HHS Public Access

Author manuscript

Comput Stat Data Anal. Author manuscript; available in PMC 2018 September 01.

Published in final edited form as:

Comput Stat Data Anal. 2017 September ; 113: 125–135. doi:10.1016/j.csda.2016.10.026.

Model-based clustering for assessing the prognostic value of imaging biomarkers and mixed type tests

Zheyu Wang[a,*], **Krisztian Sebestyen**[a], and **Sarah E. Monsell**[b]

[a]Johns Hopkins University, Baltimore, MD, USA

[b]University of Washington, Seattle, WA, USA

Abstract

A model-based clustering method is proposed to address two research aims in Alzheimer's disease (AD): to evaluate the accuracy of imaging biomarkers in AD prognosis, and to integrate biomarker information and standard clinical test results into the diagnoses. One challenge in such biomarker studies is that it is often desired or necessary to conduct the evaluation without relying on clinical diagnoses or some other standard references. This is because (1) biomarkers may provide prognostic information long before any standard reference can be acquired; (2) these references are often based on or provide unfair advantage to standard tests. Therefore, they can mask the prognostic value of a useful biomarker, especially when the biomarker is much more accurate than the standard tests. In addition, the biomarkers and existing tests may be of mixed type and vastly different distributions. A model-based clustering method based on finite mixture modeling framework is introduced. The model allows for the inclusion of mixed typed manifest variables with possible differential covariates to evaluate the prognostic value of biomarkers in addition to standard tests without relying on potentially inaccurate reference diagnoses. Maximum likelihood parameter estimation is carried out via the EM algorithm. Accuracy measures and the ROC curves of the biomarkers are derived subsequently. Finally, the method is illustrated with a real example in AD.

Keywords

Finite mixture; Latent variable model; Diagnostic tests; Biomarkers; Differential covariate effect; Imperfect gold standard

1. Introduction

Model-based clustering methods provide a means to study constructs that are not directly measurable (Goodman, 1974; Fraley and Raftery, 2002; McLachlan and Peel, 2000), and have significant applications in medical diagnostic studies (Hui and Zhou, 1998; Pepe and Janes, 2007; van Smeden et al., 2013; Collins and Huynh, 2014). Such studies aim to evaluate the accuracy of applying a particular test or procedure for disease diagnosis or prognosis. The difficulty is that disease is a complex process; one may not have a perfect

*Corresponding author. wangzy@jhu.edu (Z. Wang).

reference standard (the gold standard) to base upon the evaluation. On the other hand, using an imperfect reference standard can cause biased assessments (Zhou et al., 2009). In particular, standard references only represent our current understanding about the underlying disease process and the technique to uncover it. New discovery and technology advancement can introduce biomarkers that have higher accuracy than current tests. Nonetheless, because standard references are often developed based on existing tests, performing biomarker evaluation against the reference can give an unfair advantage to the existing tests and mask the value of a new marker. In this situation, clustering methods provide perhaps the most realistic view that reflects the unobservable nature of the true disease status, while adopting information from a reference standard to help differentiate diseased and healthy subjects. Many clustering methods use all observed variables indistinguishably. However, in medical biomarker studies, one often has a basic understanding of which variables are risk factors, which are biomarkers and which are covariates that may affect biomarker levels. In order to adopt this knowledge, finite mixture models are often used. These models have a general form as follows,

$$f(\vec{t}_i)=\sum_{d=0}^{L-1}\pi_{di}\gamma_d(\vec{t}_i), \quad (1)$$

where $\pi_{di} = P(D_i = d)$ is the probability of group membership for the ith individual, with $d = 0, \ldots, L-1$ denoting the L possible group memberships, such as diseased group ($D = 1$), healthy group ($D = 0$) or groups of different disease severity levels or subtypes;

$\gamma_d=P(\vec{t}_i|D_i=d)$ is the conditional distribution of $\vec{T}_i$ at value $\vec{t}_i$, with $\vec{T}_i=(T_{i1},\ldots,T_{iK})$ being K manifest variables that reflect underlying group membership, such as new diagnostic biomarkers, existing reference standards or other information that help reveal the underlying disease status.

In many applications of the clustering methods, the focus is on π_{di}, i.e., to make correct classification, or to infer the relationship between an individual's group membership and other factors $\boldsymbol{Z}$. In the latter case, π_{di} becomes a function of $\boldsymbol{Z}_i$, $\pi_{di}(\boldsymbol{Z}_i)$. The structure and modeling assumption on $\gamma_d(\vec{t}_i)$ are not of high priority and may be chosen for computational convenience as long as the model produces robust results related to group membership D. For example, conditional independence between T_k's within the same group or strong parametric assumptions are often assumed (Vermunt and Magidson, 2002). By contrast, applications in medical diagnostic studies are interested in the exact form of $\gamma_d(\vec{t}_i)$, i.e., the relationship between the manifest test results and the underlying disease status. In such cases one may need to interpret the results in terms of diagnostic accuracy measures, such as sensitivity of a new biomarker T_k at cut-off value t, $P(T_k \geq t|D = 1)$ and the corresponding specificity $P(T_k < t/D = 0)$. Many models were proposed to address these needs in diagnostic studies (Henkelman et al., 1990; Branscum et al., 2005; Zheng et al., 2005; Branscum et al., 2015). Fixed effect and random effect models were adopted to relax the conditional independence assumption (Qu et al., 1996; Xu and Craig, 2009; Wang and Zhou, 2012). Different longitudinal models were proposed for repeated test measurements

(Cook et al., 2000; Jones et al., 2012; Wu et al., 2016). And Albert and Dodd (2004) investigated the influence on parameter estimates introduced by different modeling assumption for $\gamma_d(\vec{t}_i)$.

The increasing availability and fast discovery of imaging and other biomarkers further promoted the application of a model-based clustering method in diagnostic studies. This is because the lack of gold standard is particularly prevalent in biomarker evaluation where the goal of the biomarker is often to facilitate early detection, to reduce cost, and/or to improve diagnostic accuracy, and the current diagnostic standard may not be able to provide conclusive or accurate evidence sufficiently quickly and may be too costly to conduct for all at risk individuals. For example, in the motivating project for this work, the goal is to assess the diagnostic ability of imaging biomarkers for early Alzheimer's disease (AD) detection. Due to the long preclinical phase of AD, current diagnosis procedures, which are based on clinical symptoms and neuropsychological tests, detect abnormality often decades after initial AD pathology starts which is, unfortunately, too late for existing treatments to be effective. Imaging biomarkers have strong diagnostic ability and, as such, have been increasingly used in conjunction with other clinical tests for early AD detection (Nestor et al., 2004; Dubois et al., 2007; Sperling et al., 2011). However, their diagnostic performance has only been assessed against the imperfect clinical diagnosis because a definite AD diagnosis requires a brain autopsy which can only be done upon a patient's consent and death and is also a costly procedure. In addition, because diagnostic studies require careful handling of $\gamma_d(\vec{t}_i)$ as explained previously, features introduced by biomarker studies compared to traditional diagnostic test studies need to be taken into account, including: (1) biomarker values can be highly influenced by subject characteristics. Instead of using a random effect to explain the correlation among biomarkers within a disease group, it is often of interest to know how covariates $\boldsymbol{X}$ affect biomarkers' diagnostic performance so that a more personalized diagnosis can be given. One solution is to model $\gamma_d(\vec{t}_i)$ explicitly as a function of covariate $\gamma_d(\vec{t}_i|\boldsymbol{X}_i)$; (2) multiple biomarkers may be examined and used together with other tests, which may have vastly different distributions. In particular, existing tests are usually discrete while biomarkers are often continuous and skewed. This requires the model $\gamma_d(\vec{t}_i|\boldsymbol{X}_i)$ to handle mixed type t_i's; (3) in contrast to traditional diagnostic tests, biomarkers are often proposed to be used together with, rather than to replace, existing procedures. Therefore, in addition to evaluate the diagnostic accuracy of the biomarkers alone, it is also interesting to assess the diagnostic incremental value obtained by introducing new biomarkers. In this paper, we propose an extension to current model-based clustering methods to address these common issues in biomarker diagnostic studies.

