BIOMARKER ASSESSMENT AND COMBINATION WITH DIFFERENTIAL COVARIATE EFFECTS AND AN UNKNOWN GOLD STANDARD, WITH AN APPLICATION TO ALZHEIMER'S DISEASE

By Zheyu Wang[*], and Xiao-Hua Zhou[†]

Johns Hopkins University[*] and University of Washington[†]

The continued efforts to evaluate biomarkers' predictive abilities and identify optimal biomarker combinations are often challenged by the absence of a gold standard, i.e. the true disease status. Current methods that address this issue are mostly developed for binary or ordinal diagnostic tests, which do not fully utilize information provided by continuous biomarkers, or require strong parametric assumptions. Moreover, limited methods exist to allow for the inclusion of covariates – despite their crucial role in facilitating the accurate evaluation of biomarkers. In this paper, we proposed a latent profile approach to evaluating diagnostic accuracy of biomarkers without a gold standard. The method allows for flexible biomarker distributions and incorporation of previous knowledge about risk factors while simultaneously permitting researchers to model paticipants' characteristics that putatively affect biomarker levels, and therefore provides information needed to develop more personalized diagnostic procedures. Additionally, the proposed method presents a potential strategy for biomarker combination when gold standard information is unavailable, as it derives a composite risk score for the underlying disease status. The method is applied to evaluate different cerebral spinal fluid (CSF) biomarkers for Alzheimer's disease (AD) detection. The results show that CSF biomarkers hold significant potential for facilitating early AD detection and for continuous disease monitoring. Furthermore, they call attention to biomarker variability in subgroups and reexamination of CSF biomarker distributions. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

1. Introduction. In recent years, advances in biomarker discovery have re-energized the field of diagnostic medicine, as researchers continuously strive to obtain more convenient, economical, accurate, and/or timely diagnoses, by adding or combining various biomarkers to create novel diagnostic procedures. If one considers these biomarkers for use in creating diagnostic tests, traditional methods for diagnostic testing provide a means of evalu-

ation. Challenges arise when true disease status, commonly referred to as the gold standard, is not available due to cost constraints, ethics issue, or a lack of necessary biotechnology. In traditional diagnostic test setting, many methodologies have been proposed to address this issue (van Smeden et al., 2013; Collins and Huynh, 2014). For example, Zhou et al. (2005) proposed a nonparametric approach to evaluating ordinal tests without a gold standard, which generalized the first model introduced by Hui and Walter (1980). Qu et al. (1996) and Albert et al. (2001) introduced random effects to allow for dependence structure among the tests. Several authors explored Bayesian approach (Branscum et al., 2008, 2015; Wu et al., 2016). Wang et al. (2011), Wang and Zhou (2012), discussed the evaluation when the gold standard is ordinal to capture additional severity or subtype information. Different longitudinal models were proposed for repeated test measurements (Cook et al., 2000; Jones et al., 2012).

There are three major issues when applying these methods to biomarker evaluation. First, these methods were developed for binary or ordinal tests, yet most biomarkers are continuous. Categorizing biomarkers can result in a loss of important information. In fact, Albert and Dodd (2004) showed that different models can become indistinguishable when only a small number of binary tests are used. Moreover, many biomarkers lack an established threshold for use in delineating clinically meaningful categories. Current methods for continuous tests without a gold standard usually require strong parametric assumptions, such as multivariate normality. Secondly, existing work rarely discusses the inclusion of covariates. Some exceptions can be found in Bandeen-Roche et al. (1997), Branscum et al. (2008) and Pfeiffer et al. (2008), where researchers included covariates to allow for a varying prevalence. However, covariates can affect accuracy of the assessments in at least one other way: they may influence the magnitude and accuracy of test results in different subgroups. These effects are prevalent among biomarker studies, as most biomarkers values can be highly sensitive to the protocol and assay applied, as well as other inter-individual differences such as age and gender. A pooled analysis ignoring covariates can therefore lead to biased accuracy estimates (Pepe, 2003; Janes and Pepe, 2009). Inclusion of covariates that affect test results also allows researchers to relax the conditional independence assumption. In fact, this approach is generally preferred in biomarker evaluations, as compared with a random effects approach. This is due to the potential for covariates to explain, more explicitly, the mechanism by which dependence occurs, and to show how biomarker levels vary with covariates, as well as how their diagnostic performance differs among subpopulations – all are crucial issues in biomarker evaluations. However, due to computa-

tional difficulty and identifiability issues, only a limited body of work exists (Branscum et al., 2015; Huang and Bandeen-Roche, 2004), and the methods are restricted to a single biomarker with covariate effect only in the disease group, or including only a small number of categorical covariates, or to impose modeling constraints, such as assuming constant covariate effect across disease groups. Thirdly, biomarker studies usually have slightly different foci than traditional diagnostic testing studies: In addition to assessing test accuracy, biomarker studies often aim to propose novel biomarkers to combine with, rather than to replace, current markers. Nevertheless, biomarker variation across subgroups introduces another key question, regarding how to synthesize the information to achieve more personalized diagnoses.

Our research is motivated by such questions in AD biomarker studies. A definite diagnosis of AD needs a neuropathological brain autopsy, which requires proper medical resources and can only be conducted after a patient has died and provided consent. Moreover, AD-related pathophysiological changes are believed to begin 10 years or more before any detectable clinical symptom, and decades before sufficient cognitive impairment accrues to warrant a clinical diagnosis of AD – an event that usually occurs 5 to 30 years before a patient's death (Jack Jr et al., 2010). Therefore, all other difficulties aside, merely having sufficient follow-up data to make a potential gold standard diagnosis is challenging. On the other hand, this long time-gap provides an opportunity for early diagnosis and intervention. Guided by the amyloid hypothesis (Selkoe, 1991), researchers seek to use biomarkers to detect preclinical AD patients before extensive neuron damage and synaptic loss occurs. Additionally, biomarker information can be useful for monitoring AD progression, as well as for evaluating new AD treatments and recruiting specific subsets of patients for inclusion in AD clinical trials. Existing studies often rely on clinical AD diagnoses as the gold standard in their evaluations, due to inadequate levels of follow-up data. Additionally, existing knowledge about risk factors and covariates that affect biomarker level is often ignored. In this paper, we propose a latent profile approach to simultaneously handling the above issues. The approach further provides a risk score based on multiple biomarkers, risk factors and subjects' characteristics, and therefore provides the opportunity for a more personalized diagnosis.

2. Model. Let D denote the unobserved true disease status. For binary disease status, $D = 1$ indicates disease and $D = 0$ nondisease. For ordinal disease status, $D = 0, \ldots, L - 1$ denotes level of severity, with higher value of D indicating more severe status. Let $\vec{T} = (T_1, \ldots, T_K)$ de-

note K continuous biomarkers. We assume that for all biomarkers, a higher value is more indicative of a more severe status. A latent profile model (or a finite mixture model with continuous component distributions) (Lazarsfeld and Henry, 1968) without covariates is $f(\vec{t}) = \sum_{d=0}^{L-1} \pi_d \boldsymbol{\gamma}_d(\vec{t})$, where $\vec{t} = (t_1, \ldots, t_K)$, $\pi_d = P(D = d)$ which is usually referred to as the latent structure part, and the conditional density $\boldsymbol{\gamma}_d(\vec{t}) = P(\vec{T} = \vec{t} \mid D = d)$ which is usually referred to as the measurement part. The conditional independence assumption among the biomarkers within each severity group is commonly adopted to simplify the joint measurement model $\boldsymbol{\gamma}_d(\vec{t})$ as its univariate products $\prod_{k=1}^{K} \gamma_{dk}(t_k)$. When this assumption does not hold, a random effect approach can be used to account for the dependence (Qu et al., 1996; Albert et al., 2001).

