Biometrics 73, 646–655
June 2017

DOI: 10.1111/biom.12583

Estimation of Diagnostic Accuracy of a Combination of Continuous Biomarkers Allowing for Conditional Dependence between the Biomarkers and the Imperfect Reference-Test

Leandro García Barrado,[1,*] Els Coart,[2] and Tomasz Burzykowski[1,2]

[1]Hasselt University, I-BioStat, Agoralaan, B-3590 Diepenbeek, Belgium
[2]International Drug Development Institute (IDDI), Avenue Provinciale 30, 1340 Louvain-la-Neuve, Belgium
*email: leandro.garciabarrado@uhasselt.be

Summary. Estimating biomarker-index accuracy when only imperfect reference-test information is available is usually performed under the assumption of conditional independence between the biomarker and imperfect reference-test values. We propose to define a latent normally-distributed tolerance-variable underlying the observed dichotomous imperfect reference-test results. Subsequently, we construct a Bayesian latent-class model based on the joint multivariate normal distribution of the latent tolerance and biomarker values, conditional on latent true disease status, which allows accounting for conditional dependence. The accuracy of the continuous biomarker-index is quantified by the AUC of the optimal linear biomarker-combination. Model performance is evaluated by using a simulation study and two sets of data of Alzheimer's disease patients (one from the memory-clinic-based Amsterdam Dementia Cohort and one from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database). Simulation results indicate adequate model performance and bias in estimates of the diagnostic-accuracy measures when the assumption of conditional independence is used when, in fact, it is incorrect. In the considered case studies, conditional dependence between some of the biomarkers and the imperfect reference-test is detected. However, making the conditional independence assumption does not lead to any marked differences in the estimates of diagnostic accuracy.

Key words: Alzheimer's disease; Bayesian Estimation; Biomarker; Conditional dependence; Imperfect reference-test.

1. Introduction

An important issue in the development of diagnostic tests is the availability of the correct case and control labels. However, sometimes such labels based on a gold standard (GS) reference-test may not be available or useful. For example, in the context of dementia or Alzheimer's disease (AD), only post-mortem pathological confirmation on brain tissue can be regarded as a GS reference-test (Scheltens and Rockwood, 2011); obviously, the confirmation is useless from a diagnostic perspective.

Hence, case and control labels for diagnostic-test development may be based on the result of an imperfect reference-test, which may misclassify cases and controls. If the misclassification is ignored in the development of a diagnostic test, the estimates of the parameters describing its accuracy may be severely biased (Lu et al., 2010).

To take into account the absence of a gold standard when assessing the accuracy of diagnostic tests, latent-class models with two latent classes have been proposed (Rindskopf and Rindskopf, 1986). The models allow inclusion of the imperfect reference-test information in the form of covariate information, but require certain strict identifiability restrictions.

Preferably, one would like to weigh the information present in the results of an imperfect reference-test according to the prior knowledge about the accuracy of the test. For example, in AD-biomarker research, clinical diagnosis of AD can be regarded as an imperfect reference-test. Reports about the accuracy of the diagnosis are available in the literature (Wollman and Prohovnik, 2004; Beach et al., 2012). Using this information might be instrumental in obtaining more reliable estimates of biomarker accuracy and in mitigating issues of identifiability. This is possible within the Bayesian framework.

In case a GS reference-test is available, a fully parametric Bayesian method to estimate the accuracy of univariate continuous diagnostic tests was introduced by O'Malley et al. (2001). For the case of an imperfect reference-test, several Bayesian models were proposed, including Bayesian latent-class mixture models. In particular, Wang et al. (2006) developed a Bayesian latent-class mixture model for a single continuous test, which considers use of a dichotomous imperfect reference-test.

Often, diagnostic tests are developed based on biomarkers. A biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to therapeutic interventions" (Biomarkers Definition Working Group, 2001). An often-used method to summarize properties of a continuous biomarker-based diagnostic test is the area under the receiver operating characteristic curve (AUC) (Zou et al., 2012).

Combining continuous biomarkers has been shown to increase the diagnostic accuracy over considering biomarkers alone. In particular, such a combination can be constructed when maximizing the AUC by using a model-free approach (Huang, Qin, and Fang, 2011), a discriminant-function approach (O'Malley and Zou, 2006), or a fully parametric

approach (Su and Liu, 1993). Yu, Zhou, and Bandinelli (2011) developed a Bayesian latent-class mixture model to estimate the optimal linear-combination of multiple continuous tests.

All of the aforementioned Bayesian methods of estimation of the diagnostic-test accuracy of continuous biomarkers in the presence of an imperfect reference-test are based on the "conditional independence" assumption, i.e., they assume that, conditionally on the true disease status, the misclassification error of the reference-test is independent of the error of the candidate diagnostic test. As the methods do not allow formal testing of this assumption, they have to rely on heuristic arguments to enforce its plausibility.

In the present article, we propose a Bayesian latent-class mixture model to develop a diagnostic test based on a linear combination of multiple continuous biomarkers when an imperfect reference-test is available. Moreover, we allow for conditional dependence between the continuous biomarkers and the imperfect reference-test. On one hand, our model is an extension of the approaches developed by O'Malley and Zou (2006) for the case of a GS reference-test and by Yu, Zhou, and Bandinelli (2011) for the case when no reference test is available. On the other hand, we extend the model proposed by Wang et al. (2006) by developing an optimal linear-combination of continuous tests/biomarkers maximizing the AUC of the combination. Finally, we show that the proposed model could prove an important tool in AD-biomarker research, where admitting the imperfect nature of clinical diagnosis could be essential in obtaining reliable estimates of biomarkers' diagnostic accuracy.

2. Methods

2.1. *Model*

In the remainder of the article, the following assumptions and notation will be used. The K-dimensional vector $\boldsymbol{y}$ contains observations of K biomarkers. Conditional on the true disease status D ($D = 0$ for controls and $D = 1$ for cases), $\boldsymbol{y}$ is assumed to be normally distributed:

$$\boldsymbol{y}_{|D=d} \sim N_K\left(\boldsymbol{\mu}_{Y_d}, \boldsymbol{\Sigma}_d\right), \tag{1}$$

with $\boldsymbol{\mu}_{Yd}$ and $\boldsymbol{\Sigma}_d$ denoting the underlying biomarker mean-vector and variance–covariance matrix, respectively. Under the assumption of normality, it can be shown that the linear combination of biomarkers with maximal AUC has coefficients $\boldsymbol{a} \propto (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}\left(\boldsymbol{\mu}_{Y_1} - \boldsymbol{\mu}_{Y_0}\right)$ (Su and Liu, 1993). The AUC of this linear combination is then

$$AUC_{\boldsymbol{a}} = \Phi\left(\sqrt{\left(\boldsymbol{\mu}_{Y_1} - \boldsymbol{\mu}_{Y_0}\right)^T (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}\left(\boldsymbol{\mu}_{Y_1} - \boldsymbol{\mu}_{Y_0}\right)}\right), \tag{2}$$

where $\Phi(\cdot)$ represents the cumulative standard-normal distribution function. From equation (2) it can be seen that $AUC_{\boldsymbol{a}}$ is a rather complex function of the biomarker distribution parameters.