2. Methodology

2.1. Model specification

Following similar notation used above, let T_k, $k = 1, \ldots, K$ be K manifest variables, including biomarkers and tests. For the rest of the paper, we do not distinguish between biomarkers and tests, and refer to all of them as tests. Without loss of generality, we assume

that the first k_1 tests are continuous, and the remaining $K - k_1$ tests are categorical with values 0, …, J_k. Let $D = 0, \ldots, L - 1$ be the true disease status, X_β, X_γ and Z be the covariates related to continuous tests, categorical tests and disease prevalence with length p, q and r, respectively. These three sets of covariates can be overlapping or mutually exclusive. We assume that the tests are conditionally independent given true disease status and covariates, $T_k \perp T_j \mid D, X_\beta, X_\gamma, k, j = 1, \ldots, K$. In other words, we assume the dependence among tests are due to disease status and other covariates. The conditional distribution $\gamma_d(\vec{t}_i)$ can then be modeled by its univariate marginals $\gamma_{dk}(t_{ki})$. We specify the model in the finite mixture form (1) but with covariates, i.e.,

$$f(t_i|X_{\beta i}, X_{\gamma i}, Z_i) = \sum_{d=0}^{L-1} \pi_{di}(Z_i) \prod_{k=1}^{K} \gamma_{dk}(t_{ki}|X_{\beta i}, X_{\gamma i}). \tag{2}$$

Disease group membership (or prevalence) is modeled by a multinomial regression model to allow for ordinal or nominal disease status and their dependence on risk factors as follows:

$$\begin{aligned} \eta_Z[\pi_{di}(Z_i)] = log\, \tfrac{\pi_{di}(Z_i)}{\pi_{0i}(Z_i)} = Z_i^T \alpha_d = \alpha_{d0} + \alpha_{d1} Z_{i1} + \cdots + \alpha_{dr} Z_{ir}, \quad d = 1, \ldots, L-1 \\ \pi_{di}(z) = P(D_i = d | Z_i = z) = \eta_Z^{-1}(z^T \alpha_d) = \frac{exp(z^T \alpha_d)}{\sum_{l=0}^{L-1} exp(z^T \alpha_l)} \end{aligned} \tag{3}$$

where $D = 0$ is the baseline group with parameters $\alpha_0 = (\alpha_{00}, \ldots, \alpha_{0p}) = (0, \ldots, 0)$. Transformation regression models and cumulative link models are used for continuous and for categorical tests, respectively as follows:

$$\begin{aligned} H_k(T_{ki}, \lambda_k) = X_{\beta i}^T \beta_{kd} + \varepsilon_{ik} = \beta_{kd0} \chi_{\beta i0} + \beta_{kd1} \chi_{\beta i1} + \cdots + \beta_{kdp} \chi_{\beta ip} + \varepsilon_{ik}, \\ \varepsilon_{ik} \sim^{i.i.d.} N(0, \sigma_k^2), \quad k = 1, \ldots, k_1, \end{aligned} \tag{4}$$

$$\begin{aligned} \eta_X(P(T_{ki} \le j | D_i = d, X_\gamma)) = \text{logit}(P(T_{ki} \le j | D_i = d, X_\gamma)) = X_{\gamma i}^T \gamma_{kdj} \\ = \gamma_{kdj0} + \gamma_{kd1} \chi_{\gamma i1} + \cdots + \gamma_{kdq} \chi_{\gamma iq}, \quad k = k_1 + 1, \ldots, K, j = 1, \ldots, J_k, \end{aligned} \tag{5}$$

where $H_k(\cdot, \lambda_k)$ is parametric function with unknown parameters λ_k. These transformations allow tests to have more flexible distributions. In the remaining derivation for this paper, we assume the Box–Cox transformation. However, any monotonic function can be used. Regression models for $\gamma_{dk}(t_k|X_\beta, X_\gamma)$ take into account the extra variation and dependence among the tests, and allow for quantitative examination of test accuracy in different subgroups to facilitate personalized diagnosis. Note that regression coefficients β_{kd} and γ_{kdj} are allowed to vary with test T_k and disease status D. This permits an interaction effect between X and D, or differential covariate effect on test performance. However, one can also introduce constraints on these coefficients, such as assuming constant covariate effect across

disease group. Rather than having to perform a constraint maximization, the estimation procedure described in the next section can be easily adapted in such a situation and proceed as an unconstrained estimation.

Based on Eq. (4), we have

$\gamma_{dk}(t|X_{\beta i}, X_{\gamma i}) = P(T_{ki}=t|D_i=d, X_{\beta i},) = \phi(\frac{H_k(t) - X_{\beta i}^T \beta_{kd}}{\sigma_k}) detJ(T_{ki}, \lambda_k)$, $K = 1, \ldots, k_1$, where ϕ is the standard normal density function, and det $J(T_{ki}, \lambda_k)$ is the Jacobian determinate of $H_k(T_{ki}, \lambda_k)$, det $J = \exp((\lambda_k - 1) \log T_{ki})$. Additionally, Eq. (5) suggests $\gamma_{dk}(j|X_{\beta i}, X_{\gamma i}) = P(T_{ki}=j|D_i=d, X_{\gamma i}) = \eta_X^{-1}(X_{\gamma i}^T \gamma_{kdj}) - \eta_X^{-1}(X_{\gamma i}^T \gamma_{kd(j-1)})$, $K = k_1, + 1, \ldots, K$, $j = 1, \ldots, J_k$. Combining the above with Eq. (2), the mixture model is,

$$f(\theta) = \sum_{d=0}^{L-1} \left\{ \eta_Z^{-1}(Z_i^T \alpha_d) \prod_{k=1}^{k_1} detJ(T_{ki}, \lambda_k) \phi \left(\frac{H_k(T_{ki}) - X_{\beta i}^T \beta_{kd}}{\sigma_k} \right) \prod_{k=k_1+1}^{K} [\eta_X^{-1}(X_{\gamma i}^T \gamma_{kdj_k}) - \eta_X^{-1}(X_{\gamma i}^T \gamma_{kd(j_k-1)})] \right\}, \quad (6)$$

where $\theta = \{\alpha_d, \beta_{kd}, \gamma_{kdj_k}, \lambda_k, \sigma_k \mid d = 0, \ldots, L-1; k = 1, \ldots, K; j = 0, \ldots, J_k\}$.

2.2. Estimation via the EM algorithm

The maximum likelihood estimates of parameter θ can be obtained by directly maximizing the likelihood function of model (6). However, the EM algorithm provides a more convenient way of estimating the MLE of parameters in a finite mixture model. Specifically, by considering the true disease status D as missing value, the complete data log likelihood based on model (6) is,

$$\begin{aligned} l_c(\theta) = & \sum_{i=1}^{N} \sum_{d=0}^{L-1} I(D_i=d) log\, P(D_i=d|Z_i) \\ & + \sum_{i=1}^{N} \sum_{d=0}^{L-1} \sum_{k=1}^{k_1} I(D_i=d) log\, P(H_k(T_{ki}, \lambda_k)|D_i=d, X_{\beta i}) + \sum_{i=1}^{N} \sum_{k=1}^{k_1} (\lambda_k - 1) log\, T_{ki} \\ & + \sum_{i=1}^{N} \sum_{d=0}^{L-1} \sum_{k=k_1+1}^{K} \sum_{j=1}^{J_k} I(D_i=d) I(T_{ki}=j) log\, P(T_{ki}=j|D_i=d, X_{\gamma i}), \end{aligned} \quad (7)$$

where $I(\cdot)$ is an indicator function that equals 1 if true and 0 otherwise.

