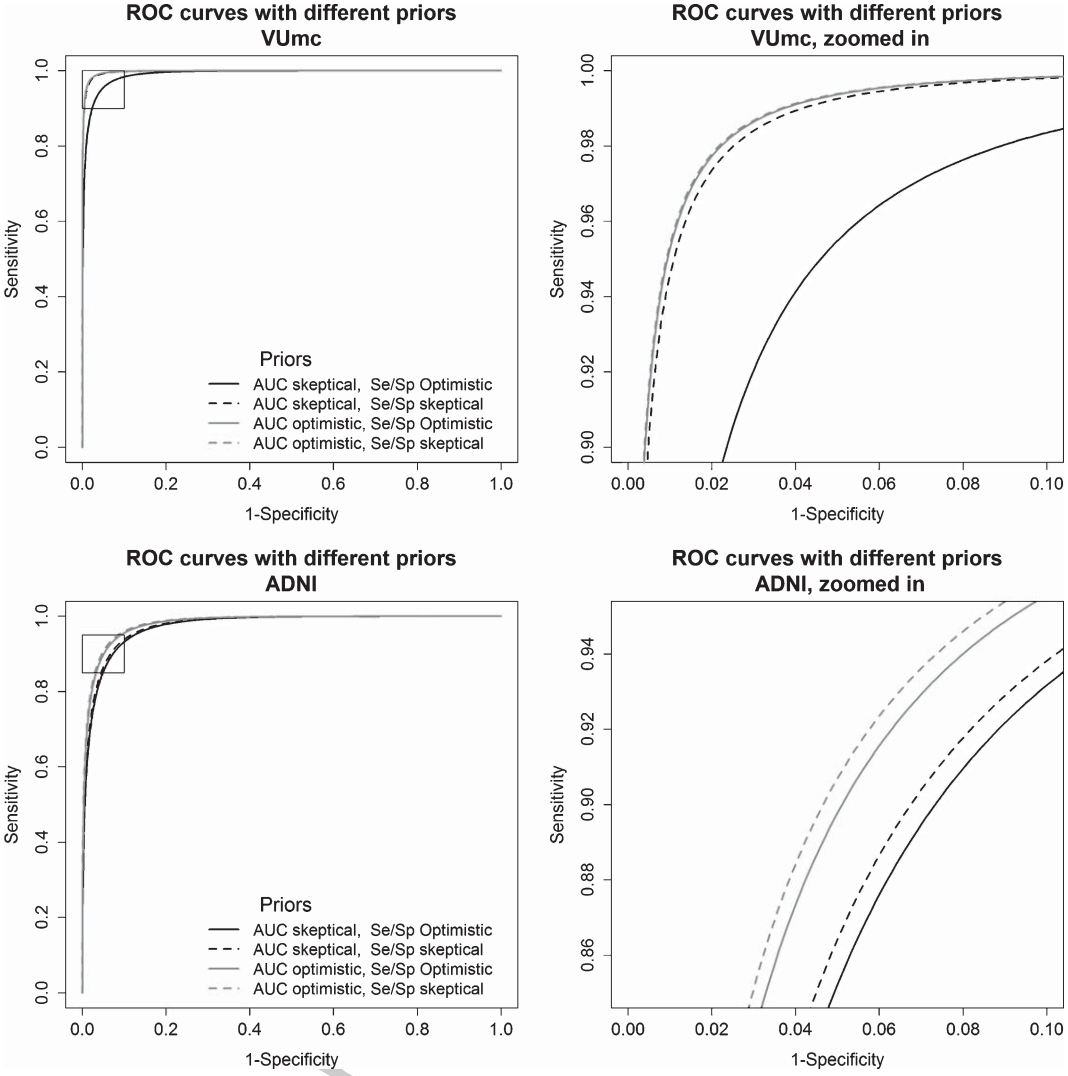

Fig. 5. Sensitivity analysis: ROC curves when using different priors for VUmc (top row) and ADNI dataset (bottom row). The graphs on the right represent the same ROC curves as the graphs on the left, but are zoomed in to the rectangle in the upper left corner. Note that, for VUmc, the solid and dashed grey line overlap.

performance of the biomarkers or clinical diagnosis as found in literature. Figure 5 shows that the different combinations of these prior distributions resulted in minimally different ROC-curves. This implies that our conclusions are robust to reasonable changes in prior distributions for the diagnostic performance of the biomarkers and clinical diagnosis. Put differently, the results presented in Fig. 4 are mainly driven by the data and not by the prior information.