Assume next that only imperfect information on the true disease status is available in the form of an imperfect reference-test T. In other words, D is not observed, but is considered a latent variable, from now on denoted by $\tilde{D}$. We assume that

$$T \mid \tilde{D} = d \sim Bern\left(\pi_d\right), \tag{3}$$

$$\pi_0 = P\left(T = 1 \mid \tilde{D} = 0\right) \equiv 1 - Sp_T,$$

$$\pi_1 = P\left(T = 1 \mid \tilde{D} = 1\right) \equiv Se_T,$$

where Sp_T and Se_T denote, respectively, the specificity and sensitivity of the imperfect reference T. A consequence of not observing true disease status $\tilde{D}$ is that estimation of the linear combination coefficients $\boldsymbol{a}$ is not straightforward. Toward this end, we propose a Bayesian latent-class mixture model.

2.2. *Likelihood*

The full-data likelihood $P(\boldsymbol{y}, T, \tilde{D})$ can be decomposed as follows:

$$P\left(\boldsymbol{y}, T, \tilde{D}\right) = P\left(T \mid \boldsymbol{y}, \tilde{D}\right) \times P\left(\boldsymbol{y} \mid \tilde{D}\right) \times P\left(\tilde{D}\right),$$

where $P\left(\boldsymbol{y} \mid \tilde{D}\right)$ is the density function of the K-variate normal distribution given in equation (1); $P\left(\tilde{D}\right)$ is the probability of the latent disease status, with $P\left(\tilde{D} = 1\right) \equiv \theta$, the prevalence of disease; and $P\left(T \mid \boldsymbol{y}, \tilde{D}\right)$ is the probability of the result of the imperfect reference-test T conditional on biomarker values $\boldsymbol{y}$ and $\tilde{D}$. If, conditionally on $\tilde{D}$, $\boldsymbol{y}$ and T are independent, then $P\left(T \mid \boldsymbol{y}, \tilde{D}\right) \equiv P\left(T \mid \tilde{D}\right)$ defined by equation (3).

To allow dependence between the imperfect reference-test T and biomarkers $\boldsymbol{y}$, we propose to model the dependence through a latent continuous tolerance-variable $\tilde{T}$, underlying T. In particular, we assume that T is the result of dichotomizing $\tilde{T}$, with

$$\tilde{T} \mid \tilde{D} = d \sim N\left(\mu_{\tilde{T}_d}, 1\right).$$

Note that, without loss of generality, the variance of the tolerance distribution can be fixed at 1 (see, e.g., Renard et al., 2002).

Consequently, π_0 and π_1 from equation (3) are expressed as $\pi_0 = 1 - \Phi\left(-\mu_{\tilde{T}_0}\right)$ and $\pi_1 = 1 - \Phi\left(-\mu_{\tilde{T}_1}\right)$, where $\mu_{\tilde{T}_d}$ denotes the mean of $\tilde{T}$ for the true disease group d.

By considering the joint-distribution of $\tilde{T}$ and $\boldsymbol{y}$, their correlation can be introduced directly. Assume that, conditionally on $\tilde{D}$, $\tilde{T}$ and $\boldsymbol{y}$ are jointly normally distributed:

$$\begin{pmatrix} \tilde{T} \\ \boldsymbol{y} \end{pmatrix} \Bigg| \; \tilde{D} = d \sim N_{K+1}\left(\begin{pmatrix} \mu_{\tilde{T}_d} \\ \boldsymbol{\mu}_{Y_d} \end{pmatrix}, \overline{\boldsymbol{\Sigma}}_d\right) \tag{4}$$

$$\text{with} \quad \overline{\boldsymbol{\Sigma}}_d = \begin{pmatrix} 1 & \boldsymbol{\tau}^T{}_d \\ \boldsymbol{\tau}_d & \boldsymbol{\Sigma}_d \end{pmatrix} \text{ and } \boldsymbol{\tau}_d = \left(\rho_{1,d}\sigma_{1,d}, \; \ldots, \rho_{K,d}\sigma_{K,d}\right)^T.$$

In equation (4), the covariance of $\tilde{T}$ and the k-th biomarker is expressed as the product of the correlation coefficient $\rho_{k,d}$ and biomarker's standard deviation $\sigma_{k,d}$.

By using the joint normal distribution in equation (4), it follows that

$$T \mid \boldsymbol{y},\ \tilde{D} = d\ \sim Bern\left(\pi_d(\boldsymbol{y})\right), \tag{5}$$

where

$$\begin{aligned} P\left(T=1|\boldsymbol{y},\tilde{D}=0\right) &= \pi_0(\boldsymbol{y}) \\ &= 1-\Phi\left(\frac{-\mu_{\tilde{T}_0}+\boldsymbol{\rho}_0^T\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}_{Y_0}\right)}{\sqrt{1-\boldsymbol{\rho}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\rho}_0}}\right) \\ &= 1-Sp_T(\boldsymbol{y}), \end{aligned}$$

$$\begin{aligned} P\left(T=1|\boldsymbol{y},\tilde{D}=1\right) &= \pi_1(\boldsymbol{y}) \\ &= 1-\Phi\left(\frac{-\mu_{\tilde{T}_1}+\boldsymbol{\rho}_1^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}_{Y_1}\right)}{\sqrt{1-\boldsymbol{\rho}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\rho}_1}}\right) \\ &= Se_T(\boldsymbol{y}). \end{aligned}$$

From equation (5) it follows that the imperfect reference-test T has different sensitivity $Se_T(\boldsymbol{y})$ and specificity $Sp_T(\boldsymbol{y})$ for each possible value $\boldsymbol{y}$, which introduces the dependence between T and $\boldsymbol{y}$ conditional on $\tilde{D}$.