Let $\theta^{(t)}$ denote the parameter estimate updated from the t^{th} EM iteration, with initial value $\theta^{(0)}$. The following procedure describes the t^{th} iteration.

The E step computes the expected value of $I(D_i = d)$ based on parameter estimates from the previous step, $P(d) = E[I(D_i = d)|\theta^{(t-1)}, T, X_\beta, X_\gamma, Z]$:

$$P^{(t)}(d)=P^{(t)}(D_i=d|\theta^{(t-1)},T_i,X_{\beta i},X_{\gamma i},Z_i)=\frac{P(T_i|D_i=d,X_{\beta i},X_{\gamma i},)P(D_i=d|Z_i)}{\sum_{d=0}^{L-1}P(T_i|D_i=d,X_{\beta i},X_{\gamma i})P(D_i=d|Z_i)}$$

$$=\frac{[\prod_{k=1}^{k_1}\phi(\frac{h_k(T_{ki},\lambda_k^{(t)})-X_{\beta i}^T\beta_{kd}^{(t)}}{\sigma_k^{(t)}})\prod_{k=k_1+1}^{K}\prod_{j=1}^{J}\left(\eta_X^{-1}(X_{\gamma i}^T\gamma_{kdj_k})-\eta_X^{-1}(X_{\gamma i}^T\gamma_{kd(j_k-1)})\right)^{I(T_{ki}=j)}]\eta_Z^{-1}(Z_i^T\alpha_d^{(t)})}{\sum_{d=0}^{L-1}\{[\prod_{k=1}^{k_1}\phi(\frac{h_k(T_{ki},\lambda_k^{(t)})-X_{\beta i}^T\beta_{kd}^{(t)}}{\sigma_k^{(t)}})\prod_{k=k_1+1}^{K}\prod_{j=1}^{J}\left(\eta_X^{-1}(X_{\gamma i}^T\gamma_{kdj_k})-\eta_X^{-1}(X_{\gamma i}^T\gamma_{kd(j_k-1)})\right)^{I(T_{ki}=j)}]\eta_Z^{-1}(Z_i^T\alpha_d^{(t)})\}}.$$

The M step maximizes the expected value of $l_c(\theta)$ obtained by substituting $I(D_i = d)$ by $P^{(t)}(d)$ in function (7). We denote the result as function (7*) with the same arrangement of the three terms as in function (7). Note that the parameters in the three terms of (7*) are disjoint, so the maximization can be done separately. Next, we show that these terms are proper log likelihood functions of the weighted version of the corresponding regression models. Let W be a $N \times L$ by $N \times L$ diagonal matrix $W^{(t)} = diag\{P^{(t)}(0), \ldots, P^{(t)}(L-1)\}$, where $P^{(t)}(d)=(P_1^{(t)}(d),\ldots,P_N^{(t)}(d)),\ d=0,\ldots,L-1$. Let $Y_k = H_k(T_k, \lambda_k)$. Let $\mathbf{Z}$, $\mathbf{X}_\beta(D)$ and $\mathbf{X}_\gamma(D)$ be the design matrices defined by models (3)–(5), with possible interactions between X and D. Create stacked outcome vectors and design matrices as follows:

$$Y_k^*=\begin{pmatrix} Y_k \\ \cdots \\ Y_k \end{pmatrix},\quad Z^*=\begin{pmatrix} \boldsymbol{Z} \\ \ldots \\ \boldsymbol{Z} \end{pmatrix},\quad X_\beta^*=\begin{pmatrix} \mathbf{X}_\beta(D=0) \\ \cdots \\ \mathbf{X}_\beta(D=L-1) \end{pmatrix},\quad X_\gamma^*=\begin{pmatrix} \mathbf{X}_\gamma(D=0) \\ \cdots \\ \mathbf{X}_\gamma(D=L-1) \end{pmatrix}.$$

Then the first term in function (7*) becomes $W^{(t)} log\, \eta_X^{-1}(Z^*\alpha_d^{(t)})$. It is a log likelihood function of a weighted multinomial regression with $N \times L$ observations and weight $W^{(t)}$. Similarly, we can show that the second term and the third term in function (7*) are the log likelihood functions of a weighted Box–Cox regression model and a weighted cumulative link model, respectively. Therefore, maximization can be carried out by standard routines of these models.

Note that the creation of the stacked design matrices also makes it easy to impose constraints on regression coefficients. For example, to constrain some covariate effects to be the same across some disease status, one can simply remove the corresponding interaction terms in the design matrices and carry out the estimation in an unconstrained fashion. Similarly, if the outcome vectors and the design matrices are also stacked for each test T_k, constraints on covariate effects across different tests can be easily adopted.

3. Computational consideration

3.1. EM initialization

The following two procedures are used to obtain the initial values of the EM algorithm.

The first procedure is based on existing clustering methods. It first performs a crude clustering on all test results T_k's ignoring the covariates. Then the initial values are obtained

by performing the corresponding multinomial or cumulative link regression as specified in models (3)–(5) using results from the clustering method as values for the unknown disease status. Due to tests having mixed types, clustering methods based on Euclidean distance may not perform well, but any clustering method that allows for mixed-type variables can be used here. In particular, the method does not have to provide a hard classification, a fuzzy clustering method can also be used. In such case, the regressions to obtain the initial values become a weighted one. Clustering methods described in (Kaufman and Rousseeuw, 2009) (which are implemented in R package “cluster”) are adopted in our program. Fuzzy clustering method “fanny” appears to have better performance than others based on our experience.

Clustering methods work well when the influence of disease on test results is stronger than the influence of covariates, so that the groups are relatively separated. This may not always be the case, especially when biomarkers are involved. In general, it is recommended that one always chooses a wide range of initial values because the likelihood functions of finite mixture models usually have multiple local maxima (McLachlan and Peel, 2000). In our work, we always have 200 runs with random starting values, obtained by randomly assigning disease status to each individual and then performing the corresponding regressions. Results from all runs are then examined and the result with the highest likelihood value is chosen provided it is not a spurious local maximizer (more discussion are provided in Section 3.6).

3.2. Scaling the outcome vectors

For transformation regression, scaling the continuous test results T_k's by the nth root of the Jacobian determinant det J of transformation H_k can often make computation more stable, especially when T_k's have very different ranges or shapes. The scaling for the Box–Cox transformation is $Z_k = H_k(T_k)/(detJ)^{1/n} = Y_k/exp[(\sum_{i=1}^{N}(\lambda_k - 1)log\, T_{ik})/n]$. Other than numerical stability, this scaling also makes the Jacobian term in the second line of Eq. (7) disappear: $\sum_{i=1}^{N}\sum_{k=1}^{K} log\, J(t_{ki};\lambda_k) = \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k - 1)log\, 1 = 0$.

3.3. Unbounded likelihood

Finite mixture models may have unbounded likelihood functions. A simple example is two normal mixtures with heteroscedastic variances (Lehmann and Casella, 1998). This is because in the partial likelihood that relates to the continuous tests, the variance in the denominator can become very small. In our model for the continuous tests, the inclusion of transformation and covariates further complicated the situation. In this case, although the maximum likelihood estimates do not exist as a global maximizer, the work in Redner and Walker (1984) and Cheng and Traylor (1995) showed that there still exists a sequence of roots of the likelihood function corresponding to local maxima of the model, and they are consistent, efficient and asymptotically normal. Therefore, consistent estimates can still be obtained via an EM algorithm as long as we reach the consistent local maximum.