To extend this model to incorporate covariate information, let $\mathbf{X} = (X_1, \ldots, X_q)$ denote the covariates that may affect biomarker values within each severity group and $\mathbf{Z} = (Z_1, \ldots, Z_p)$ denote the covariates related to disease prevalence. The elements in $\mathbf{X}$ and $\mathbf{Z}$ can be overlapping or mutually exclusive. We assume that the biomarker values are conditionally independent, given the true disease status D and the covariates $\mathbf{X}$. In other words, we assume that the possible dependence among biomarker values within each disease severity group can be explained by covariates $\mathbf{X}$, such as some subjects' characteristics. Under this assumption, the latent profile model with covariates is given as follows:

$$(2.1) \qquad f(\vec{t} \mid \mathbf{X}, \mathbf{Z}) = \sum_{d=0}^{L-1} \big[\pi_d(\mathbf{Z}) \prod_{k=1}^{K} \gamma_{dk}(t_k \mid \mathbf{X}) \big],$$

where $\pi_d(\mathbf{Z}) = P(D = d \mid \mathbf{Z})$ and $\gamma_{dk}(t_k \mid \mathbf{X}) = P(T_k = t_k \mid D = d, \mathbf{X})$.

A natural way to model the varying prevalence $\pi_d(\mathbf{Z}) = P(D = d \mid \mathbf{Z})$ is to use polytomous regression. Specifically, let $\eta(\vec{z}_i^T \boldsymbol{\alpha}_d) = P(D_i = d \mid \vec{Z}_i = \vec{z})$ as in a generalized linear model, subject to normalizing conditions $\sum_{d=0}^{L-1} \eta(\vec{z}_i^T \boldsymbol{\alpha}_d) = 1, \forall \vec{z}_i \in \mathcal{X}$, where $i = 1, \ldots, N$ is the subject ID and $\mathcal{X}$ denotes the covariate space. The model for the latent structure part is:

$$\eta(\vec{z}_i^T \boldsymbol{\alpha}_d) = \frac{exp(\vec{z}_i^T \boldsymbol{\alpha}_d)}{1 + \sum_{l=1}^{L-1} exp(\vec{z}_i^T \boldsymbol{\alpha}_l)}, \quad \log \frac{\eta(\vec{z}_i^T \boldsymbol{\alpha}_d)}{\eta(\vec{z}_i^T \boldsymbol{\alpha}_0)} = \vec{z}_i^T \boldsymbol{\alpha}_d, \quad d = 1, ..., L-1,$$

where $D = 0$ is the baseline group with parameters $\boldsymbol{\alpha}_0 = (\alpha_{00}, \ldots, \alpha_{0p}) = (0, \ldots, 0)$.

A transformation regression to describe the measurement part $T_k \mid D, \mathbf{X}$ is defined as:

$$H_k(T_{ik}; \lambda_k) = \vec{X}_i^T \boldsymbol{\beta}_{kd} + \epsilon_{ik}, \quad \epsilon_{ik} \overset{i.i.d.}{\sim} G(\cdot; \nu_k),$$

where $G(\cdot; \nu_k)$ is a distribution function with parameter ν_k, $H_k(\cdot; \lambda_k)$ is a pre-specified monotonic function with parameter λ_k. When the error distribution is assumed normal, the transformation accounts for possible skewness of the biomarker values within each severity group.

Then, the latent profile model with covariates is given as follows:

$$(2.2) \quad f(\vec{t} \mid \mathbf{X}, \mathbf{Z}) \;\; = \;\; \sum_{d=0}^{L-1} \left\{ \pi_d(\mathbf{Z}) \prod_{k=1}^{K} \left[J(t_k; \lambda_k) \tilde{\gamma}_{dk}\big(H_k(t_k; \lambda_k) \mid \mathbf{X} \big) \right] \right\},$$

where $J(t_k; \lambda_k) = \partial H_k(t_k; \lambda_k)/\partial t_k$ is the Jacobian of transformation $H_k(\cdot; \lambda_k)$, and $\tilde{\gamma}_{dk}(H_k(t_k; \lambda_k) \mid \mathbf{X}) = P(H_k(t_k; \lambda_k) \mid D = d, \mathbf{X}) = g(H_k(t_k; \lambda_k) - \mathbf{X}^T \boldsymbol{\beta}_{kd}; \nu_k)$ with $g(\cdot; \nu_k)$ denoting the corresponding density function of $G(\cdot; \nu_k)$.

The advantage of this model is that it allows for a varying disease prevalence based on risk factors and simultaneous inclusion of differential covariates that affect biomarker values. Currently, such models are limited and often require constant covariate effects to ensure model identifiability (Huang and Bandeen-Roche, 2004). In contrast, we allow covariate effect $\boldsymbol{\beta}_{kd}$ to depend on both biomarker index k and disease status d, that is, covariates can have differential effects among biomarkers and among disease groups.

To better understand this improvement, we consider an example when D is binary. The covariate-specific receiver operating characteristic (ROC) curve based on the measurement model is: $\mathrm{ROC}_k(t, \vec{X}) = 1 - G_1(G_0^{-1}(1 - t) + \vec{X}^T \boldsymbol{\beta}_{k0} - \vec{X}^T \boldsymbol{\beta}_{k1})$. If covariate effects are assumed to be the same for every disease group with only intercept terms differing, i.e., $\boldsymbol{\beta}_{kd} = \{\beta_{kd0}, \beta_{k1}, \ldots, \beta_{kq}\}$, the ROC curve no longer depends on covariates. In other words, when D is binary, assuming a constant covariate effect across disease groups is equivalent to assuming that the covariate-specific ROC curves are the same as the marginal ROC curves, which is rarely the case in practice. A similar conclusion can be made about generalized disease status, D. Justification for model identifiability with this flexible structure is provided in section 4.

3. Estimation.

3.1. *An EM Algorithm.* The maximum likelihood estimates of parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_d, \boldsymbol{\beta}_{kd}, \lambda_k, \nu_k \mid d = 0, \ldots, L-1 \text{ and } k = 1, \ldots, K\}$ in model (2.2) can be obtained by directly maximizing the observed data likelihood function, or by considering the true disease status as missing and applying

the EM algorithm to the complete data likelihood function below:

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{d=0}^{L-1} I_d(D_i) \log \pi_d(\vec{Z}_i) + \sum_{i=1}^{N} \sum_{k=1}^{K} \log J(t_{ik}; \lambda_k) + \sum_{i=1}^{N} \sum_{d=0}^{L-1} \sum_{k=1}^{K} I_d(D_i) \log \tilde{\gamma}_{dk}(t_k^{(\lambda_k)} \mid \mathbf{X}),$$

where $I_d(D_i)$ is an indicator function. It equals 1 if $D_i = d$ and equals 0 otherwise.

Details of the EM algorithm can be found in the appendix. In particular, the maximization in the M step can be broken down into two disjoint parts to simplify the computation, one for α_d and the other for λ_k, β_{kd} and ν_k. In addition, these two maximizations can be rewritten as a maximization of a weighted polytomous regression likelihood and a maximization of a weighted transformation regression likelihood. As a result, computational routine in standard statistical package can be adopted.

As mentioned before, the proposed model does not require constraints on the covariate effects across disease groups. However, if such constraints are desired based on scientific knowledge, we discussed an easy way to incorporate them in the design matrix without having to do a constrained maximization.