Combining all the developments, we arrive at the following full-data likelihood function for a data set including observations for N individuals (indexed by i):

$$\begin{aligned} &L\left(\boldsymbol{\mu}_{Y_0},\boldsymbol{\mu}_{Y_1},\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1,\mu_{\tilde{T}_0},\mu_{\tilde{T}_1},\boldsymbol{\rho}_0,\boldsymbol{\rho}_1,\theta|\boldsymbol{Y},\boldsymbol{t},\widetilde{\boldsymbol{d}}\right) \\ &= \prod_{i=1}^{N}\left(\theta\left\{1-Se_T\left(\boldsymbol{y}_i\right)\right\}^{(1-t_i)}\left\{Se_T\left(\boldsymbol{y}_i\right)\right\}^{t_i}\frac{1}{\sqrt{(2\pi)^K\left|\boldsymbol{\Sigma}_1\right|}}exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i-\mu_{\boldsymbol{Y}_1}\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y}_i-\mu_{\boldsymbol{Y}_1}\right)\right\}\right)^{d_i} \\ &\quad\times\left((1-\theta)\left\{1-Sp(\boldsymbol{y}_i)\right\}^{t_i}\left\{Sp_T(\boldsymbol{y}_i)\right\}^{(1-t_i)}\frac{1}{\sqrt{(2\pi)^K\left|\boldsymbol{\Sigma}_0\right|}}exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i-\mu_{\boldsymbol{Y}_0}\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{y}_i-\mu_{\boldsymbol{Y}_0}\right)\right\}\right)^{(1-d_i)}, \end{aligned}$$

where $\widetilde{\boldsymbol{d}}=(d_1,\ldots,d_N)'$ and $\boldsymbol{t}=(t_1,\ldots,t_N)'$ are the vectors containing, respectively, the true (unobserved) disease status indicators and observed reference-test results for the N individuals, while $\boldsymbol{Y}$ is the $N\times K$ matrix containing the observed biomarker values. The parameters of interest are $\boldsymbol{\mu}_{Y0}$, $\boldsymbol{\mu}_{Y1}$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_1$, because together they define $AUC_{\boldsymbol{a}}$, as indicated in equation (2).

2.3. *Priors*

Non-identifiability is an important issue for mixture models (McLachlen and Peel, 2004), as well as for the estimation of accuracy of imperfect diagnostic-tests (Dendukuri and Joseph, 2001). We propose prior distributions which include appropriate restrictions that mitigate non-identifiability while allowing introduction of available prior information. Whenever feasible, flat priors are proposed. Table 1 of Supplementary Materials summarizes all proposed prior distributions for the parameters included in the model. Non-standard or non-flat priors are discussed below.

For the prevalence parameter θ, a uniform distribution restricted between $1/N$ and $1-1/N$ is proposed. The restriction assumes that the data contain at least one control and one case. It helps to resolve convergence issues when MCMC algorithms get stuck in a one-component solution instead of a mixture, a possible result of model non-identifiability (Robert and Soubiran, 1993).

To allow a more controlled way of introducing prior biomarker-accuracy information, a prior distribution is proposed for $\boldsymbol{\delta}$, which is defined as follows:

$$\boldsymbol{\delta}=\boldsymbol{Q}\left(\boldsymbol{\mu}_{Y_1}-\boldsymbol{\mu}_{Y_0}\right)\ \text{with}\ \boldsymbol{Q}'\boldsymbol{Q}=\left(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1\right)^{-1},$$

where $\boldsymbol{Q}$ is an upper-triangular matrix. By considering a flat normal prior for $\boldsymbol{\mu}_{Y_0}$ and a multivariate normal prior for $\boldsymbol{\delta}$, it is possible to control the resulting prior for the AUC of the linear combination of the biomarkers. In particular, under the assumption that $\boldsymbol{\delta}\sim N_K(\boldsymbol{\kappa},\boldsymbol{\Psi})$ and $\boldsymbol{\kappa}=\boldsymbol{0}$, different choices for the standard deviations and correlations defining $\boldsymbol{\Psi}$ will lead to $AUC_{\boldsymbol{a}}$-priors expressing different amounts of prior information. Examples of so-constructed $AUC_{\boldsymbol{a}}$-priors are presented in Appendix A of Supplementary Materials. The prior distribution for $\boldsymbol{\mu}_{Y_1}$ is implied by the choice of priors for $\boldsymbol{\mu}_{Y_0}$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\delta}$.

We follow the proposal of Wei and Higgins (2013) to construct flat prior-distributions for the variance–covariance matrices. In particular, the overall $(K+1)\times(K+1)$ variance–covariance matrix $\overline{\boldsymbol{\Sigma}}_d$ is decomposed as $\overline{\boldsymbol{\Sigma}}_d=\boldsymbol{S}_d\boldsymbol{R}_d\boldsymbol{S}_d$, where $\boldsymbol{S}_d$ and $\boldsymbol{R}_{\mathrm{d}}$ are, respectively, the diagonal matrix of standard deviations and the correlation matrix for the disease group d. Additionally, $\boldsymbol{R}_d$ is expressed as $\boldsymbol{R}_d=\boldsymbol{L}_d\boldsymbol{L}_d^T$, where $\boldsymbol{L}_d$ is a lower-triangular matrix. Subsequently, wide uniform-distributions are put directly on the biomarker standard deviations $\sigma_{k,d}$ included in $\boldsymbol{S}_d$. Additionally, flat priors are put on K of the $((K+1)^2-(K+1))/2$ non-zero off-diagonal elements of the Cholesky decomposition-factor $\boldsymbol{L}_d$ of $\boldsymbol{R}_{\mathrm{d}}$ (see Table 1 of Supplementary Materials).

The prior distributions for $\mu_{\tilde{T}_0}$ and $\mu_{\tilde{T}_1}$ are derived from the prior distributions for Sp_T and Se_T. This way, restrictions can be enforced on $\mu_{\tilde{T}_0}$ and $\mu_{\tilde{T}_1}$ leading to a sensible interpretation of Sp_T and Se_T. In the case of case-control data, a sensible choice for Sp_T and Se_T is to use independent Beta-distributions restricted to the (0.5, 1] interval. Based on the relationship defined in equation (3), this leads to the following prior distributions ($\phi(\cdot)$ denotes the standard-normal density

function):

$$Sp_T \sim Beta\,(a, b)\,trunc[0.51, 1]$$

$$f_{\mu_{\tilde{T}_0}}(\mu_{\tilde{T}_0}) = \begin{cases} \frac{1}{B(a,b)}\left(\Phi\left(-\mu_{\tilde{T}_0}\right)\right)^{(a-1)}\left(1-\Phi\left(-\mu_{\tilde{T}_0}\right)\right)^{(b-1)}\left|-\phi\left(-\mu_{\tilde{T}_0}\right)\right| & if\ \mu_{\tilde{T}_0} \in (-\infty, -\Phi\,(0.51)] \\ 0 & otherwise \end{cases}$$

$$Se_T \sim Beta\,(c, d)\,trunc[0.51, 1]$$

$$f_{\mu_{\tilde{T}_1}}(\mu_{\tilde{T}_1}) = \begin{cases} \frac{1}{B(c,d)}\left(\Phi\left(\mu_{\tilde{T}_1}\right)\right)^{(c-1)}\left(1-\Phi\left(\mu_{\tilde{T}_1}\right)\right)^{(d-1)}\left|\phi\left(\mu_{\tilde{T}_1}\right)\right| & if\ \mu_{\tilde{T}_1} \in [\Phi\,(0.51)\,, +\infty) \\ 0 & otherwise \end{cases}$$

Other restrictions on the prior distributions of Sp_T and Se_T could be considered as well. Jones et al. (2010) propose to assume that $Sp_T + Se_T > 1$ to overcome non-identifiability. This restriction allows Sp_T or Se_T to be smaller than 0.5, but implies a dependence between Sp_T and Se_T which is not trivial to implement. For this reason, we limit ourselves to assuming that Sp_T and Se_T are both strictly larger than 0.5. This restriction resolves the label-switching problem observed for mixture models (McLachlen and Peel, 2004) and mitigates the over-parameterization with multiple imperfect reference-tests (Dendukuri and Joseph, 2001), two consequences of model non-identifiability.