In the M step, the maximization of the second line in Eq. (7) is done by root finding of the corresponding gradient function so that one can reach local maximizers and avoid diverging

to the unbounded area of the likelihood function. Specifically, λ_k is obtained by solving the gradient function of the profile log likelihood as follows:

$$-N\frac{Y_{k,new}^{*T}P_{X_{new}^*}U_k(\lambda_k)}{Y_{k,new}^{*T}P_{X_{new}^*}Y_{k,new}^*}+\frac{N}{\lambda_k}+\rightarrow log\,T_k=0\ \ \lambda_k\neq 0$$
$$-N\frac{Y_{k,new}^{*T}P_{X_{new}^*}S_k(0)}{Y_{k,new}^{*T}P_{X_{new}^*}Y_{k,new}^*}+\rightarrow log\,T_k=0\ \ \lambda_k=0$$

where $X_{new}^*=W^{\frac{1}{2}}X_\beta^*$, $Y_{k,new}^*=W^{\frac{1}{2}}Y_k^*$, $P_X=I-X_{new}^*\left(X_{new}^{*T}X_{new}^*\right)^{-1}X_{new}^{*T}$, $U_k(\lambda_k)$ is a $L\times N$ by 1 vector with element $W_i^{\frac{1}{2}}(T_{ki}^{\lambda_k}log\,T_{ki})/\lambda_k$, and $S_k(0)$ is a $L\times N$ by 1 vector with elements $W_i^{\frac{1}{2}}(log\,T_{ki})^2/2$.

Substituting λ_k into log likelihood function and taking partial derivative with respect to β, we have $\beta_k=\left(X_{k,new}^{*T}WX_{new}\right)^{*-1}X_{new}^TWY_{k,new}^*$. Similarly,

$\sigma_k^2=\frac{1}{N}\sum_{i=1}^N\sum_{d=0}^{L-1}P_i^{(t)}(d)(Y_{ki}-X_i\beta_{kd})^2=\frac{1}{N}\left(Y_{k,new}^*-X_{new}^*\beta_k\right)^TW\left(Y_k^*-X_{new}^*\beta_k\right)=\frac{1}{N}Y_{k,new}^{*T}P_{X_{new}^*}Y_{k,new}^*$.

It is possible to assume heteroscedastic variance in model (4), i.e., $\varepsilon_{ik}\sim^{i.i.d.}N(0,\sigma_{kd}^2)$. In this case, the gradient function is,

$$-\sum_{d=0}^{L-1}\left\{\left[\sum_{i=1}^N P_i(d)\right]\frac{\sum_{i=1}^N P_i(d)(Y_{ki}-X_{\beta i}\beta_{kd})V_{ki}(\lambda_k)}{\sum_{i=1}^N P_i(d)(Y_{ki}-X_{\beta i}\beta_{kd})^2}\right\}+\frac{N}{\lambda_k}+\rightarrow log\,T_k=0\ \ \lambda_k\neq 0$$
$$-\sum_{d=0}^{L-1}\left\{\left[\sum_{i=1}^N P_i(d)\right]\frac{\sum_{i=1}^N P_i(d)(Y_{ki}-X_{\beta i}\beta_{kd})V_{ki}(0)}{\sum_{i=1}^N P_i(d)(Y_{ki}-X_{\beta i}\beta_{kd})^2}\right\}+\rightarrow log\,T_k=0\ \ \lambda_k=0$$

where $V_k(\lambda_k)$ is a N by 1 vector with elements $(T_{ki}^{\lambda_k}log\,T_{ki})/\lambda_k$, and $V_k(0)$ is a N by 1 vector with elements $(log\,T_{ki})^2/2$.

Then substituting λ_k into log likelihood function and taking partial derivatives with respect to β, we have $\beta_k=(X_{new}^{*T}WdX_{new})^{*-1}X_{new}^TWdY_{k,new}^*$, where $Wd=diag\{\frac{P(0)}{\sigma_{k0}},\ldots,\frac{P(L-1)}{\sigma_{k(L-1)}}\}$. Similarly,

$\sigma_{kd}^2=\frac{1}{\sum_{i=1}^N P_i(d)}\sum_{i=1}^N P_i(d)(Y_{ki}-X_i\beta_{kd})^2=\frac{1}{\sum_{i=1}^N P_i(d)}\left(Y_{k,new}^*-X_{new}^*\beta_k\right)_{[d]}^TP(d)_{[d]}\left(Y_{k,new}^*-X_{new}^*\beta_k\right)_{[d]}$

where the subscript [d] indexes the corresponding d^{th} "block" of the corresponding vector or matrix.

3.4. Label switching

A finite mixture model has an inherited identifiability issue due to “label switching”. This is because as with most clustering methods, it only groups subjects who are similar. Group labels are assigned later according to the specific context. However, this causes the likelihood function having multiple local maxima with the same value, corresponding to different permutations of group labels. Model parameter estimates may appear to be inconsistent. On the other hand, if we can classify subjects correctly into similar groups, finding the correct group labels is usually straight forward in diagnostic testing studies, as it is often apparent which group is the diseased group and which is the healthy group. In particular, one can often assume the direction of association between tests and disease status is known. In our estimating procedure, these directions are included as input and used to assign disease group labels. Without loss of generality, we assume higher test results correspond to higher disease status all for all tests. Then, group labels are assigned according to the average rank of tests within each cluster. This approach also reduces the impact of initial values on the EM algorithm.

3.5. Acceleration and convergence

A squared polynomial extrapolation method (Varadhan and Roland, 2008) can be used to accelerate the convergence of the EM algorithm. The algorithm is developed for any fixed-point iterative procedure, including the EM algorithm. It updates based only on values from the EM algorithm without requiring gradients. Briefly, let $\theta^{(i)}$, $\theta^{(i+1)}$, and $\theta^{(i+2)}$ be three consecutive EM updates on previous extrapolation $\tilde{\theta}^{(i)}$, then $\tilde{\theta}^{(i+1)}$ is updated as $\theta^{(i)} - 2\alpha(\theta^{(i+1)} - \theta^{(i)}) - \alpha^2(\theta^{(i+2)} - 2\theta^{(i+1)} + \theta^{(i)})$, where step length $\alpha = \|\theta^{(i+1)} - \theta^{(i)}\|/\|\theta^{(i+2)} - 2\theta^{(i+1)} + \theta^{(i)}\|$.

Both absolute and relative tolerances can be used as convergence criteria. Absolute tolerance is defined as $abstol = |\theta^{(i+1)} - \theta^{(i)}|$, and relative tolerance is defined as $reltol = (|\theta^{(i+1)} - \theta^{(i)}|)/(1 + |\theta^{(i)}|)$.

3.6. Spurious local maximizers

If the data have a group of outliers or a few data points that are overly close, the algorithm may fallaciously consider them as a group and results in a spurious local maximizer. This issue is usually eased when sample size is large. In diagnostic testing studies, since we usually have a good sense of group sizes for each disease group, spurious local maximizers can be easily identified. One can also impose constraints on group size or other parameters in the estimating procedure to avoid this issue. We did not choose this approach for the following reasons. With the relabeling procedure described in Section 3.4, the program mostly converges to the same place and rarely results in more than one solution. If it does occur, we believe it is worthwhile examining all the results since they may suggest an unexpected structure, incorrect modeling assumption of lack of model identifiability, rather than just a spurious solution. In addition, we also gain some computational simplicity.