All statistical analyses rely on assumptions. Bayesian statistics has the advantage to encourage a thorough consideration and presentation of the assumptions underlying the performed analysis. We have avoided the assumption that the reference test is perfect, because this has been reported to cause biased diagnostic accuracy results [7, 36, 37]. The validity of the presented approach relies on the assumption that the clinical diagnosis and AD biomarkers do not misclassify the same subjects (the 'conditional independence assumption'). At this point, mainly heuristic arguments can be offered for the plausibility of this assumption. As long as the clinical diagnosis is not based on the CSF biomarkers, we can assume that the biomarkers and clinical diagnosis do not tend to misclassify the same subjects. Furthermore, our findings are in line with the reports on lower diagnostic performance of CSF biomarkers when evaluated against the clinical

diagnosis instead of the pathology confirmed diagnosis [10].

There is no gold standard for a complex disease like AD [3]. We show that this is no longer an issue as the developed Bayesian methodology can deal with the absence of a perfect reference test. The new approach is constructed by assembling components of methods that have been proposed for the evaluation of the diagnostic performance of a combination of markers [39] when no perfect reference test is available [36]. To our knowledge, this is the first report of the use of a Bayesian approach to define the diagnostic performance of AD biomarkers that acknowledges the absence of a perfect reference test.

The new methodology is based on well-established statistical concepts, but is more complicated than a simple comparison with the clinical diagnosis or dichotomized PET data as outcome. It is, however, the complexity of a dementia diagnosis that calls for appropriate, more advanced analysis methods.

The reported diagnostic accuracy results are relevant only for discrimination between the two well-defined groups in this study namely AD versus SMC/control. These estimates of diagnostic accuracy are often higher than expected in the target patient population which contains difficult-to-diagnose subjects (e.g., MCI patients) [7]. This is not an issue for the purpose of our manuscript, as our goals were to develop a new method that allows for an imperfect reference test and to compare the resulting estimates of diagnostic accuracy with those obtained by currently applied methodologies. In practice, these extremely high accuracy estimates will not be achieved because the target patient population will contain difficult-to-diagnose subjects (such as MCI patients) and patients with different types of dementia.

However, the estimates of diagnostic accuracy are expected to be higher in the target patient population when estimated with the Bayesian analysis as compared to a classical analysis with the clinical diagnosis as perfect reference test. Although the patterns of differences between the results for the different models (Fig. 4) were identical for VUmc and ADNI datasets, the numerical values of the AUC estimates were not. For each of the three models, the combined biomarkers' AUC was higher for the VUmc data than for the ADNI data. This difference is most likely due to the higher age of the ADNI subjects (on average about 10 years older than VUmc subjects), as it is well-known that the diagnostic accuracy of CSF AD biomarkers decreases with age [38].

The new methodology can now be used for re-investigation of the clinical value of existing AD biomarkers to determine which CSF biomarkers are needed for maximum discriminate between stable and progressing MCI patients or for a differential dementia diagnosis. The cut-offs that would be derived from the ROC-curve of the new method will be different from the current cut-offs values that are set with the clinical diagnosis as perfect reference test. Also the comparison of the clinical value between CSF biomarkers measured using different platforms or Aβ PET deposition measured with different tracers could be addressed. Importantly, the new analysis method also supports the direct comparison of the diagnostic value of CSF and imaging biomarkers for Aβ deposition. In this way, the interchangeability (assumed in the (preclinical) AD criteria [1, 5]) or complementarity (as suggested by the reported proportion of discordant cases [12–14]) of the two *in vivo* biomarkers could be determined. We anticipate that the use of the new Bayesian framework will lead to a more accurate diagnosis based on biomarkers and hence more diagnostic confidence in early stages of AD.

ACKNOWLEDGMENTS

This research was conducted within the framework of the European EUROTRANS-BIO – ERA-NET project 'B4AD', a collaborative project of International Drug Development Institute (Louvain-la-Neuve, Belgium), PamGene International (Den Bosch, The Netherlands) and the VU University medical center and the Alzheimer center (Amsterdam, The Netherlands). We thank Riet Hilhorst, Faris Naji, Rik de Wijn, and Rinie van Beuningen (PamGene) for helpful discussions.

Research of the VUmc Alzheimer center and the Department of Pathology is part of the Neurodegeneration research program of the Neuroscience Campus Amsterdam. The VUmc Alzheimer center is supported by Alzheimer Nederland and Stichting VUmc fonds. The VUmc clinical database structure was developed with funding from Stichting Dioraphte.

Authors' disclosures available online (http://j-alz.com/manuscript-disclosures/14-2886r2).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the

National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is Rev December 5, 2013 coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-142886.

REFERENCES

[1] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (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.

[2] Vos SJB, Xiong C, Visser PJ, Jasielec MS, Hassenstab J, Grant EA, Cairns NJ, Morris JC, Holtzman DM, Fagan AM (2013) Preclinical Alzheimer's disease and its outcome: A longitudinal cohort study. *Lancet Neurol* **12**, 957-965.