2.4. *Application*

To investigate the performance of the model, we carried out a simulation study and applied the model to two AD data sets: the VUmc (VU University Medical Center) data set, which consists of patients from the memory-clinic-based Amsterdam Dementia Cohort (see Figure 1), and the publically available data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) (see Figure 2). In these studies, cerebro-spinal-fluid (CSF) levels of Aβ1-42, Total-tau, and P-tau181p were collected for all individuals, as well as a clinical diagnosis of AD. More detailed information about the data sets can be found in Appendix B of Supplementary Materials.

2.4.1. Simulation study. The goal of the simulation study was to show adequate model performance in the conditional dependence case and, subsequently, to investigate the impact of violating the conditional independence assumption. In particular, 400 data sets of size 600 were simulated under the conditional dependence setting. The underlying parameter settings are summarized in Appendix C of Supplementary Materials. They yielded biomarker data (for $K = 3$ biomarkers) with underlying true AUC_a of 0.787 and an imperfect reference-test with $Se_T = Sp_T = 0.85$. The underlying latent tolerance and biomarkers' correlation coefficients were set as follows: $\rho_{1,0} = \rho_{1,1} = 0$, $\rho_{2,0} = \rho_{2,1} = 0.7$, and $\rho_{3,0} = \rho_{3,1} = 0.3$. An example of a randomly selected simulation data set under the conditional dependence setting is shown in Figure 1 of Supplementary Materials.

2.4.2. Analysis settings. The proposed model allowing for conditional dependence was fitted to the simulated data and to the two case studies. The model assuming conditional independence was fitted to the data as well. The latter model was obtained by fixing the correlation parameters $\rho_{k,d}$ between the latent continuous tolerance-variable $\tilde{T}$ and biomarkers to be equal to zero. Prior distributions were specified as in Table 1 of Supplementary Materials. For the simulation study, parameters a, b, c, and d of the Sp_T and Se_T Beta prior distributions were all set to 1, leading to flat uniform-prior information. For the two case-study data examples, the Beta-prior parameters for Sp_T and Se_T were set so that they allowed capturing the available information from literature. In particular, $Beta(a = 2.69, b = 1.99)$ and $Beta(c = 4.15,\ d = 2.54)$ distributions were used, truncated to the $(0.51, 1]$ interval. The prior for AUC_a was varied to investigate sensitivity of the results to the choice of the prior distribution. In the "optimistic" case, the prior was defined by setting $\boldsymbol{\kappa} = (0, 0, 0)^T$ and defining $\boldsymbol{\Psi}$ by standard deviations equal to 0.7 and correlation coefficients set to 0.6. As a result, a prior distribution disfavoring small values of AUC_a was obtained (see panel a of Figure 3). A "conservative" prior, favoring moderate values of AUC_a (see panel c of Figure 3), was constructed by setting $\boldsymbol{\kappa} = (0, 0, 0)^T$ and defining $\boldsymbol{\Psi}$ by standard deviations equal to 0.5 and correlation coefficients equal to 0.3.

In the simulation study, to compensate for the lack of prior information for Sp_T and Se_T, only the "optimistic" AUC_a prior was used together with a more restrictive $U(0.1, 0.9)$ prior for the prevalence parameter θ.

For both the simulation study and the real-data analyses, 10,000 MCMC samples were retained after a burn-in of 10,000 samples. In both cases, five independent MCMC chains were used. After fitting the models, general diagnostic tools were applied to the results of all 400 data sets as part of the results-analyzing R-script. Convergence over chains was investigated by the Gelman–Rubin convergence index, for which a cut-off value of 1.1 was applied (Gelman and Rubin, 1992). To ensure that at least 3 out of 5 chains converged, chain-by-chain convergence was monitored by using the Geweke convergence criterion (Geweke, 1992).

The models were fitted by using OpenBUGS 3.2.1 (Lunn et al., 2009). The code is presented in Appendix D of Supplementary Materials. Results were analyzed and summarized using R 3.0.1 (x64) (R Core Team, 2013). The R-package R2OpenBUGS (Sturtz, Ligges, and Gelman, 2005) was used as an interface between R and OpenBUGS. For the proposed conditional-dependence model, fitting times were equal to 20 h for a single simulated and the VUmc data set, and 15 h for the ADNI-data, on a 64-bit, 2.8 GHz, 8 GB RAM machine. For the purpose of the manuscript, all simulations and analyses