3.2. *Biomarker Accuracy and Combination.* After obtaining parameter estimates of the model, accuracy measures of each biomarker can be easily derived. For example, sensitivity, specificity and the covariate-specific ROC curve can be calculated based on $P(T_k \mid D, \mathbf{X})$. Positive predictive value, negative predictive value and other measures can be obtained with additional quantity $P(d)$. Furthermore, the model based risk $P(D_i \mid \vec{T}_i, \vec{X}_i, \vec{Z}_i, \theta)$, obtained by applying Bayes' Formula to $P(T_k \mid D, \mathbf{X})$ and $P(d)$, provides a potential way for biomarker combination to improve accuracy. This risk score not only summarize diagnostic information from multiple biomarkers when a gold standard is not available, but also takes into consideration risk factors and subjects' characteristics, hence offering a more personalized diagnosis. However, if one wishes to evaluate the accuracy performance of the combination rule, one needs to have an independent test sample.

3.3. *Computational Issues.*

3.3.1. *Initial parameter values.* A wide choice of initial values are often needed for latent profile models because the likelihood functions usually have multiple local maxima (McLachlan and Peel, 2004). Accordingly, the following two methods are used in our computation.

Initial values by clustering: perform a crude clustering on the biomarker values ignoring covariates (e.g. K-means clustering). Label the resulting clusters such that higher values of D corresponds to clusters with higher biomarker values. Initial values are then obtained by polytomous regression of label D on $\mathbf{Z}$ and transformation regression of $\vec{T}$ on $\mathbf{X}$ and D.

Random starting values: if prior information on disease severity group is available, the prevalence is generated accordingly. Otherwise, randomly generate $L-1$ numbers $a_1, \ldots, a_{L-1}$ from the standard uniform distribution $U(0,1)$. Set the prevalence as $\vec{p} = \{a_1, a_2 - a_1, \ldots, a_{L-1} - a_{L-2}, 1 - a_{L-1}\}$ and generate disease label for each subject from a multinomial distribution with probabilities $\vec{p}$. Then the initial values are obtained as previously described.

3.3.2. *Scaling the biomarker values.* Scaling T_k 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. For example, when the Box-Cox transformation (Box and Cox, 1964) is used, the scaling is $Z_k = T_k^{(\lambda_k)}/(\det J)^{1/n}$, where $(\det J)^{1/n} = \exp\left[(\sum_{i=1}^{N}(\lambda_k - 1)\log T_{ik})/n\right]$, $k = 1, \ldots, K$. Other than numerical stability, this scaling also makes the Jacobian term in equation (7.3) disappear: $\sum_{i=1}^{N} \sum_{k=1}^{K} \log J(t_{ik}; \lambda_k^{(t)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} (\lambda_k^{(t)} - 1)\log 1 = 0$.

3.3.3. *Spurious local maximizers.* If the error distribution in the measurement model is allowed to be heteroscedastic among disease groups, i.e. $H_k(T_{ik}) = \vec{X}_i^T \beta_{kd} + \epsilon_{ikd}$, $\epsilon_{ikd} \overset{i.i.d.}{\sim} N(0, \sigma_{kd}^2)$, a relatively large local maximum can occur as a consequence of fitting a disease group having only a few, and very close observations (therefore the variance in the denominator of the likelihood function becomes very small). The resulting parameter estimates are usually called spurious local maximizers. This problem is due to the fact that the likelihood function for a heteroscedastic mixture model can be unbounded (Lehmann and Casella, 1998).

When the likelihood is unbounded, the maximum likelihood estimates do not exist as a global maximizer. This does not invalidate our method, because maximizing the likelihood function is not our goal; rather, it is to find estimates that are consistent. In this situation, there still exists a sequence of roots of the likelihood function corresponding to local maxima for our model, and they are consistent, efficient and asymptotically normal. Further discussion can be found in Redner and Walker (1984), Cheng and Traylor (1995). As a result, consistent estimates can still be obtained via an EM algorithm. We obtain the maximizers of l_2 by finding the roots of its derivative l_2' instead of direct maximization to avoid diverging to the unbounded area. After

reaching convergence, one needs to examine the relative size or the variance of the fitted components to exclude possible spurious local maximizers. This can also be handled by imposing constraints in the maximization, such as requiring that for each given k, $\sigma_{kd}/\sum_{h=0}^{D}\sigma_{kh} \leq R/(D+1)$, $\forall d = 0, \ldots, L-1$, where R is a pre-specified positive constant that restricts the relative size of the component variances. We did not use the constraint approach because (1) in diagnostic settings spurious local maximizers can be easily identified with very minimal information about prevalence; (2) unconstrained maximization is easier and faster than constrained ones; and (3) it can be useful to examine all results because sometimes a fit with dissimilar components is not a spurious solution.

4. Model Identifiability. Due to an inherent "label switching" problem (the distribution remains identical if the labels of the latent groups are switched) and that the likelihood functions of these models usually have multiple local maxima, research on latent variable model identifiability has mainly focused on local identifiability, which considers whether the likelihood can uniquely determine a set of parameter values in its neighborhood. Specifically, a model is a function $\boldsymbol{p} = f(\boldsymbol{\theta})$ that maps points in the parameter space into the data space. The model is locally identifiable at $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ if there exists some neighborhood $U_{\boldsymbol{\theta}_0}$ of $\boldsymbol{\theta}$, such that $f(\boldsymbol{\theta}) \neq f(\boldsymbol{\theta}_0), \forall \ \boldsymbol{\theta} \in U_{\boldsymbol{\theta}_0} \setminus \{\boldsymbol{\theta}_0\}$. This suggests that F is locally invertable in $U_{\boldsymbol{\theta}_0}$. By the weak inversion theorem, this is equivalent to the Jacobian matrix $J(\boldsymbol{\theta}_0)$ having full column rank.

For latent class models, McHugh (1956) proposed sufficient conditions for local identifiability of models with dichotomous observed variables. Goodman (1974) extended these conditions to polytomous variables. Here we borrow these results to obtain identifiability conditions for models with covariates.

The link function $\eta(\cdot)$ in the structure model and transformation $H_k(\cdot; \lambda)$ are both monotonic functions and thus do not affect model identifiability. Therefore, without loss of generality, we assume that they are both the identity function. Model (2.1) can then be rewritten as follows,

$$(4.1) \quad f(\vec{T} = \vec{t} \mid \mathbf{X}, \mathbf{Z}) = \sum_{d=0}^{L-1} \Big[\pi_d(\mathbf{Z}) \prod_{k=1}^{K} \prod_{j=1}^{J_k} \gamma_{dkj}(t_k \mid \mathbf{X})^{I[t_k=j]}\Big],$$

with $\pi_d(\mathbf{Z}) = \mathbf{Z}\boldsymbol{\alpha}_d$ and $\gamma_{dkj}(t_k \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_{kdj}$.

Let $\{\Pi_d = \partial f/\partial \pi_d \mid d = 1, \ldots, L-1\}$ and $\{\Gamma_{dkj} = \partial f/\partial \gamma_{dkj} \mid d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$ be the column vectors of the Jacobian matrix J of the corresponding latent class model without covariates. Ex-

plicit expressions of these vectors are not necessary for the following derivation, but can be found in Wang (2013); Wang and Zhou (2014). When covariates are included, by the chain rule, we have $\partial f/\partial\alpha_d = \partial f/\partial\pi_d \times \partial\pi_d/\partial\alpha_d$ and $\partial f/\partial\beta_{dkj} = \partial f/\partial\gamma_{dkj} \times \partial\gamma_{dkj}/\partial\beta_{dkj}$. Therefore the column vectors of the Jacobian matrix of model (4.1) is

$$
A_d = \begin{pmatrix} \Pi_d(\vec{X}_1)\pi'_d(\vec{Z}_1\alpha_d)\vec{Z}_1 \\ \vdots \\ \Pi_d(\vec{X}_N)\pi'_d(\vec{Z}_1\alpha_d)\vec{Z}_N \end{pmatrix}, \quad B_{kjd} = \begin{pmatrix} \Gamma_{kjd}(\vec{X}_1,\vec{Z}_1)g'_{kjd}(\vec{X}_1\beta_{kjd})\vec{X}_1 \\ \vdots \\ \Gamma_{kjd}(\vec{X}_N,\vec{Z}_N)g'_{kjd}(\vec{X}_1\beta_{kjd})\vec{X}_N \end{pmatrix},
$$

where A_d is a $N\left[(\prod_{k=1}^{K} J_k)-1\right]$ by p matrix and B_{kjd} is a $N\times((\prod_{k=1}^{K} J_k)-1)$ by q matrix, $d = 0,\ldots,L-1$, $k = 1,\ldots K$, $j = 1,\ldots,J_k$.