4. Model identifiability

When using a model-based clustering method to study unobserved structures, one should always check for model identifiability to avoid reaching a faulty conclusion. Due to the inherent "label switching" problem, a finite mixture model is not globally identifiable. As a result, establishing identifiability of such a model has mostly focused on local area around the result (Goodman, 1974; Jones et al., 2010). In particular, a model $\boldsymbol{p} = f(\boldsymbol{\theta})$ is locally identifiable at $\boldsymbol{\theta}_0$ if there exists some neighborhood $U_{\boldsymbol{\theta}_0}$ of $\boldsymbol{\theta}_0$, such that $f(\boldsymbol{\theta}) \neq f(\boldsymbol{\theta}_0)$, $\forall \boldsymbol{\theta} \in U_{\boldsymbol{\theta}_0} \setminus \{\boldsymbol{\theta}_0\}$. This is equivalent to the Jacobian matrix of the modeling function f having full column rank. In the proposed model, one should evaluate the partial derivatives of model (6) at the converged value, and then check the column rank of the Jacobian matrix. Both of these can be done numerically.

It is worth mentioning that, our model includes covariates in both the models for mixing proportion $\pi_d(Z)$ and for component distribution $\gamma_{dk}(X)$, and allows for differential covariate effects that vary across disease status. This flexibility does not cause much identifiability problem as compared with a similar model without covariates. This is because the covariate effects are modeled linearly on some scale in Eqs. (3)–(5). In particular, when the design matrices have full column rank, they may help restore model identifiability even when the corresponding model without covariate is not identifiable. More discussion can be found in Forcina (2008) and Wang (2013).

5. Simulations

We conducted various simulations to assess model performance. In all simulations, there were at least 3 continuous tests and 3 three-level categorical tests, denoted by $\overrightarrow{T}$. We simulated 3 disease groups with a binary variable $Z \sim Bernoulli(0.5)$ that affects disease prevalence such that prevalence of the three disease groups $\overrightarrow{p} \approx (0.51, 0.31, 0.19)$ among subjects with $Z = 0$ and $\overrightarrow{p} \approx (0.23, 0.38, 0.38)$ among subjects with $Z = 1$. In other words, subjects with $Z = 0$ were most healthy, whereas risk factor $Z = 1$ leads to more subjects having mild or severe conditions. We compared the area under the ROC curve (AUC) values of each of the tests estimated based on the proposed model to those calculated using true disease group information. Mean biases and mean square error (MSE) are shown in Table 1. To keep the results concise, we only presented the results regarding AUC values for distinguishing disease level 1 versus disease level 0, denoted by $AUC_{1\,vs.0}$, from 3 representative tests. Results for other tests and for $AUC_{2\,vs.1}$ are similar.

In most medical biomarker studies, as well as in our real data example, the biomarkers selected in the study are reflective of disease progression, i.e., they are informative biomarkers, with biological, in-vivo and even clinical evidence. In addition, the researchers usually have a good understanding of covariates that may affect the diagnostic performance of these biomarkers. This setting is investigated in the first two rows in Table 1. We can see that the AUC estimates based on the proposed model are fairly close to the AUC value calculated using true disease information. Both mean biases and MSEs are small, and decreasing with increasing sample size. In addition, we investigated settings where irrelevant

covariates (U_1 and U_2 in the table) or non-informative tests (iT_1, …, iT_3 in the table) were included in the model. We can see that including limited number of irrelevant covariates does not seem to affect the AUC estimates much. On the other hand, including non-informative tests does affect AUC estimates slightly, possibly due to spurious association in finite samples. This is evident by two observations: first, the bias and MSE increase with higher number of non-informative tests; second, the bias and MSE decreases faster with increasing sample size, and at $N = 1500$ it is comparable to those in the first setting where there is no irrelevant covariates or non-informative tests. Nevertheless, all the aforementioned biases and MSEs are small, suggesting certain amount of robustness of the proposed model to these mis-specifications. However, one should note that including large number of non-informative tests can blur the clusters, as suggested by the slightly increased incidence of fitting errors and non-convergence in these settings. Finally, we examined situations where higher number of tests were included in the model. Rather than the 6 tests in $\overrightarrow{T}$, we added 3 to 18 additional tests in the model. The results are shown in the bottom part of Table 1. We can see that, with small sample size, including a large number of tests can cause bias; whereas with larger sample size, more tests can be included before increasing the bias. These results suggest that, when applying the proposed method, one should carefully choose a selected number of informative tests and consider increasing the sample size by roughly 100 for each additional test to be included.

6. Application to real data

We applied this method to a real data example concerning the clinical progression of Alzheimer's disease. As clinical trials aim to begin testing therapies to reduce or eliminate AD pathologic lesions in participants with mild cognitive impairment (MCI) and even normal cognition, accurate identification of those with underlying AD pathology driving their cognitive decline will be critical. The inclusion of participants who do not have AD pathology and are unlikely to progress to dementia would reduce power to detect a positive effect of the drug as these participants are not "at risk" to benefit from the therapy. Therefore, accurate diagnosis is a necessary step in designing clinical trials to evaluate new therapies to treat AD.

The Clinical Dementia Rating scale (CDR) is one of the most commonly used assessment instrument for staging dementia severity (Morris, 1993). The CDR examines cognitive functioning in 6 domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. CDR scores usually detect impairment long after the onset of pathological changes in the brain when the deterioration is sufficiently severe and has manifested into clinically detectable symptoms. Moreover, because the CDR is obtained through interviews, many factors may affect its accuracy. For example, subjects who are classified as having questionable, mild or even moderate impairment may be assessed as not impaired at later CDR assessments, suggesting possibly incorrect, or at least somewhat unstable, assessments. Therefore, although commonly used, the CDR is far from being a gold standard. On the other hand, imaging biomarkers are becoming increasingly available and have been used in conjunction with other clinical tests to improve diagnostic accuracy. This is because imaging biomarkers provide a more direct reflection of the

underlying pathological changes in the brain and have exhibited promising diagnostic ability when compared against the clinical diagnosis (Jack et al., 1999; Whitwell et al., 2008; Vemuri et al., 2009). It is expected that with imaging biomarkers, one can better distinguish subjects with true AD pathology from those who have a high CDR score due to transitory factors, such as depression, and therefore predict more accurately which subjects are likely to experience cognitive and/or functional decline at a follow-up visit. In this section, we apply the proposed method to assess this prognostic performance of combining Magnetic Resonance Imaging (MRI) biomarkers and CDR tests, to avoid biases due to imperfect reference standard and to estimate covariate effects on biomarkers/tests, therefore facilitating a personalized prognosis.

We used data from the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS) from 2005 to present. Described previously (Beekly et al., 2007), the UDS is an approximately annual assessment performed at one of the 34 past and present Alzheimer's Disease Centers (ADCs). At each assessment, demographic and clinical data are collected for subjects with normal cognition, mild cognitive impairment, AD dementia, and other dementias. Imaging data are available for a subset of UDS subjects who undergo imaging and have their images submitted voluntarily to NACC. Image summary calculations were performed by the IDEA lab (Director: Charles DeCarli, MD; University of California, Davis; http://idealab.ucdavis.edu/), following ADNI protocols (Jack et al., 2008). Only the first 5 domains of the CDR items are included in this analysis; personal care was excluded because deterioration in this domain usually happens in later stages of AD. The CDR items take on values of 0, 0.5, 1, 2, and 3 denoting increased impairment. We collapsed scores of 1, 2, and 3 into one category for all tests to restrict our attention to subjects who have less impairment and thus more difficult to prognose. Three MRI-based biomarkers are included: volumes of hippocampus, white matter hyperintensities and temporal lobe gray matter. Many studies have suggested that these biomarkers are predictive of AD progression and provide possibly complementary prognostic information (Chetelat and Baron, 2003; Storandt et al., 2012). Covariates included in the model that relate to CDR items are age (in 10-year unit), education (in years), and depression status within the past two years (binary). In the model that relates to the biomarker values, the covariates included are age and total brain volume (in 10^2 cc). Finally, in the model that relates to the prevalence, the covariates included are risk factors age and number of ApoE4 alleles (Blacker et al., 1997).