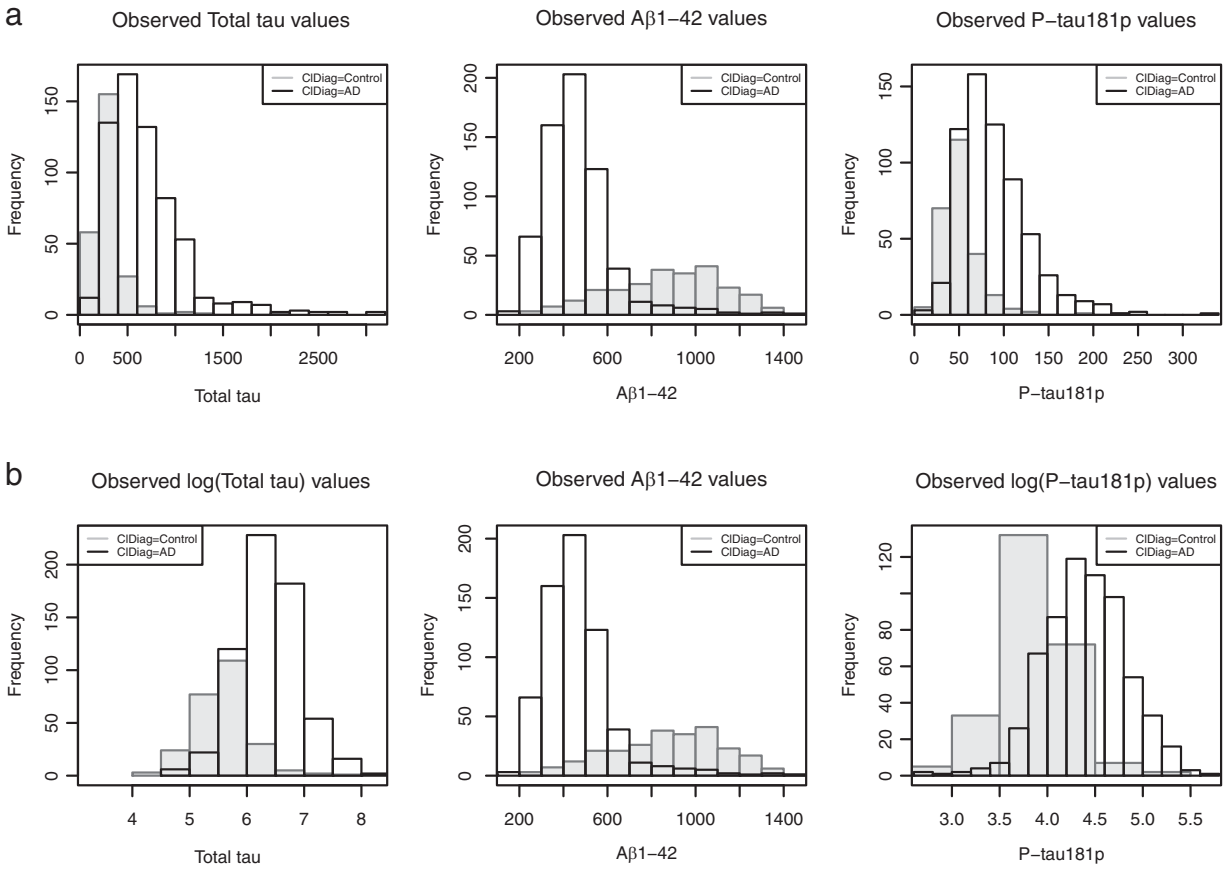

Figure 1. Observed total tau, Aβ1-42, and P-tau181p values from the VUmc data set. Clinically diagnosed controls are indicated by the gray histograms, clinical cases by the black histograms. (a) Raw values. (b) Log-transformed values of Total tau and P-tau181p-values, raw values for Aβ1-42.

were run on the infrastructure provided by the VSC (Flemish Supercomputer Center).

3. Results

The results of the simulation study are summarized in Table 1. For the proposed conditional-dependence model (the third column of the table), the mean of the 400 posterior medians is very close to the true underlying values for all parameters.

The fourth column of Table 1 presents the results for the conditional-independence model. It can be concluded that, on average, AUC_a (true value of 0.787), Sp_T (true value of 0.85), and Se_T (true value of 0.85), are substantially overestimated with mean posterior-medians equal to 0.892, 0.995, and 0.982, respectively.

Table 2 presents the medians of the posterior distributions obtained for the VUmc data, together with their posterior standard deviations and 95%-credible intervals. The medians of the AUC_a distributions are similar irrespectively of the model (conditional-dependence or independence) and the AUC_a-prior ("conservative" or "optimistic"). Hence, in what follows, only the results for the "optimistic" AUC_a-prior are discussed.

Allowing for conditional dependence leads to posterior median Sp_T and Se_T estimates of 0.823 and 0.940, respectively, with respective 95%-credible intervals equal to [0.768;0.870] and [0.915;0.960]. The posterior distribution for θ has a median of 0.706 with a 95%-credible interval of [0.672;0.739]. The posterior distributions, shown in Appendix E of Supplementary Materials, indicate that the assumed prior distributions for Sp_T, Se_T, and θ are well updated by the data, suggesting a successful mitigation of model non-identifiability (Garrett and Zeger, 2000).

Moreover, the results show dependence between the clinical diagnosis of AD and Total-tau: the 95%-credible intervals for the correlation between the latent tolerance and Total-tau in the control ($\rho_{1,0}$) and AD ($\rho_{1,1}$) groups are equal to [0.156;0.530] and [0.068;0.451], respectively, and they both exclude the value of zero. Despite the correlation between the latent tolerance and Total-tau, no important difference in the posterior medians of AUC_a is found for the conditional-dependence and conditional-independence models.

Table 3 presents the results for the ADNI data. In this case, the resulting MCMC-samples for the conditional-dependence model defined by using the "conservative" AUC_a-prior require some attention. Even after 200,000 iterations, the OpenBUGS MCMC-algorithms do not seem to have converged (Appendix F of Supplementary Materials). This may be taken as implying that the use of the "conservative" prior for AUC_a does not provide enough information to overcome potential non-identifiability of the model given the limited size of the data