Therefore, the Jacobian matrix J^* of model (4.1), has similar pattern as the Jacobian matrix J of the corresponding latent class model without covariates, but has N times as many rows as J. By applying condition (20-21) in Goodman (1974), we have the following theorem:

Theorem 1 Model (2.1) is locally identifiable at parameter $\theta = \{\alpha_d, \beta_{kjd} \mid d = 0,\ldots,L-1; k = 1,\ldots K; j = 1,\ldots,J_k\}$ if the following conditions hold:
(i) $N\left[(\prod_{k=1}^{K} J_k) - 1\right] \geq q \times L[\sum_{k=1}^{K}(J_k - 1)] + p(L-1)$;
(ii) $P(\vec{T} = \vec{t} \mid \mathbf{X}_i, \mathbf{Z}_i) > 0, \quad \forall \vec{T}$;
(iii) Column vectors in matrices $\{\Pi_d(\mathbf{X}_0) \mid d = 1,\ldots,L-1\}$, $\{\Gamma_{kjd}(\mathbf{X}_0,\mathbf{Z}_0) \mid d = 0,\ldots,L-1; k = 1,\ldots K; j = 1,\ldots,J_k\}$ all together are linearly independent for some $\mathbf{X}_0 \in \{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ and $\mathbf{Z}_0 \in \{\mathbf{Z}_1,\ldots,\mathbf{Z}_N\}$; and
(iv) Design matrix $\mathbb{X}$ and $\mathbb{Z}$ both have full rank.

The first condition simply requires the degrees of freedom in the data is greater than the number of free parameters in the model, a necessary condition for model identifiability. The second condition means that all biomarker values in the domain are potentially observable. Conditions (iii) and (iv) guarantee that column vectors A_d and B_{kjd} are linearly independent, which means the model Jacobian matrices J^* has full column rank. Therefore, the model is locally invertable at θ_0.

5. Simulation Studies.

5.1. *Regular settings.* This section assessed model performance under various correctly specified settings. In all simulations, we chose 3 biomarkers ($K = 3$) and 3 disease categories ($L = 3$). In the latent structure model, we assumed that a binary variable, $Z \sim$ Bernulli(0.5), affected the disease prevalence $P(D_i = d \mid Z_i = z) = \eta(\alpha_{d0} + \alpha_{d1}z)$, $d = 1,2$. We chose $\alpha = (\alpha_{10}, \alpha_{11}, \alpha_{20}, \alpha_{21})$=(-0.5, 1, -1, 1.5). This parameter value resulted in

prevalence $\vec{p} \approx$ (0.51, 0.31, 0.19) among subjects with $Z = 0$ and $\vec{p} \approx$ (0.23, 0.38, 0.38) among subjects with $Z = 1$, where $\vec{p} = (P(D = 0 \mid Z), P(D = 1 \mid Z), P(D = 2 \mid Z))$. As a result, subjects with $Z = 0$ were most healthy with about 20% having severe conditions, whereas risk factor $Z = 1$ leads to more subjects having mild or severe conditions.

We used the Box-Cox and the Yeo-Johnson transformations for $H_k(\cdot; \lambda_k)$. We assumed that a normally distributed variable, $X \sim N(0, 1)$, affected the biomarker performance via regression model $H_k(T_{ik}) = \tilde{\beta}_{kd0} + \tilde{\beta}_{kd1} X_i + \epsilon_{ik}$, $d = 0, 1, 2$, with $\lambda = 1$, 0.5 or 0. This can be reparameterized as $H_k(T_{ik}) = \beta_{k0} + \beta_{k1} X_i + \beta_{k2} I_1(D_i) + \beta_{k3} I_2(D_i) + \beta_{k4} X_i I_1(D_i) + \beta_{k5} X_i I_2(D_i) + \epsilon_{ik}$, $k = 1, \ldots, K$. Error distributions were generated as $\epsilon_{ik} \sim N(0, 0.5^2)$ for all biomarkers. Two sets of coefficient values were used: $\beta_k =$ (5, -1, 1, 2.5, 0.5, 1.5) representing biomarkers with good discriminating ability, and $\beta_k =$ (5, -1, 0.5, 1, 0.5, 1.5) representing biomarkers with fair discriminating ability. For the first scenario, the covariate-specific AUC for discriminating $D = 0$ versus $D = 1$ is $\Phi(\frac{\beta_{k2} + \beta_{k4} X}{\sigma_k}) = \Phi(2 + X)$ (Φ denotes the standard normal cumulate distribution function) with an average value of the covariate-specific AUC in this population being $\Phi(\frac{\beta_{k2}}{\sqrt{\sigma_k^2 + \beta_{k4}^2}}) = \Phi[(0.5^2 + 0.5^2)^{-1/2}] \approx 0.92$. The covariate-specific AUC for discriminating $D = 1$ versus $D = 2$ is $\Phi(3 + 2X)$, with an average value in this population being $\Phi[1.5(1 + 0.5^2)^{-1/2})] \approx 0.91$. Similarly, for the second scenario, the covariate-specific AUC for discriminating $D = 0$ versus $D = 1$ is $\Phi(1 + X)$, with the average value in this population being $\Phi[0.5(0.5^2 + 0.5^2)^{-1/2}] \approx 0.76$. The covariate-specific AUC for discriminating $D = 1$ versus $D = 2$ is $\Phi(1 + 2X)$, with the average value in this population being $\Phi[0.5(0.5^2 + 1^2)^{-1/2}] \approx 0.67$.

Simulation results for biomarkers that have good or fair diagnostic performances with the Box-Cox transformation were shown in Table 1 and Table 2. Results for the Yeo-Johnson transformation are similar (not shown). In each scenario, we considered sample sizes $N = 500$ and $N = 800$. All results were based on 1000 simulation replicates.

These results suggest that the proposed method converged to the true parameter values. In the current settings, parameters λ_k and α_d can be well estimated with a sample size of 500. Increasing the sample size to 800 did not change the results much. On the other hand, bias in the estimates for β_{kd} decreased when the sample size increased. This is because, in our model β_{kd} was assumed to differ across biomarkers and disease groups. The effective sample size for estimating β_{kd} was less than that for estimating λ_k and α_d. In addition, the biases for β_{kd} and σ_k are smaller with smaller λ_k. Comparing these two tables, the strength of biomarkers' diagnostic ability seem to reduce the bias and standard error of the estimates. Note that bias

TABLE 1

Mean and standard error (in parentheses) estimates based on 1000 simulations for biomarkers with good performance.