[3] Scheltens P, Rockwood K (2011) How golden is the gold standard of neuropathology in dementia? *Alzheimers Dement* **7**, 486-489.

[4] Beach TG, Monsell SE, Phillips LE, Kukull W (2012) Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *J Neuropathol Exp Neurol* **71**, 266-273.

[5] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR Jr, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH (2011) Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging and the Alzheimer's Association workgroup. *Alzheimers Dement* **7**, 280-292.

[6] Salloway S, Sperling R, Fox NC, Blennow K, Klunk W, Raskind M, Sabbagh M, Honig LS, Porsteinsson AP, Ferris S, Reichert M, Ketter N, Nejadnik B, Guenzler V, Miloslavsky M, Wang D, Lu Y, Lull J, Tudor IC, Liu E, Grundman M, Yuen E, Black R, Brashear HR; Bapineuzumab 301 and 302 Clinical Trial Investigators (2014) Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *N Engl J Med* **370**, 322-333.

[7] Reitsma JB, Rutjes AWS, Khan KS, Bossuyt PM (2009) A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol* **62**, 797-806.

[8] Valenstein PN (1990) Evaluating diagnostic tests with imperfect standards Am. *J Clin Pathol* **93**, 252-258.

[9] Zhou X-H, Obuchowski NA, McClish DK (2002) *Statistical Methods in Diagnostic Medicine,* Wiley Inc., New York.

[10] Toledo JB, Brettschneider J, Grossman M, Arnold SE, Hu WT, Xie SX, Lee VM, Shaw LM, Trojanowski JQ (2012) CSF biomarkers cutoffs: The importance of coincident neuropathological diseases. *Acta Neuropathol* **124**, 23-35.

[11] De Meyer G, Shapiro F, Vanderstichele H, Vanmechelen E, Engelborghs S, De Deyn PP, Coart E, Hansson O, Minthon L, Zetterberg H, Blennow K, Shaw L, Trojanowski JQ, Alzheimer's Disease Neuroimaging Initiative (2010) Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol* **67**, 949-956.

[12] Fagan AM, Shaw LM, Xiong C, Vanderstichele H, Mintun MA, Trojanowski JQ, Coart E, Morris JC, Holtzman DM (2011) Comparison of analytical platforms for cerebrospinal fluid measures of β-amyloid 1-42, Total tau, and P-tau181 for identifying Alzheimer disease amyloid plaque pathology. *Arch Neurol* **68**, 1137-1144.

[13] Palmqvist S, Zetterberg H, Blennow K, Vestberg S, Andreasson U, Brooks DJ, Owenius R, Hägerström D, Wollmer P, Minthon L, Hansson O (2014) Accuracy of brain amyloid detection in clinical practice using cerebrospinal fluid β-amyloid 42. A cross-validation study against amyloid positron emission tomography. *JAMA Neurol* **71**, 1282-1289.

[14] Landau SM, Lu M, Joshi AD, Pontecorvo M, Mintun MA, Trojanowski JQ, Shaw LM, Jagust WJ, Alzheimer's Disease Neuroimaging Initiative (2013) Comparing positron emission tomography imaging and cerebrospinal fluid measurements of β-amyloid. *Ann Neurol* **74**, 826-836.

[15] Spiegelhalter DJ, Abrams KR, Myles JP (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluations.* Wiley Inc., New York.

[16] Gelman A, Carlin JB, Stern HS, Rubin DB (2003) *Bayesian Data Analysis, second edition.* Chapman & Hall/CRC, New York.

[17] Berry SM, Berry DA, Natarajana K, Lina C-S, Hennekens CH, Belder R (2004) Bayesian survival analysis with nonproportional hazards. *J Am Stat Assoc* **99**, 36-44.

[18] Anonymous (2010) Guidance for Industry and FDA Staff; Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. http://www.fda.gov/downloads/

MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf

[19] Adamina M, Tomlinson G, Guller U (2009) Bayesian statistics in oncology. *Cancer* **115**, 5371-5381.

[20] Schoenfeld DA, Zheng H, Finkelstein DM (2009) Bayesian design using adult data to augment pediatric trials. *Clin Trials* **6**, 297-304.

[21] Broemeling LD (2007) *Bayesian Biostatistics and Diagnostic Medicine.* Chapman & Hall/CRC, New York.

[22] Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P (2007) Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS–ADRDA criteria. *Lancet Neurol* **6**, 734-746.