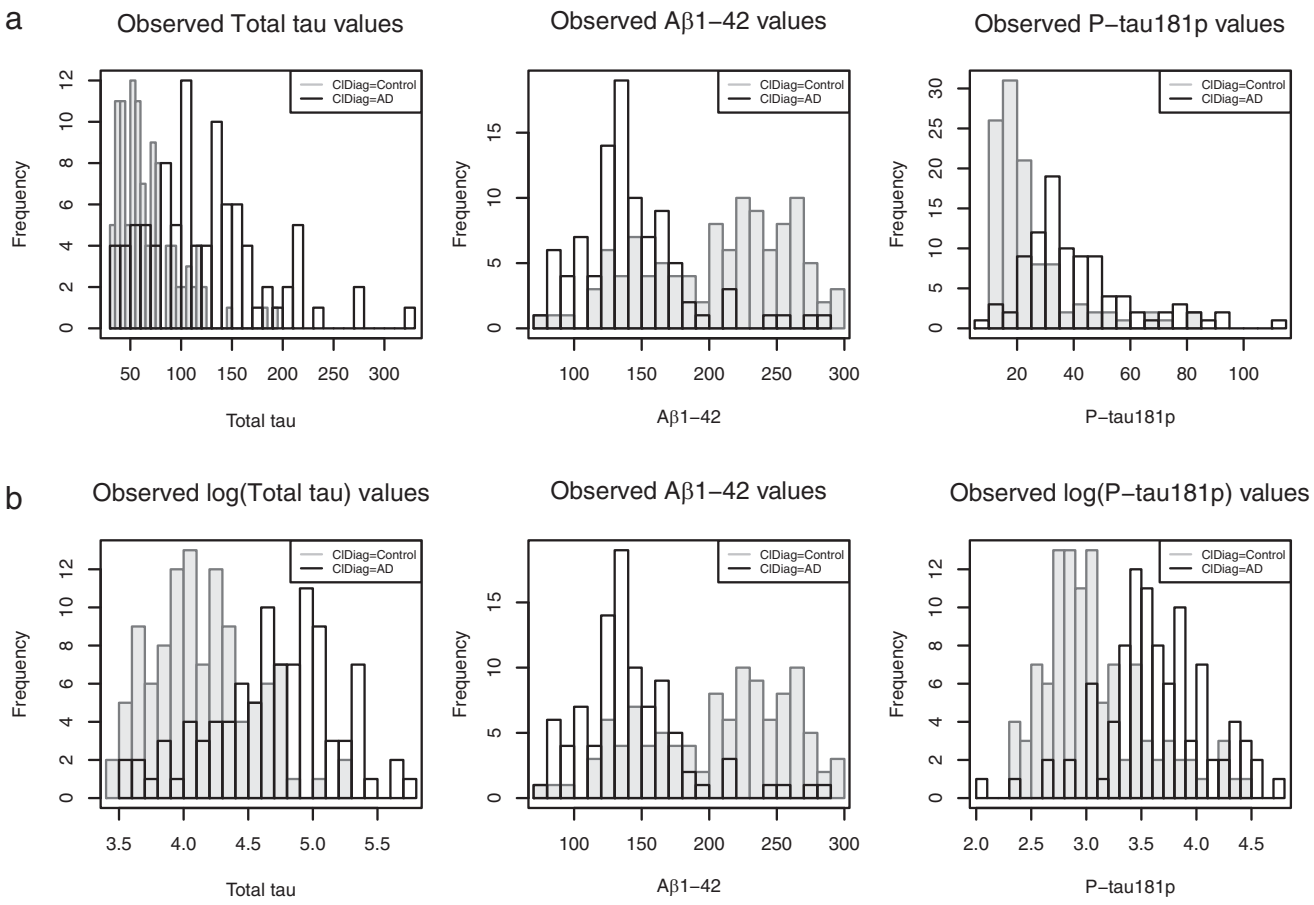

Figure 2. Observed total tau, Aβ1-42, and P-tau181p values from the ADNI data set. Clinically diagnosed controls are indicated by the gray histograms, clinical cases by the black histograms. a. Raw values. (b) Log-transformed values of Total tau and P-tau181p-values, raw values for Aβ1-42.

set. For this reason, the data were also analyzed with a model using an "intermediate" AUC_a-prior (see Panel *b* of Figure 3). Table 3 contains the results obtained with the "optimistic," as well as the "intermediate" AUC_a-prior distribution.

No difference between the results obtained with the "intermediate" and "optimistic" AUC_a-prior were observed as the 95% credible intervals show substantial overlap. Therefore, only the results from the "optimistic" AUC_a-prior setting will be discussed. The results obtained for the proposed conditional-dependence model do not indicate any correlation between the biomarkers and the latent tolerance underlying the AD diagnosis. The posterior AUC_a distributions for the

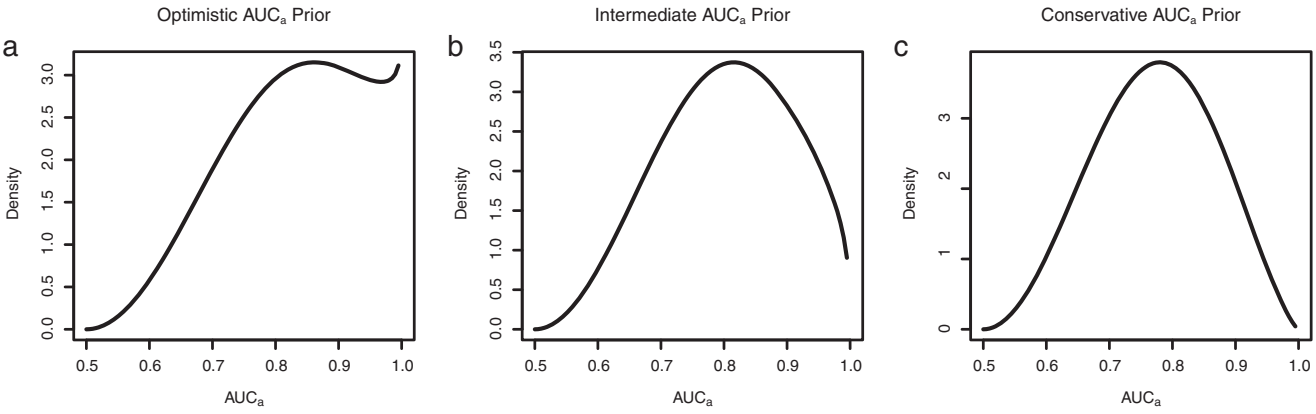

Figure 3. Considered AUC_a-prior distributions. (a) "Optimistic" AUC_a-prior distribution. (b) "Intermediate" AUC_a-prior distribution. (c) "Conservative" AUC_a-prior distribution.

Table 1

Mean of posterior medians, (standard deviations) and [empirical 95%-confidence intervals] for the simulated data (400 data sets of size N=600). Results are shown for both the conditional-dependence and conditional-independence model considering the "optimistic" AUC_a-prior distribution. Correlation coefficients: $\rho_{1,0}$, Total-tau in the control group; $\rho_{1,1}$, Total-tau in the AD group; $\rho_{2,0}$, Aβ1-42 in the control group; $\rho_{2,1}$, Aβ1-42 in the AD group; $\rho_{3,0}$, P-tau181p in the control group; $\rho_{3,1}$, P-tau181p in the AD group.

		Model	
Parameter	True	Conditional Dep.	Conditional Ind.
AUC_a	0.787	0.777 (0.023) [0.728;0.819]	0.892 (0.011) [0.869; 0.912]
Se_T	0.85	0.834 (0.031) [0.770;0.894]	0.982 (0.013) [0.940; 0.993]
Sp_T	0.85	0.837 (0.027) [0.782;0.887]	0.995 (0.001) [0.993; 0.996]
θ	0.5	0.505 (0.023) [0.458;0.554]	0.510 (0.017) [0.479; 0.542]
$\rho_{0,1}$	0	0.022 (0.090) [-0.146;0.213]	0
$\rho_{0,2}$	0.7	0.702 (0.046) [0.608;0.782]	0
$\rho_{0,3}$	0.3	0.318 (0.079) [0.165;0.470]	0
$\rho_{1,1}$	0	0.020 (0.088) [-0.147;0.197]	0
$\rho_{1,2}$	0.7	0.705 (0.052) [0.596;0.795]	0
$\rho_{1,3}$	0.3	0.310 (0.078) [0.158;0.457]	0

conditional-dependence and conditional-independence models show only a slight difference. In particular, the posterior medians of AUC_a are equal to 0.979 and 0.983, respectively, with the respective 95% credible intervals of [0.939;0.994] and [0.961;0.994].

To further investigate applicability of the model to the case of data structures similar to the ADNI data, an additional simulation setting was considered (Appendix G of Supplementary Materials). In particular, the underlying true parameter values were chosen so that a bi-modal distribution would be observed, as is the case for the ADNI data (Figure 2). Results from this additional simulation study show that also for bi-modal observed data, the conditional-dependence model provides correct estimates of the true parameter values.