$N = 500$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.498 (0.18)	1.005 (0.26)	-1.007 (0.19)	1.513 (0.26)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 1$	$\beta_{k3} = 2.5$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.131 (0.89)	-1.044 (0.27)	1.052 (0.29)	2.644 (0.77)	0.515 (0.14)	1.573 (0.44)
$k = 2$	5.080 (0.72)	-1.030 (0.23)	1.035 (0.25)	2.600 (0.67)	0.510 (0.12)	1.553 (0.38)
$k = 3$	5.020 (0.30)	-1.008 (0.11)	1.014 (0.14)	2.539 (0.36)	0.501 (0.09)	1.516 (0.21)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.007 (0.15)	0.503 (0.09)	0.001 (0.02)	0.522 (0.15)	0.513 (0.13)	0.502 (0.07)
$N = 800$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.502 (0.14)	1.006 (0.20)	-0.998 (0.15)	1.501 (0.20)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 1$	$\beta_{k3} = 2.5$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.064 (0.70)	-1.023 (0.21)	1.026 (0.23)	2.571 (0.59)	0.508 (0.11)	1.540 (0.34)
$k = 2$	5.022 (0.53)	-1.009 (0.17)	1.014 (0.18)	2.540 (0.49)	0.503 (0.09)	1.516 (0.28)
$k = 3$	5.019 (0.24)	-1.007 (0.09)	1.008 (0.11)	2.527 (0.28)	0.502 (0.07)	1.515 (0.16)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.001 (0.12)	0.499 (0.07)	0.001 (0.02)	0.511 (0.12)	0.504 (0.10)	0.502 (0.06)

and standard error appears to be smaller for β_{k2} and β_{k3} in Table 2 as compared to Table 1 because the magnitude of the true parameter values are different between these two tables (in order to have biomarkers with different discriminating ability).

In the simulations reported in Table 1 and Table 2, although we used the same parameters to generate biomarker results for all three biomarkers, we allowed β_{kd} to be different in the estimations. Therefore, the results displayed here represent the model's performance when biomarkers had different diagnostic accuracy. The same parameters β_{kd} in the data generation were used for the sake of clarity in the presentation, and to help examine the possible effects of different transformations on the estimates. As a reference, Table 6 in Appendix B shows the results with one good biomarker and two fair biomarkers.

5.2. *Model mis-specfications.*   This section investigates the model performance under two types of mis-specified situations: deviations from the conditional independence assumption and inclusion of non-informative biomarkers.

For deviations from the conditional independence assumption, we used the same parameter specifications as Table 2 in the previous section, but adopted a multivariate error term to induce additional correlations among the biomarkers. We considered positive or negative correlation between two of the biomarkers, as well as among all three biomarkers, after adjusting for all covariates. Results on mean bias and mean square error (MSE) were

TABLE 2
Mean and standard error (in parentheses) estimates based on 1000 simulations for biomarkers with fair performance.

$N = 500$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.505 (0.25)	1.008 (0.35)	-1.002 (0.22)	1.513 (0.30)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 0.5$	$\beta_{k3} = 1$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.136 (1.08)	-1.051 (0.33)	0.527 (0.19)	1.055 (0.34)	0.520 (0.18)	1.580 (0.51)
$k = 2$	5.084 (0.89)	-1.036 (0.29)	0.513 (0.16)	1.039 (0.29)	0.519 (0.16)	1.556 (0.45)
$k = 3$	5.025 (0.41)	-1.011 (0.16)	0.510 (0.11)	1.020 (0.17)	0.502 (0.11)	1.518 (0.25)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.001 (0.18)	0.499 (0.11)	0.000 (0.03)	0.521 (0.17)	0.513 (0.15)	0.502 (0.09)
$N = 800$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.499 (0.19)	1.006 (0.27)	-0.999 (0.17)	1.502 (0.23)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 0.5$	$\beta_{k3} = 1$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.082 (0.86)	-1.030 (0.26)	0.517 (0.14)	1.035 (0.27)	0.512 (0.14)	1.548 (0.40)
$k = 2$	5.025 (0.65)	-1.012 (0.21)	0.508 (0.11)	1.016 (0.21)	0.506 (0.12)	1.521 (0.33)
$k = 3$	5.023 (0.32)	-1.001 (0.12)	0.507 (0.09)	1.010 (0.13)	0.502 (0.08)	1.518 (0.20)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.000 (0.14)	0.497 (0.08)	0.001 (0.02)	0.514 (0.14)	0.504 (0.11)	0.503 (0.07)

summarized in the top portion of Table 3. To keep the results concise, we only present the results regarding AUC values for distinguishing disease level 1 versus disease level 0, denoted by $AUC_{1vs.0}$. Results for $AUC_{2vs.1}$ are similar.

These results suggest that, residual correlations can lead to biased AUC estimates and the biases are not diminishing even with increased sample size. On the other hand, magnitude of the biases is minimal, for example, even for correlation 0.8, the average bias in AUC estimates are less than 8%, suggesting the proposed model has adequate robustness with respect to this type of violation.

We also investigated the effect of including non-informative biomarkers in the model. The results are shown in Table 3. We found that including non-informative biomarkers can lead to biased results and higher MSE. This effect is greater if more non-informative biomarkers are included. Larger sample size can help reducing these biases. Although the proposed model exhibit adequate robustness with respect to including a small number of non-informative biomarkers, we would like to emphasize that the model was not designed for variable selection. For application purposes of such model, the researchers should have adequate understanding of the biomarkers and select only the ones that have established predictive ability of the disease based on biological information or previous knowledge.

5.3. *Setting that mimics the real data application.* This section conducted simulations that intend to mimic the settings in our real data application in