The analytic data set included subjects who had at least one MRI evaluation, had volumetric MRI biomarkers calculated, and had a UDS visit within 2 years of the MRI visit. This last requirement was imposed so that the CDR items were obtained at a time comparable to that of the MRI biomarkers. For the rest of the section, we refer to subjects' first MRI visit and the corresponding UDS visit as their baseline visits. We apply the proposed method to evaluate the ability of MRI markers and CDR items to detect AD-related changes and therefore provide prognostic information, and validate the results with the subjects latest cognitive assessment.

A total of 359 subjects with complete information on aforementioned covariates are included in the analysis. Among them, there are 166 subjects with normal cognition, 29 subjects who are impaired but do not meet criteria for MCI, 123 subjects with MCI of any type (Roberts

and Knopman, 2013), and 41 subjects who have AD dementia, based on cognitive evaluation at the baseline visit. We consider two latent groups in our analysis defined by whether or not a subject has AD-related changes and thus is deteriorating towards AD. Results are shown in Table 2, where 95% confidence intervals (CIs) were obtained from bootstrap method with 2000 bootstrap replicates.

These results suggest that the risk of developing AD related changes is associated with higher age and more ApoE4 alleles. Specifically, the odds of having AD related changes increases exp(0.57) = 1.77-fold (95% CI: 1.27 to 2.51) on average for every 10 years increase in age, and increases 2.27-fold (95% CI: 1.52 to 3.39) on average for every additional ApoE4 allele. The volumes of hippocampus and temporal gray are significantly lower in subjects with AD related changes after adjusting for total brain volume, suggesting possible diagnostic ability. Volume of white matter hyperintensities increase with AD related changes, consistent with existing studies, although this association is not significant based on these data. Moreover, increase in age is significantly associated with lower volume of hippocampus and higher volume of white matter hyperintensities among subjects in the same disease status group. This indicates the normal aging effect on brain structure, which should be taken into account when utilizing MRI biomarkers for AD diagnosis. From the test model, we find that AD related changes are significantly associated with lower CDR scores, as expected based on the validity of CDR tests. Additionally, education is significantly associated with higher test scores for subjects of the same disease status. Again, we see normal aging effect and additionally the effect of having depression within 2 years in lowing CDR scores, suggesting covariate effects need to be taken into consideration when developing threshold or other diagnostic rules regarding AD.

To have a more intuitive understanding of MRI markers' diagnostic ability, we plot their receiver operating characteristic (ROC) curves based on model estimates in Fig. 1. This plot shows that when only a single MRI marker (adjusted by total brain volume) is used, the volume of hippocampus has highest diagnostic ability, followed by the volume of temporal gray. We can also see the clear advantage of including covariates when using these biomarkers as reflected by the ROC curves getting closer to the top left corner. In particular, the AUC is 0.71 (95% CI: 0.64 to 0.78) for hippocampal volume, without accounting for subjects' characteristics; whereas the AUC increases to 0.74 (95% CI: 0.67 to 0.81) with covariate adjustment. Similarly, the AUC increases from 0.57 (95% CI: 0.49 to 0.63) to 0.65 (95% CI: 0.55 to 0.73) for volume of white matter hyperintensities, and from 0.67 (95% CI: 0.61 to 0.76) to 0.72 (95% CI: 0.62 to 0.79) for volume of temporal gray with covariate adjustment.

The model-based risk $P(D|T, X_\beta, X_\gamma, Z)$ can be obtained from the Bayes rule and model coefficient estimates. This risk score provides a convenient way of combining MRI markers and CDR items while taking into account varying prevalence and other covariate effects. As a partial validation, we compare the prognosis of the model based on baseline information with subjects' latest cognitive status determined by clinicians. It is a partial validation because although the latest clinical diagnosis uses longitudinal and all available information, it is still far from a gold standard and may be incorrect especially if the follow-up time is short. Nevertheless, three prognostic scenarios are examined: among subjects who are

diagnosed as not normal at baseline, how well does the model distinguish those who convert back to normal from those who are MCI or AD at latest follow-up; among subjects who are diagnosed as any type of MCI, how well the model distinguishes those who improve to normal or impaired but not MCI from whose who deteriorate to AD; and among subjects who are diagnosed as normal, how well the model distinguishes those who stay normal from those who progress (impaired but not MCI, any MCI or AD). In this evaluation, the model-based risk predicts that a subject will decline if the subject's probability of being in disease group $D = 1$ is greater than 0.5, based on the estimate from the model. The results are shown in Table 3. The results suggest that the baseline CDR items and MRI biomarkers have prognostic ability for distinguishing between subjects who will deteriorate and those who will not. Specifically, among subjects who are diagnosed as not normal, the model correctly predicts 28 out of 30 subjects who revert back to normal, and 108 out of 160 subjects who deteriorate to MCI or AD, resulting in a specificity of 0.93 and a sensitivity of 0.67. For MCI subjects, the model correctly predicts 10 out of 12 subjects who improve and 36 out of 55 subjects who decline (specificity = 0.83 and sensitivity = 0.69). The prediction is less optimal among normal subjects. This may be caused by the fact that long follow-ups are needed to examine subjects who are normal at baseline, and the prognostic performance of the model is masked by imperfection of latest clinical diagnosis. Similarly, we can also derive model-based risk without the MRI biomarkers, i.e., a personalized prognostic strategy based on CDR items and the subjects' characteristics such as age. Similar results are also included in Table 3. This comparison allows us to see how much MRI biomarkers improve upon CDR items. We find that for the same specificity, introducing MRI biomarkers improves sensitivity slightly, although this increase is not statistically significant. The results also suggest that when the subjects' characteristics are also included, the CDR items have good prognostic performance and can further distinguish among clinically similar subjects.

7. Discussion

A model-based clustering method for mixed type variables was introduced via the finite mixture modeling framework. The method represents an extension to current latent variable models. It was developed for two challenges in biomarker evaluations: (1) biomarkers are often used in combination with other tests; (2) a gold standard is often not available and other reference tests may be biased towards current tests and therefore mask the potential value of a biomarker. The proposed method allows for inclusion of possible differential covariate effects in both the mixture components and the mixing proportions. It also adopted a transformation model to relax the distributional assumption on the continuous manifest variables. We did not discuss the choice of number of latent classes, because research questions in medical studies usually make it apparent how many disease groups or subtypes one should consider. Otherwise, one can consider model selection criteria, such as BIC, to choose the number of latent classes. An R package "latentreg" has been developed for the proposed method. A preliminary version can be downloaded at: https://sites.google.com/site/wangzylab/resources. Future work will involve the extension of the categorical part of the model to allow for a variety of link functions and possible nominal covariate effects, and the extension of the continuous component of the model to allow for other parametric transformations as well semiparametric transformation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Wang was supported partially by NACC project 2015-JI-05 and by NIH/NCI Grant P30CA006973. The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs (listed individually in the Supplement (see Appendix A)).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.10.026.

References

Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics. 2004; 60(2):427–435. [PubMed: 15180668]

Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, Hubbard JL, Koepsell TD, Morris JC, Kukull WA, et al. The national Alzheimer's coordinating center (nacc) database: the uniform data set. Alzheimer Dis Assoc Disord. 2007; 21(3):249–258. [PubMed: 17804958]

Blacker D, Haines J, Rodes L, Terwedow H, Go R, Harrell L, Perry R, Bassett S, Chase G, Meyers D, et al. Apoe-4 and age at onset of Alzheimer's disease the nimh genetics initiative. Neurology. 1997; 48(1):139–147. [PubMed: 9008509]

Branscum A, Gardner I, Johnson W. Estimation of diagnostic-test sensitivity and specificity through bayesian modeling. Prev Vet Med. 2005; 68(2):145–163. [PubMed: 15820113]

Branscum AJ, Johnson WO, Hanson TE, Baron AT. Flexible regression models for ROC and risk analysis, with or without a gold standard. Stat Med. 2015; 34(30):3997–4015. [PubMed: 26239173]

Cheng R, Traylor L. Non-regular maximum likelihood problems. J R Stat Soc Ser B Stat Methodol. 1995:3–44.