Table 2

Posterior medians, (standard deviations) and [95%-credible intervals] for VUmc data. In the respective columns results are shown for the conditional-dependence and conditional-independence models for both the 'conservative' (Cons.) and 'optimistic' (Opt.) AUC_a-prior distribution. Correlation coefficients: $\rho_{1,0}$, Total-tau in the control group; $\rho_{1,1}$, Total-tau in the AD group; $\rho_{2,0}$, Aβ1-42 in the control group; $\rho_{2,1}$, Aβ1-42 in the AD group; $\rho_{3,0}$, P-tau181p in the control group; $\rho_{3,1}$, P-tau181p in the AD group.

Parameter	Conditional Dep. (Cons. AUC)	Conditional Dep. (Opt. AUC)	Conditional Ind. (Cons. AUC)	Conditional Ind. (Opt. AUC)
AUC_a	0.996 (0.002) [0.992;0.998]	0.997 (0.001) [0.993;0.998]	0.995 (0.002) [0.990;0.997]	0.995 (0.002) [0.991;0.998]
Se_T	0.940 (0.012) [0.915;0.960]	0.940 (0.012) [0.915;0.960]	0.955 (0.010) [0.934;0.973]	0.954 (0.010) [0.934;0.971]
Sp_T	0.818 (0.027) [0.762;0.868]	0.823 (0.026) [0.768;0.870]	0.850 (0.024) [0.799;0.894]	0.850 (0.024) [0.799;0.894]
θ	0.705 (0.017) [0.671;0.738]	0.706 (0.017) [0.672;0.739]	0.700 (0.016) [0.668;0.732]	0.701 (0.016) [0.668;0.732]
$\rho_{1,0}$	0.366 (0.091) [0.176;0.520]	0.358 (0.093) [0.156;0.530]	0	0
$\rho_{2,0}$	0.110 (0.103) [−0.100; 0.306]	0.104 (0.103) [−0.103; 0.301]	0	0
$\rho_{3,0}$	0.034 (0.095) [−0.154; 0.206]	0.030 (0.100) [−0.171; 0.235]	0	0
$\rho_{1,1}$	0.293 (0.091) [0.102;0.454]	0.280 (0.097) [0.068;0.451]	0	0
$\rho_{2,1}$	0.186 (0.093) [−0.009; 0.353]	0.186 (0.092) [−0.007; 0.356]	0	0
$\rho_{3,1}$	0.199 (0.091) [0.014;0.366]	0.186 (0.098) [−0.018; 0.366]	0	0

Table 3

Posterior medians, (standard deviations) and [95%-credible intervals] for ADNI data. Results are shown for both the conditional-dependence and conditional-independence models using the 'intermediate' (Int.) and 'optimistic' (Opt.) AUC_a-prior distribution. Correlation coefficients: $\rho_{1,0}$, Total-tau in the control group; $\rho_{1,1}$, Total-tau in the AD group; $\rho_{2,0}$, Aβ1-42 in the control group; $\rho_{2,1}$, Aβ1-42 in the AD group; $\rho_{3,0}$, P-tau181p in the control group; $\rho_{3,1}$, P-tau181p in the AD group.

Parameter	Conditional Dep. (Int. AUC)	Conditional Dep. (Opt. AUC)	Conditional Ind. (Int. AUC)	Conditional Ind. (Opt. AUC)
AUC_a	0.976 (0.022) [0.912;0.994]	0.979 (0.014) [0.939;0.994]	0.982 (0.009) [0.958;0.993]	0.983 (0.009) [0.961;0.994]
Se_T	0.805 (0.061) [0.669;0.906]	0.808 (0.056) [0.688;0.905]	0.818 (0.044) [0.726;0.898]	0.818 (0.044) [0.724;0.896]
Sp_T	0.829 (0.077) [0.627;0.940]	0.835 (0.061) [0.691;0.930]	0.880 (0.037) [0.798;0.942]	0.879 (0.037) [0.798;0.941]
θ	0.463 (0.091) [0.230;0.614]	0.466 (0.071) [0.317;0.597]	0.499 (0.043) [0.413;0.582]	0.498 (0.043) [0.414;0.582]
$\rho_{1,0}$	0.142 (0.225) $[-0.297; 0.549]$	0.121 (0.185) $[-0.264; 0.459]$	0	0
$\rho_{2,0}$	0.295 (0.250) $[-0.323; 0.637]$	0.277 (0.223) $[-0.288; 0.584]$	0	0
$\rho_{3,0}$	0.186 (0.222) $[-0.257; 0.560]$	0.161 (0.183) $[-0.226; 0.494]$	0	0
$\rho_{1,1}$	0.322 (0.182) $[-0.073; 0.621]$	0.314 (0.177) $[-0.073; 0.610]$	0	0
$\rho_{2,1}$	0.053 (0.195) $[-0.352; 0.414]$	0.048 (0.182) $[-0.318; 0.394]$	0	0
$\rho_{3,1}$	−0.079(0.273) $[-0.580; 0.468]$	−0.083(0.252) $[-0.543; 0.412]$	0	0

4. Discussion

We have proposed a Bayesian latent-class model for construction of a diagnostic biomarker-index in the presence of a dichotomous imperfect reference-test. Importantly, the model does not require the conditional-independence assumption because it explicitly allows for a correlation between the results of the reference test and biomarkers.

The simulation study results showed adequate model performance leading to unbiased estimates of the model parameters. Moreover, the results showed that falsely assuming conditional independence may lead to substantial bias in the estimates of biomarker-index accuracy. The observed overestimation of AUC_a may be related to the substantial overestimation of Sp_T and Se_T. As suggested by one of the reviewers, this can be explained by the model trying to capture the excess correlation of the conditional dependence by increasing the imperfect reference-test sensitivity estimate. Since the marginal positive rate for the reference test is fixed together with a stable disease prevalence estimate, specificity is overestimated as well. A similar observation was made for a dichotomous test by Pepe and Janes (2007). A supplementary simulation study showed that the proposed model is able to accommodate bi-modal data, similar to those observed in the ADNI data set.

For the VUmc data set, the chosen prior distributions for Sp_T, Se_T, θ, and AUC_a addressed the model non-identifiability issues. The form of the prior distribution for AUC_a did not affect the results. Interestingly, the model suggested dependence between clinical diagnosis and Total-tau. Despite this, there was not much difference in the posterior-median AUC obtained under conditional independence or dependence. A possible explanation could be that the bias in AUC_a is too small to be picked up by the model (Baker et al., 2014). Another explanation could be that the importance of Total-tau in the linear combination comprising the diagnostic index is only limited (McKhann et al., 2011). In fact, the posterior median for coefficient $a_{Total-tau}$ was equal to 9.43 and was smaller as compared to the medians for the two other biomarkers ($\hat{a}_{A\beta 1-42} = 15.04$; $\hat{a}_{P-tau181p} = 21.92$). It is also possible that the lack of difference for the conditional-dependence analyses is due to a misspecification of the correlation structure in the former. In fact, Albert and Dodd (2004) showed that estimates of the diagnostic-performance measures may be biased if the conditional-dependence structure is misspecified and it may be difficult to distinguish between different dependence structures based on the observed data. However, these conclusions were formulated for the case of multiple dichotomous tests and a frequentist analysis. Bayesian approaches based on good prior information may be less sensitive to this issue, as also indicated by Albert and Dodd (2004).