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

Models	$N = 300$				$N = 500$				$N = 800$			
	T_1	T_2	T_3	iT_1^*	T_1	T_2	T_3	iT_1	T_1	T_2	T_3	iT_1
$\rho(T_1,T_2) = 0.1$	$0.25^\dagger$	0.18	-0.05	NA	0.24	0.17	-0.07	NA	0.25	0.16	-0.08	NA
	$(0.02)^\ddagger$	(0.02)	(0.02)	(NA)	(0.01)	(0.01)	(0.01)	(NA)	(0.01)	(0.01)	(0.01)	(NA)
$\rho(T_1,T_2) = 0.2$	0.26	0.26	-0.06	NA	0.25	0.25	-0.07	NA	0.25	0.25	-0.08	NA
	(0.02)	(0.02)	(0.02)	(NA)	(0.01)	(0.02)	(0.01)	(NA)	(0.01)	(0.01)	(0.01)	(NA)
$\rho(T_1,T_2)$ $= -0.2$	0.23	-0.06	-0.03	NA	0.23	-0.07	-0.05	NA	0.23	-0.08	-0.07	NA
	(0.02)	(0.02)	(0.03)	(NA)	(0.01)	(0.01)	(0.01)	(NA)	(0.01)	(0.01)	(0.01)	(NA)
$\rho(T_1,T_2) = 0.3$	0.28	0.34	-0.06	NA	0.27	0.33	-0.07	NA	0.27	0.33	-0.08	NA
	(0.02)	(0.03)	(0.02)	(NA)	(0.01)	(0.02)	(0.01)	(NA)	(0.01)	(0.02)	(0.01)	(NA)
$\rho(T_1,T_2) = 0.5$	0.32	0.51	-0.06	NA	0.31	0.51	-0.07	NA	0.31	0.50	-0.08	NA
	(0.02)	(0.04)	(0.02)	(NA)	(0.02)	(0.04)	(0.01)	(NA)	(0.01)	(0.03)	(0.01)	(NA)
$\rho(T_1,T_2) = 0.8$	0.40	0.79	-0.04	NA	0.40	0.78	-0.06	NA	0.40	0.78	-0.07	NA
	(0.03)	(0.08)	(0.03)	(NA)	(0.02)	(0.07)	(0.01)	(NA)	(0.02)	(0.07)	(0.01)	(NA)
$\rho(T_1,T_2,T_3)$ $= 0.1$	0.21	0.16	0.07	NA	0.21	0.14	0.06	NA	0.21	0.14	0.04	NA
	(0.02)	(0.02)	(0.02)	(NA)	(0.01)	(0.01)	(0.01)	(NA)	(0.01)	(0.01)	(0.01)	(NA)
$\rho(T_1,T_2,T_3)$ $= 0.3$	0.17	0.29	0.34	NA	0.18	0.28	0.33	NA	0.17	0.28	0.31	NA
	(0.02)	(0.03)	(0.04)	(NA)	(0.01)	(0.02)	(0.02)	(NA)	(0.01)	(0.01)	(0.02)	(NA)
$\rho(T_1,T_2,T_3)$ $= 0.6$	0.04	0.49	0.77	NA	0.12	0.54	0.82	NA	0.18	0.56	0.83	NA
	(0.09)	(0.07)	(0.12)	(NA)	(0.03)	(0.05)	(0.10)	(NA)	(0.01)	(0.04)	(0.08)	(NA)
$\vec{T} + iT_1 \sim X$	0.22	0.10	-0.06	-0.03	0.20	0.08	-0.05	0.00	0.14	0.08	-0.06	-0.01
	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\vec{T} + iT_1 + iT_2$ $\sim X$	-0.30	-0.13	-0.27	0.06	0.10	0.03	-0.09	0.01	0.02	-0.03	-0.16	0.02
	(0.42)	(0.14)	(0.10)	(0.03)	(0.14)	(0.05)	(0.03)	(0.02)	(0.18)	(0.06)	(0.04)	(0.01)
$\vec{T} + iT_1 + iT_2$ $+iT_3 \sim X$	-0.35	-0.15	-0.27	0.04	-0.21	-0.13	-0.22	0.03	-0.13	-0.07	-0.23	0.02
	(0.48)	(0.15)	(0.10)	(0.03)	(0.38)	(0.12)	(0.08)	(0.02)	(0.39)	(0.12)	(0.07)	(0.01)
Real data setting	-0.06	0.05	0.05	NA	-0.07	0.05	0.06	NA	-0.07	0.05	0.06	NA
	(0.00)	(0.00)	(0.00)	(NA)	(0.00)	(0.00)	(0.00)	(NA)	(0.00)	(0.00)	(0.00)	(NA)

[*] iT_k denotes the kth non-informative biomarker.

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

[‡] MSE on scale of $\times 10^{-3}$

section 6. The model has two disease levels but more covariates, including categorical and continuous. The simulation parameters are specified according to the model estimates based on real data application. Results are shown in the bottom rows of Table 3. We found that both bias and MSE are small, for example, the average bias is less than 0.6% even with a sample of size 300.

6. Example: Application to Alzheimer's Disease Data. AD is a progressive and fatal neurodegenerative disorder that affected 5.2 million people in the United States in 2013. Without advanced therapy, this number is predicted to rise to 13.8 million by 2050 (Hebert et al., 2013). The reason for the current lack of effective AD therapy is the decades-long pre-clinical stage where too much neurodegeneration occurs before any clinical diagnosis can be made. Therefore, great effort has been put into preclinical

AD detection. Studies suggested that the CSF biomarkers total tau (t-tau), phosphorylated tau (p-tau$_{181p}$), and amyloid β 1-42 (Aβ_{42}) changes about 15 years before clinical AD onset and are particularly relevant to tracking the pathological onset and preclinical stages of AD (Bateman et al., 2012). Appropriate assessment and effective combination of these biomarkers for early AD detection can provide a critical opportunity for therapeutic intervention, disease monitoring and to lessen the time and cost for clinical trials. We applied the proposed method to avoid using imperfect clinical diagnoses and to take into account AD risk factors and patients' characteristics that may affect biomarker levels.

Our analyses used data from the Alzheimer's Disease Neuroimaging Initiative (ANDI) (http://adni.loni.ucla.edu/). The ADNI study is a multicenter longitudinal observation study launched in 2004 in the United States and Canada. It is conducted to examine whether biomarkers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early AD patients. ADNI enrolled subjects from three groups: cognitively normal (CN) subjects, subjects with MCI, and subjects with early AD based on clinical assessments. Detailed information about the study's design and inclusion and exclusion criteria can be found at http://www.adni-info.org. Briefly, all participants were recruited between the ages of 55 and 90 years and had at least 6 years of education. Participants who took specific psychoactive medications, or who had other neurological disorders, were excluded. CSF biomarkers, including Aβ_{42}, t-tau, and p-tau$_{181p}$, along with subjects' demographic and clinical information were obtained at baseline. After the baseline visit, subsequent visits occurred at 6- or 12-month intervals but mostly without additional CSF measurements.

We applied the proposed model to compare and combine CSF biomarkers, Aβ_{42}, t-tau, and p-tau$_{181p}$, for detecting preclinical AD related pathophysiological changes, which we will call the AD signature. Our analysis used a cross sectional dataset at baseline of all currently available ADNI subjects as of April 7th, 2015 with at least one CSF biomarker measurement. There were 808 subjects (223 CN, 457 MCI and 127 early AD). Two latent disease groups were used for this problem, defined by the presence or absence of an AD signature. CSF Aβ_{42}, t-tau, and p-tau$_{181p}$ are believed to reflect different processes in AD pathology (Storandt et al., 2012), so we assumed the correlation among them within subjects with the same AD signature status can be explained by covariates. Covariates in the latent structure model included risk factors age, presence of ApoE 4 alleles and education. Covariates in the measurement model were age, gender and the interaction term between age

and D. Thus, age could have different impacts on biomarkers among subjects with or without an AD signature. We used the Box-Cox transformation for $H_k(\cdot; \lambda_k)$. A total of 790 subjects with complete biomarker and covariate information were included in the analysis. Results were shown in Table 4, where $D1$ stands for $D = 1$, denotes the group with an AD signature. Confidence intervals are obtained based on nonparametric bootstrap (Efron, 1981) with 1000 bootstrap replications. We note here that bootstrap confidence intervals are approximation of the actual confidence intervals (Efron, 1987).

TABLE 4

Estimates and 95% bootstrap CI (in parentheses) for CSF $A\beta_{42}$, t-tau and p-tau$_{181p}$ (Estimates in bold indicate a significant effect).

	Structural (prevalence) model $P(D \mid Z)$			
	(Intercept)	Age($\times 10$)	ApoE4	Education
$D = 1$	**-5.52** (-7.75, -3.33)	**0.67** (0.42, 0.93)	**2.60** (2.23, 3.03)	-0.03 (-0.09, 0.04)
	Measurement model $P(T \mid D, X)$			
	$A\beta_{42}$		T-tau	P-tau$_{181p}$
(Intercept)	**13.88** (7.96, 25.19)		**2.40** (2.07, 2.87)	**2.87** (2.40, 3.52)
D1	**-1.83** (-5.26, -0.23)		**0.81** (0.43, 1.36)	**1.04** (0.54, 1.72)
Age($\times 10$)	-0.11 (-0.04, 0.45)		**0.05** (0.02, 0.09)	-0.01 (-0.06, 0.03)
Gender=Male	-0.15 (-0.51, 0.02)		**-0.03** (-0.07, -0.01)	-0.02 (-0.07, 0.03)
D1$\times$Age($\times 10$)	-0.10 (-0.41, 0.11)		**-0.07** (-0.12, -0.03)	**-0.06** (-0.14, -0.00)
λ	0.33 (0.15, 0.49)		-0.19 (-0.30, -0.09)	-0.07 (-0.17, 0.04)
σ	1.10 (0.45, 2.51)		0.17 (0.11, 0.27)	0.31 (0.21, 0.44)