[23] Duits FH, Teunissen CE, Bouwman FH, Visser PJ, Mattsson N, Zetterberg H, Blennow K, Hansson O, Minthon L, Andreasen N, Marcusson J, Wallin A, Rikkert MO, Tsolaki M, Parnetti L, Herukka SK, Hampel H, De Leon MJ, Schröder J, Aarsland D, Blankenstein MA, Scheltens P, van der Flier WM (2014) The cerebrospinal fluid 'Alzheimer profile': Easily said, but what does it mean? *Alzheimers Dement* **10**, 713-723.

[24] Su JQ, Liu JS (1993) Linear combinations of multiple diagnostic markers. *J Am Stat Assoc* **88**, 1350-1355.

[25] Bloudek LM, Spackman DE, Blankenburg M, Sullivan SD (2011) Review and meta-analysis of biomarkers and diagnostic imaging in Alzheimer's disease. *J Alzheimers Dis* **26**, 627-645.

[26] Ibach B, Binder H, Dragon M, Poljansky S, Haen E, Schmitz E, Koch H, Putzhammer A, Kluenemann H, Wieland W, Hajak G (2006) Cerebrospinal fluid tau and beta-amyloid in Alzheimer patients, disease controls and an age-matched random sample. *Neurobiol Aging* **27**, 1202-1211.

[27] Kapaki E, Liappas I, Paraskevas GP, Theotoka I, Rabavilas A (2005) The diagnostic value of tau protein, beta-amyloid (1-42) and their ratio for the discrimination of alcohol-related cognitive disorders from Alzheimer's disease in the early stages. *Int J Geriatr Psychiatry* **20**, 722-729.

[28] Schoonenboom NS, Reesink FE, Verwey NA, Kester MI, Teunissen CE, van de Ven PM, Pijnenburg YA, Blankenstein MA, Rozemuller AJ, Scheltens P, van der Flier WM (2012) Cerebrospinal fluid markers for differential dementia diagnosis in a large memory clinic cohort. *Neurology* **3**, 47-54.

[29] Toledo JB, Cairns NJ, Da X, Chen K, Carter D, Fleisher A, Householder E, Ayutyanont N, Roontiva A, Bauer RJ, Eisen P, Shaw LM, Davatzikos C, Weiner MW, Reiman EM, Morris JC, Trojanowski JQ; Alzheimer's Disease Neuroimaging Initiative (ADNI) (2013) Clinical and multimodal biomarker correlates of ADNI neuropathological findings. *Acta Neuropathol Commun* **1**, 65.

[30] Wollman DE, Prohovnik I (2003) Sensitivity and specificity of neuroimaging. *Dialogues Clin Neurosci* **5**, 89-99.

[31] Hansson O, Zetterberg H, Buchhave P, Londos E, Blennow K, Minthon L (2006) Association between CSF biomarkers and incipient Alzheimer's disease in subjects with mild cognitive impairment: A follow-up study. *Lancet Neurol* **5**, 228-234.

[32] Mattsson N, Zetterberg H, Hansson O, Andreasen N, Parnetti L, Jonsson M, Herukka SK, van der Flier WM, Blankenstein MA, Ewers M, Rich K, Kaiser E, Verbeek M, Tsolaki M, Mulugeta E, Rosén E, Aarsland D, Visser PJ, Schröder J, Marcusson J, de Leon M, Hampel H, Scheltens P, Pirttilä T, Wallin A, Jönhagen ME, Minthon L, Winblad B, Blennow K (2009) CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA* **302**, 385-393.

[33] R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[34] Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: Evolution, critique and future directions (with discussion). *Stat Med* **28**, 3049-3082.

[35] Jack CR Jr, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Pankratz VS, Donohue MC, Trojanowski JQ (2013) Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *Lancet Neurol* **12**, 207-216.

[36] Scott AN, Joseph L, Bélisle P, Behr MA, Schwartzman K (2007) Bayesian modeling of tuberculosis clustering from DNA fingerprint data. *Stat Med* **27**, 140-156.

[37] Lu Y, Dendrukuri N, Schiller I, Joseph L (2010) A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Stat Med* **29**, 2532-2543.

[38] Mattsson N, Rosén E, Hansson O, Andreasen N, Parnetti L, Jonsson M, Herukka SK, van der Flier WM, Blankenstein MA, Ewers M, Rich K, Kaiser E, Verbeek MM, Olde Rikkert M, Tsolaki M, Mulugeta E, Aarsland D, Visser PJ, Schröder J, Marcusson J, de Leon M, Hampel H, Scheltens P, Wallin A, Eriksdotter-Jönhagen M, Minthon L, Winblad B, Blennow K, Zetterberg H (2012) Age and diagnostic performance of Alzheimer disease CSF biomarkers. *Neurology* **78**, 468-476.

[39] O'Malley AJ, Zou KH, Fielding JR, Tempany CMC (2001) Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: Prostrate biopsy and spiral CT of uteral stones. *Acad Radiol* **8**, 713-725.