For the ADNI data, convergence issues due to non-identifiability were noted for the "conservative" AUC_a-prior distribution. They were resolved when the "intermediate" or "optimistic" AUC_a-priors were used. This indicates that the use of the model may require a substantial sample size or, otherwise, a substantial amount of prior information to provide reliable results.

Although this was not anticipated, the underlying settings of the simulations seem to mimic the results from the real-data applications. Nevertheless, the results of the simulation study

and real-data applications are different. The most important difference is the overall accuracy of the combination which is much better for the real-data applications. Moreover, the substantial conditional dependence included in the simulation study is affecting the estimates of AUC_a, while this is not the case in the real-data applications.

The proposed model explicitly assumes that biomarkers are normally distributed in order to get a biomarker combination which maximizes the AUC of the combination. The real-data application shows that applying suitable transformation (e.g., a logarithmic one) can help in obtaining data that are compliant with the assumption. Another solution would be to include a Box–Cox transformation directly into the likelihood, as proposed by, e.g., O'Malley and Zou (2006). Extending the model along these lines remains a topic for further research.

To check whether valid inference could be obtained even with a minimum of prior information, in the presented analyses we used as uninformative prior distributions as possible. Of course, as demonstrated in the real-data application, if scientifically accepted prior information is available, it can be used. While we have restricted ourselves to include information on Sp_T, Se_T, and θ, informative priors for other parameters could be added as well if the application at hand allows it. For example, in case of the particular application in AD, Welge et al. (2009) present AUC estimates of the linear combination of the three considered biomarkers.

Another extension of the model could be to allow inclusion of continuous imperfect reference-test information. If such information is available, one could think of summarizing its accuracy information in terms of AUC and including this information as a prior, using similar parameterization as expressed in the Methods section.

5. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 3, and 4 are available with this article at the *Biometrics* website on Wiley Online Library. OpenBUGS model-code for the implementation of the proposed models is also available with this article at the Biometrics website on Wiley Online Library.

Acknowledgements

This Research has been conducted with the financial support of the Walloon Government under the European ERA-NET EUROTRANS-BIO framework (Project B4AD, Agreement no. 1017106) It was conducted in collaboration with 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).

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.

The computational resources and services used in the current work were provided by the Hercules Foundation and the Flemish Government—department EWI.

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

References

Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60**, 427–435.

Baker, S. G., Schuit, E., Steyerberg, E. W., Pencina, M. J., Vickers, A., Moons, K. G. M., et al. (2014). How to interpret a small increase in AUC with an additional risk prediction marker: Decision analysis comes through. *Statistics in Medicine* **22**, 3946–3959.

Beach, T. G., Monsell, S. E., Philips, L. E., and Kukull, W. (2012). Accuracy of the clinical diagnosis of Alzheimer diseases at the national institute on aging Alzheimer disease centers, 2005–2010. *Journal of Neuropathology & Experimental Neurology* **71**, 566–573.

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* **69**, 89–95.

Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57**, 158–167.

Garret, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics* **56**, 1055–1067.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. In *Bayesian Statistics 4*, J. M. Bernardo, J. Berger, A. P. Dawid, A. F. M. Smith (eds), 169–193. Oxford UK: Clarendon Press.

Huang, X., Qin, G., and Fang, Y. (2011). Optimal combinations of diagnostic tests based on AUC. *Biometrics* **67**, 568–576.

Jones, G., Johnson, W. O., Hanson, T. E., and Christensen, R. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics* **66**, 855–863.

Lu, Y., Dendukuri, N., Schiller, I., and Joseph, L. (2010). A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Statistics in Medicine* **29**, 2532–2543.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine* **28**, 3049–3067.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R. Jr.,, 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. *Alzheimer's & Dementia* **7**, 263–269.

McLachlen, G., and Peel, D. (2004). *Finite Mixture Models*. Toronto, Canada: Wiley Inc

O'Malley, A. G., Zou, K. H., Fielding, J. R., and Tempany, C. M. C. (2001). Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: Prostate biopsy and spiral ct of ureteral stones. *Academic Radiology* **8**, 731–725.

O'Malley, A. J. and Zou, H. K. (2006). Bayesian multivariate hierarchical transformation models for ROC analysis. *Statistics in Medicine* **25**, 459–479.

Pepe, M. S. and Janes, H. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8**, 474–484.

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

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical journal* **44**, 921–935.

Rindskopf, D. and Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine* **5**, 21–27.

Robert, C. P. and Soubiran, C. (1993). Estimation of a normal mixture model through Gibbs sampling and prior feedback. *Test* **2**, 125–146.

Scheltens P. and Rockwood K. (2011). How golden is the gold standard of neuropathology in dementia. *Alzheimer's & Dementia* **7**, 486–489.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Packages for Running WinBUGS from R. *Journal of Statistical Software* **12**, 1–16.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.

Wang, C., Turnbull, B. W., Gröhn, Y. T., and Nielsen, S. S. (2006). Estimating receiver operating characteristic curves with covariates when there is no perfect reference test for diagnosis of Johne's Disease. *Journal of Diary Science* **89**, 3038–3046.

Welge, V., Fiege, O., Lewczuk, P., Mollenhauer, B., Esselmann, H., Klafki, H.-W., et al. (2009). Combined CSF, tau, p-tau181 and amyloid-β 38/40/42 for diagnosing Alzheimer's disease. *Journal of Neural Transmission* **116**, 203–212.

Wei, Y. and Higgins, P. T. (2013). Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* **32**, 2911–2934.

Wollman, D. E. and Prohovnik, I. (2003). Sensitivity and specificity of neuroimaging for the diagnosis of Alzheimer's disease. *Dialogues in Clinical Neuroscience* **5**, 89–99.

Yu, B., Zhou, C., and Bandinelli, S. (2011). Combining multiple continuous test for the diagnosis of kidney impairment in the absence of a gold standard. *Statistics in Medicine* **30**, 1712–1721.

Zou, K. H., Liu, A., Bandos, A. I., Ohno-Machado, L., and Rockette, H. E. (2012). *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Boca Raton: Chapman & Hall.

Received April 2015. *Revised June* 2016. *Accepted July* 2016.