These results are consistent with current knowledge that older age and ApoE 4 allele are risk factors for AD. Specifically, the results suggest that the log odds of having an AD signature is about 0.7 (95% bootstrap CI: 0.4, 0.9) higher for a 10-year increase in age, adjusting for ApoE 4 and education. Subjects with ApoE 4 alleles are estimated to have about exp(2.60)=13.5 fold higher odds of developing an AD signature (95% bootstrap CI: 9.3, 20.7), adjusting for age and education. Additionally, education seems to have a protective effect against AD related pathophysiological changes, but this effect was not significant based on these data. On the biomarker side, we found that subjects with AD pathology have significantly lower levels of $A\beta_{42}$ and higher levels of t-tau and p-tau$_{181p}$, consistent with the amyloid hypothesis. Normal aging process is revealed by lower level of $A\beta_{42}$ and p-tau$_{181p}$, and higher levels of t-tau in older adults, suggesting that different diagnostic criteria should be considered for different age groups. Males apprear to have lower levels than women for all three biomarkers. In addition, the estimates for λ were about 0.33 for $A\beta_{42}$, -0.19 and -0.07 for t-tau and p-tau$_{181p}$, which challenge the conventional assumption that $A\beta_{42}$ has normal distribution, and that tau measures are log normal.
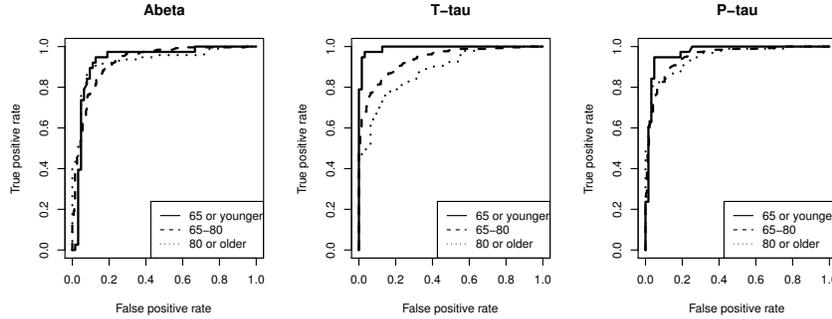
FIG 1. *Age-specific ROC curves for CSF $A\beta_{42}$, t-tau and p-tau$_{181p}$ in detecting an AD signature*

Based on the model, one can compute the estimated covariate-specific ROC curves. As an example, we plotted the estimated age-specific ROC curves for CSF biomarker $A\beta_{42}$, t-tau and p-tau$_{181p}$ in Figure 1. The plots suggest that the ROC curves differ among age groups for both t-tau and p-tau$_{181p}$, with higher AUCs for younger groups. This results also suggest that including covariates in the model and allowing for the covariate effect to vary between disease groups are important. On the other hand, the age-specific ROC curves for $A\beta_{42}$ do not vary much among these age groups. One explanation is that $A\beta_{42}$ changes much earlier than tau in AD pathophysiology; it is possible that, at this stage it is already relatively stabilized.

Combinations of CSF biomarkers have not been well examined due to the unobserved AD signature. Current research mostly uses CSF $A\beta$ or tau protein separately or CSF tau/$A\beta$ ratio as measures to classify subjects with or without an AD signature. In addition to assess the diagnostic accuracy of these biomarkers, our model suggested a way to synthesize information from different biomarkers to reach a potentially better diagnosis. Specifically, we can use the model based risk $P(D_i \mid \vec{T}_i, \vec{X}_i, \vec{Z}_i, \theta)$ as a score for diagnosing whether AD related pathophysiological change has started. This risk score summarizes diagnostic information from multiple biomarkers, risk factors and subjects' characteristics, and therefore offers a more personalized diagnosis.

Due to the unobserved gold standard, we evaluated the risk scores against clinical diagnoses at the last available follow-up using a subset of individuals who enrolled during the first phase of ADNI study (ADNI1: 2004-2010) and thus have relatively longer follow-up. The risk score results in an AUC of 0.78, higher than that based on any one of the biomarkers ($A\beta$: 0.75, t-tau:

0.73, p-tau$_{181p}$: 0.72). The improvement is more evident at the clinically relevant region of the ROC plot, where the false positive rate is low and true positive rate is high (plot not shown). For example, when choosing cut-points that give 0.8 in specificity, sensitivity for detecting an AD signature based on Aβ_{42}, t-tau or p-tau$_{181p}$ is 0.59, 0.55 and 0.57, respectively; while sensitivity based on the risk score is 0.69. It is worth noting that the reference standard used here, the clinical diagnosis, is imperfect. This can mask the good performance of the biomarkers especially when they are in fact superior to the clinical diagnosis (Pepe, 2003). Additionally, since the clinical diagnosis is likely to miss preclinical AD subjects without manifested symptoms, it will erroneously decrease the specificity estimate of the biomarkers. Nevertheless, these results partially validate the advantage of biomarker combination based on the proposed model. Additionally, we found that the risk scores are well-separated in the population, as shown in Figure 2. The histogram of the risk scores put most of the subjects either in the very low risk end or in the very high risk end of the distribution. Only 7.7% of the subjects had an estimated risk between 0.2 and 0.8, or 13.2% between 0.1 and 0.9. This suggests that the model based-risk had good discriminant ability to separate subjects with and without an AD signature. By contrast, the histograms for single biomarkers put most subjects in the middle range.
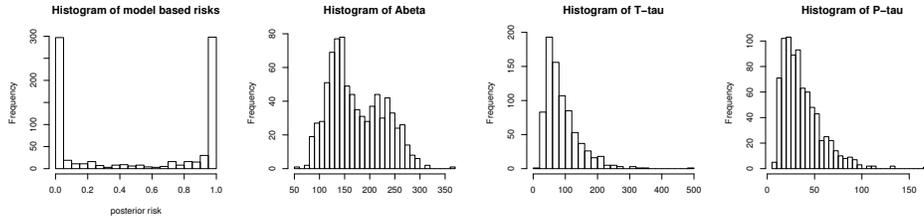

Fig 2. *Histograms of model based combined score and single biomarkers*

We also compared the estimated proportions of subjects that have an AD signature (according to model estimates) with their clinical diagnoses. The results are shown in Table 5. It is evident that the proportion of subjects with an AD signature increased from about 31% in the early MCI group to 86% in the AD group, representing a progression. Additionally, there were about 26% of subjects in the CN group that had an AD signature according to the model. This portion likely indicates that the subjects exhibit preclinical AD pathophysiological changes but still appear to have normal cognition.

TABLE 5
Numbers and proportions of subjects that have an AD signature by groups with different clinical diagnoses.

Clinical diagnosis	AD signature present	AD signature absent
CN	57 (26.3%)	160 (73.7%)
EMCI[1]	58 (30.7%)	131 (69.3%)
LMCI[2]	168 (64.9%)	91 (35.1%)
AD	107 (86.3%)	17 (13.7%)

1/2: Early/Late MCI–meet ADNI MCI criteria, objective memory loss measured by education adjusted scores on delayed recall of one paragraph from Wechsler Memory Scale Logical Memory II ($\geq$16 years: 9-11; 8-15 years: 5-9; 0-7 years: 3-6) for EMCI, ($\geq$16 years:$\leq$8; 8-15 years: $\leq$4; 0-7 years: $\leq$2) for LMCI.