Chetelat, Ga, Baron, JC. Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. Neuroimage. 2003; 18(2):525–541. [PubMed: 12595205]

Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. Stat Med. 2014; 33(24):4141–4169. [PubMed: 24910172]

Cook RJ, Ng E, Meade MO. Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. Biometrics. 2000; 56(4):1109–1117. [PubMed: 11129468]

Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. Lancet Neurol. 2007; 6(8):734–746. [PubMed: 17616482]

Forcina A. Identifiability of extended latent class models with individual covariates. Comput Statist Data Anal. 2008; 52(12):5263–5268.

Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Amer Statist Assoc. 2002; 97(458):611–631.

Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika. 1974; 61(2):215–231.

Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (roc) analysis without truth. Med Decis Making. 1990; 10(1):24–29. [PubMed: 2325524]

Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Stat Methods Med Res. 1998; 7(4):354–370. [PubMed: 9871952]

Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, L Whitwell J, Ward C, et al. The Alzheimer's disease neuroimaging initiative (adni): MRI methods. J Magn Reson Imaging. 2008; 27(4):685–691. [PubMed: 18302232]

Jack C, Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, Boeve BF, Waring SC, Tangalos EG, Kokmen E. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology. 1999; 52(7):1397–1403. [PubMed: 10227624]

Jones G, Johnson WO, Hanson TE, Christensen R. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics. 2010; 66(3):855–863. [PubMed: 19764953]

Jones G, Johnson W, Vink W, French N. A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: Application to investigating the temporal relationship between infection and disease. Biometrics. 2012; 68(2):371–379. [PubMed: 22004274]

Kaufman, L., Rousseeuw, PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Vol. 344. John Wiley & Sons; 2009.

Lehmann, EL., Casella, G. Theory of Point Estimation. Vol. 31. Springer Science & Business Media; 1998.

McLachlan, G., Peel, D. Finite Mixture Models. John Wiley & Sons; 2000.

Morris JC. The clinical dementia rating (cdr): current version and scoring rules. Neurology. 1993

Nestor PJ, Scheltens P, Hodges JR. Advances in the early detection of Alzheimer's disease. 2004

Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics. 2007; 8(2):474–484. [PubMed: 17085745]

Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics. 1996:797–810. [PubMed: 8805757]

Redner RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 1984; 26(2):195–239.

Roberts R, Knopman DS. Classification and epidemiology of MCI. Clin Geriatr Med. 2013; 29(4): 753–772. [PubMed: 24094295]

Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR, Kaye J, Montine TJ, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. 2011; 7(3):280–292. [PubMed: 21514248]

Storandt M, Head D, Fagan AM, Holtzman DM, Morris JC. Toward a multifactorial model of Alzheimer disease. Neurobiol Aging. 2012; 33(10):2262–2271. [PubMed: 22261556]

van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA. Latent class models in diagnostic studies when there is no reference standard—a systematic review. Am J Epidemiol. 2013:kwt286.

Varadhan R, Roland C. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. Scand J Stat. 2008; 35(2):335–353.

Vemuri P, Wiste H, Weigand S, Shaw L, Trojanowski J, Weiner M, Knopman D, Petersen R, Jack C, et al. MRI and CSF biomarkers in normal, MCI, and AD subjects predicting future clinical change. Neurology. 2009; 73(4):294–301. [PubMed: 19636049]

Vermunt JK, Magidson J. Latent class cluster analysis. Appl Latent Class Anal. 2002; 11:89–106.

Wang, Z. Ph D thesis. University of Washington; 2013. Latent Class and Latent Profile Analysis in Medical Diagnosis and Prognosis.

Wang Z, Zhou XH. Random effects models for assessing diagnostic accuracy of traditional chinese doctors in absence of a gold standard. Stat Med. 2012; 31(7):661–671. [PubMed: 21626532]

Whitwell J, Josephs K, Murray M, Kantarci K, Przybelski S, Weigand S, Vemuri P, Senjem M, Parisi J, Knopman D, et al. MRI correlates of neurofibrillary tangle pathology at autopsy a voxel-based morphometry study. Neurology. 2008; 71(10):743–749. [PubMed: 18765650]

Wu Z, Deloria-Knoll M, Hammitt LL, Zeger SL. Partially latent class models for case–control studies of childhood pneumonia aetiology. J R Stat Soc Ser C Appl Stat. 2016; 65(1):97–114.

Xu H, Craig BA. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. Biometrics. 2009; 65(4):1145–1155. [PubMed: 19210729]

Zheng Y, Barlow WE, Cutter G. Assessing accuracy of mammography in the presence of verification bias and intrareader correlation. Biometrics. 2005; 61(1):259–268. [PubMed: 15737102]

Zhou, XH., McClish, DK., Obuchowski, NA. Statistical Methods in Diagnostic Medicine. Vol. 569. John Wiley & Sons; 2009.

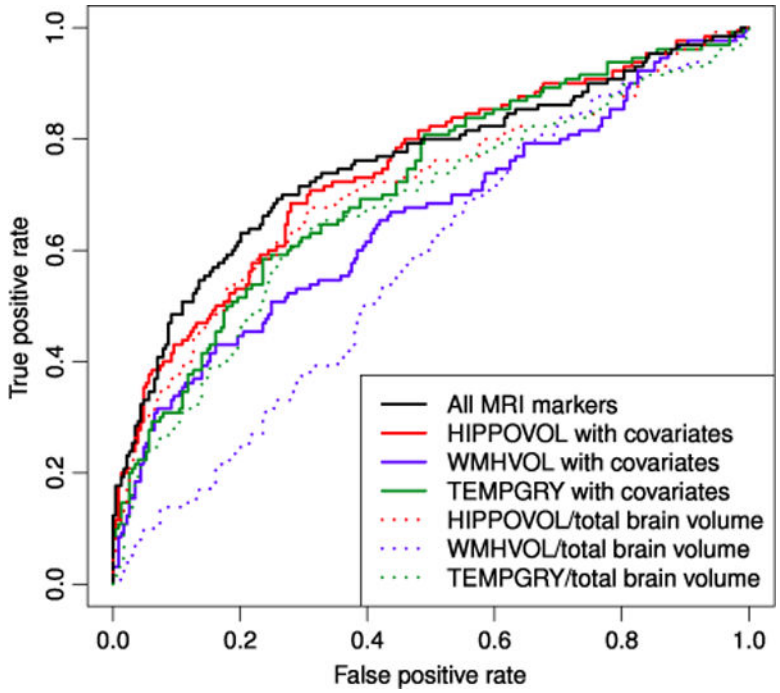

Fig. 1.
ROC curves of MRI biomarkers.

Table 1

Simulation results of mean bias ($\times 10^{-2}$) and MSE ($\times 10^{-3}$) in parenthesis of $AUC_{1\,vs.0}$ estimates based on the proposed method.

Models	N = 500				N = 800				N = 1500			
	T_{con1}	T_{cat1}	xT_1[a]	err[b]	T_{con1}	T_{cat1}	xT_1	err	T_{con1}	T_{cat1}	xT_1	err
$\overrightarrow{T}\sim X$	0.25[c] (0.08)[d]	−0.58 (0.11)	NA (NA)	0.00 0.02	0.13 (0.04)	−0.43 (0.06)	NA (NA)	0.00 0.04	0.15 (0.02)	0.24 (0.05)	NA (NA)	0.00 0.05
$\overrightarrow{T}\sim X+U_1$	0.22 (0.08)	−0.59 (0.11)	NA (NA)	0.00 0.03	0.13 (0.04)	−0.42 (0.06)	NA (NA)	0.00 0.05	0.15 (0.02)	0.24 (0.04)	NA (NA)	0.00 0.04
$\overrightarrow{T}\sim X+U_2$	0.23 (0.08)	−0.59 (0.11)	NA (NA)	0.00 0.01	0.14 (0.04)	−0.40 (0.06)	NA (NA)	0.00 0.01	0.14 (0.02)	0.24 (0.04)	NA (NA)	0.00 0.03
$\overrightarrow{T}\sim X+U_1+U_2$	0.22 (0.08)	−0.58 (0.11)	NA (NA)	0.00 0.00	0.14 (0.04)	−0.40 (0.06)	NA (NA)	0.00 0.03	0.13 (0.02)	0.23 (0.04)	NA (NA)	0.00 0.04
$\overrightarrow{T}+iT_1\sim X$	0.13 (0.11)	0.32 (0.19)	−0.33 (1.55)	0.06 0.04	0.20 (0.06)	0.47 (0.09)	−0.01 (0.08)	0.12 0.03	0.16 (0.02)	0.20 (0.04)	0.06 (0.03)	0.08 0.03
$\overrightarrow{T}+iT_1+iT_2\sim X$	0.31 (0.57)	0.45 (0.12)	−0.02 (0.13)	0.06 0.03	0.21 (0.06)	0.45 (0.09)	−0.01 (0.08)	0.12 0.02	0.15 (0.02)	0.21 (0.04)	0.05 (0.03)	0.08 0.02
$\overrightarrow{T}+iT_1+iT_2+iT_3\sim X$	0.70 (1.23)	0.60 (0.30)	0.05 (0.09)	0.06 0.07	0.45 (0.50)	0.49 (0.10)	−0.07 (0.08)	0.12 0.04	0.15 (0.02)	0.19 (0.04)	0.07 (0.03)	0.08 0.02
$\overrightarrow{T}+mT_1+\ldots+mT_3\sim X$	0.92 (1.50)	−0.31 (0.14)	0.92 (1.24)	0.02 0.04	0.33 (0.43)	−0.29 (0.03)	0.25 (0.30)	0.00 0.05	0.05 (0.01)	0.08 (0.01)	0.06 (0.01)	0.00 0.01
$\overrightarrow{T}+mT_1+\ldots+mT_6\sim X$	2.38 (4.20)	0.01 (0.27)	2.21 (3.64)	0.01 0.04	1.74 (3.08)	−0.09 (0.26)	1.61 (2.68)	0.02 0.07	0.21 (0.26)	0.07 (0.01)	0.21 (0.30)	0.00 0.02
$\overrightarrow{T}+mT_1+\ldots+mT_9\sim X$	3.57 (6.44)	0.04 (0.32)	3.38 (5.81)	0.00 0.04	2.29 (4.04)	−0.13 (0.27)	2.27 (3.97)	0.00 0.05	1.86 (3.38)	0.02 (0.56)	1.87 (3.40)	0.00 0.10
$\overrightarrow{T}+mT_1+\ldots+mT_{18}\sim X$	3.17 (5.64)	0.03 (0.36)	2.99 (5.06)	0.00 0.05	2.52 (4.31)	0.06 (0.32)	2.53 (4.40)	0.00 0.06	2.79 (5.14)	0.35 (0.71)	2.77 (5.08)	0.00 0.09

[a] xT_1 denotes either irrelevant test iT_1 or additional informative test mT_1 when appropriate.

[b] Error rates: fitting error on the top and non-convergence (within 500 iterations) on the bottom.

[c] Mean bias on scale of $\times 10^{-2}$.

[d] MSE on scale of $\times 10^{-3}$.

Comput Stat Data Anal. Author manuscript; available in PMC 2018 September 01.

Table 2

Estimates and 95% CI in parentheses based on 2000 bootstrap replicates (estimates in bold indicate a significant effect).

Prevalence model $P(D \mid Z)$			
	(Intercept)	**Age × 10**	**ApoE4**
$D = 1$	−**5.27** (−8.08, −2.73)	**0.57** (0.24, 0.92)	**0.82** (0.42, 1.22)

Biomarker model $P(T \mid D, X_\beta)$			
	HIPPOVOL	**WMHVOL**	**TEMPGRY × 10**
(Intercept)	2.39 (−0.33, 4.68)	−**3.68** (−5.73, −1.69)	−7.50 (−39.9, 0.68)
$D = 1$	−**0.89** (−2.09, −0.33)	0.13 (−0.10, 0.38)	−**1.19** (−4.21, −0.33)
Age × 10	−**0.28** (−0.79, −0.07)	**0.68** (0.48, 0.89)	0.02 (−0.49, 0.58)
BRNV × 100	**0.75** (0.28, 1.80)	0.07 (−0.04, 0.17)	**2.92** (0.88, 9.39)
λ	1.33 (0.79, 1.81)	0.06 (−0.02, 0.14)	1.42 (0.92, 1.90)
σ	1.13 (0.43, 2.61)	1.10 (0.92, 1.27)	2.19 (0.65, 6.98)

Test model $P(T \mid D, X_\gamma)$					
	MEMORY	**ORIENT**	**JUDEMENT**	**COMMUN**	**HOMEHOBB**
(Intercept1)[a]	1.8 (−0.6, 4.5)	3.1 (−0.2, 6.8)	2.7 (−0.5, 5.9)	**7.7** (3.9, 27)	**5.2** (1.9, 9.4)
(Intercept2)[b]	**6.0** (3.5, 9.2)	**5.8** (2.6, 9.7)	**5.1** (2.0, 8.4)	**9.7** (5.9, 29)	**7.5** (4.1, 12)
$D = 1$	−**4.8** (−6.3, −4.0)	−**5.0** (−6.2, −4.3)	−**4.2** (−5.1, −3.6)	−**6.7** (−23, −5.3)	−**5.1** (−6.6, −4.4)
Age × 10	−0.2, (−0.6, 0.1)	−**0.3** (−0.7, 0.2)	−0.2 (−0.6, 0.2)	−**0.5** (−1.1, −0.0)	−0.4 (−0.9, 0.0)
EDUC	**0.1** (0.0, 0.1)	**0.2** (0.1, 0.2)	**0.1** (0.0, 0.2)	**0.2** (0.1, 0.3)	**0.1** (0.0, 0.2)
DEP2YRS	−**1.0** (−1.5, −0.5)	−0.6 (−1.4, 0.0)	−0.35, (−1.0, 0.3)	−0.5 (−1.3, 0.3)	−0.5 (−1.2, 0.3)

[a] Intercept for having test score 0 vs. ≥ 0.5.

[b] Intercept for having test score ≤ 1 vs. >1.

Table 3

Comparison between prediction based on baseline MRI biomarkers and CDR tests versus cognitive status at latest follow-up.

Baseline[a]	**Not normal**		**Any type of MCI**		**Normal**	
Last follow-up[a] **Number of subjects**	**Normal** ***N*** **= 30**	**MCI or AD** ***N*** **= 160**	**no MCI no AD** ***N*** **= 12**	**AD** ***N*** **= 55**	**Normal** ***N*** **= 141**	**Not normal** ***N*** **= 25**
Prognostic accuracy	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
CDR+covariates[b]	0.93 (28[c])	0.65 (105)	0.83 (10)	0.65 (36)	0.90 (128)	0.24 (7)
CDR+MRI+covariates	0.93 (28)	0.67 (108)	0.83 (10)	0.69 (38)	0.91 (129)	0.24 (6)

[a]Based on clinician diagnosis.

[b]Based on model (6) excluding MRI markers and related covariates.

[c]Number of subjects with correct prognosis.