7. Summary and Discussion. In this paper we proposed a latent profile model for assessing the accuracy of continuous biomarkers or diagnostic tests without a gold standard. We did not discuss the choice of number of classes, because in diagnostic testing studies, the number of disease classes is often apparent and governed by the scientific questions. Compared with currently available methods, this model has several advantages. First, the model can include covariates in both the latent structure model and the measurement model. The former can help to examine the impact of risk factors on constructs that are difficult to measure, such as the pathophysiological process in a preclinical AD brain. The latter can be helpful in understanding biomarkers' performances in subgroups of patients with different characteristics. This is especially valuable for obtaining a more personalized diagnostic procedure and for achieving higher diagnostic accuracy for each individual subject. Second, the model for covariate effect is much more flexible – there is no constraint on covariate type (i.e., categorical or continuous); additionally, covariates can have different impacts on different biomarkers and within different disease severity groups. Third, the transformation on the biomarker values attempts to account for possible skewness. Last, the method can deal with categorical disease status. An R package "latentreg" has been developed for the proposed methods to accommodate more general data structures and research questions.

We present the model in the situation where the latent disease status is binary or ordinal. However, this assumption is only used to identify a unique solution among several equivalent ones caused by the "label switching" problem in latent class or latent profile models. There is no additional constraint imposed for the estimation. Therefore, this assumption is not essential if there are other pieces of information that can help to determine the labeling of the latent disease groups – for example, the group labels can be determined if the sizes of the latent groups are known. As a result, when

additional information is available to label the groups, the proposed methods can be applied to situations where the latent disease status is nominal. In this case, different intervals of the biomarker values do not translate to severity information but indicate nominal disease status, such as subtypes. As a result, it is possible for subjects with higher or lower biomarker levels to have easy to treat disease subtypes than whose with intermediate biomarker level, and vice versa.

When applying latent variable models, one should realize that all these models merely offer a tool to "cluster" the population into groups in which the manifest variables (biomarkers) are relatively homogenous. These models cannot directly identify disease groups. In other words, these models do not guarantee the resulting clusters are grouped according disease status of interest. In our application, we chose biomarkers that are biologically relevant to AD pathophysiology, so that disease status was the fundamental factor that lead to the heterogeneity of the biomarker values. However, one should carefully evaluate this assumption in each application.

One possible extension to the proposed method is to use a semi-parametric transformation with an unspecified H instead of a Box-Cox transformation. In this case, the transformation H is a nuisance parameter with infinite dimensions, the maximum likelihood approach did not apply. One possible choice is to use estimating equations in each M step to estimate H. This type of algorithm (sometimes referred to as an EM-like algorithm) has been used in practice and appears to lead to reasonable results (Benaglia et al., 2009). However, one need to prove its convergence and consistency of the results since estimating equations uses a different criterion than maximizing the likelihood and no longer guarantee the non-decreasing property in a typical EM algorithm.

The proposed model assumes that the dependence among biomarkers can be explained by disease status and other covariates. Therefore, another extension to the model is to introduce random effects if one suspects that there are remaining correlations due to unobserved covariates. While more complicated model structures may be relatively easy and perhaps tempting to adopt (aside from the potential computational burden), one should be cautious when doing so as they are often hard to validate. Additionally, the identifiability issue for such models must be resolved before adopting them in practice.

APPENDIX A

An EM algorithm. E step computes the expectation of the complete data log likelihood given the observed data:

$$
E[l_c^{(t)}(\boldsymbol{\theta}) \mid \vec{T}_i, \vec{X}_i, \vec{Z}_i, \boldsymbol{\theta}^{(t)}] = \sum_{i=1}^{N}\sum_{d=0}^{L-1} P_i^{(t)}(d)\log\eta(\vec{z}_i^T\boldsymbol{\alpha}_d^{(t)}) + \sum_{i=1}^{N}\sum_{k=1}^{K}\log J(t_{ik};\lambda_k^{(t)})
$$

$$
(7.1) \qquad\qquad + \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\log g_k^{(t)}(H_k^{(t)}(t_{ik}) - \vec{x}_i^T\boldsymbol{\beta}_{kd}^{(t)}),
$$

where superscript (t) denotes iteration number, $g_k^{(t)}(\cdot) = g_k(\cdot;\nu_k^{(t)})$, $H_k^{(t)}(\cdot) = H_k(\cdot;\lambda_k^{(t)})$, and $P^{(t)}(d)$ is the expected value of $I_d(D_i)$ at current parameter values,

$$
P_i^{(t)}(d) \equiv E\big[I_d(D_i) \mid \vec{T}_i, \vec{X}_i, \vec{Z}_i, \boldsymbol{\theta}^{(t)}\big]
$$

$$
= \frac{\big[\prod_{k=1}^{K} g_k^{(t)}\big(H_k^{(t)}(t_{ik}) - \vec{x}_i^T\boldsymbol{\beta}_{kd}^{(t)}\big)\big]\eta_d(\vec{z}_i^T\boldsymbol{\alpha}^{(t)})}{\sum_{d=0}^{L-1}\big\{\prod_{k=1}^{K}\big[g_k^{(t)}\big(H_k^{(t)}(t_{ik}) - \vec{x}_i^T\boldsymbol{\beta}_{kd}^{(t)}\big)\big]\eta_d(\vec{z}_i^T\boldsymbol{\alpha}^{(t)})\big\}}.
$$

M step maximizes the expected complete data log likelihood given in (7.1). Let

$$
l_1(\boldsymbol{\alpha}_d^{(t+1)}) \, (7.2) \sum_{i=1}^{N}\sum_{d=0}^{L-1} P_i^{(t)}(d)\log\eta(\vec{z}_i^T\boldsymbol{\alpha}_d^{(t+1)}),
$$

$$
l_2(\lambda_k^{(t)}, \boldsymbol{\beta}_{kd}^{(t)}, \nu_k^{(t)}) \, (7.3) \sum_{i=1}^{N}\sum_{k=1}^{K}\log J(t_{ik};\lambda_k^{(t)}) + \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\log g_k^{(t)}(H_k^{(t)}(t_{ik}) - \vec{x}_i^T\boldsymbol{\beta}_{kd}^{(t)}).
$$

Then the function to be maximized is $l_1(\boldsymbol{\alpha}_d^{(t)}) + l_2(\lambda_k^{(t)}, \boldsymbol{\beta}_{kd}^{(t)}, \nu_k^{(t)})$. Since parameters $\boldsymbol{\alpha}_d^{(t)}$ are only involved in $l_1(\boldsymbol{\alpha}_d^{(t)})$, and parameters $\lambda_k^{(t)}$, $\boldsymbol{\beta}_{kd}^{(t)}$

and $\nu_k^{(t)}$ are only involved in $l_2(\lambda_k^{(t)}, \boldsymbol{\beta}_{kd}^{(t)}, \nu_k^{(t)})$, the maximization can be broken down into two parts:

$$
\begin{aligned}
\{\alpha_d^{(t+1)}\}_{d=0,\ldots,L-1} &= \operatorname{argmax} l_1(\alpha_d^{(t)}), \\
\{\lambda_k^{(t+1)}, \beta_{kd}^{(t+1)}, \nu_k^{(t+1)}\}_{\substack{d=0,\ldots,L-1, \\ k=1,\ldots,K}} &= \operatorname{argmax} l_2(\lambda_k^{(t)}, \beta_{kd}^{(t)}, \nu_k^{(t)}).
\end{aligned}
$$

A close examination of equation (7.2) reveals that $l_1(\boldsymbol{\alpha}_d^{(t)})$ has the same form of the log likelihood function of a polytomous regression with N observation L categories, except that the group indicator $I_d(D_i)$ is replaced by its expected value $P^{(t)}(d)$. Since $\sum_{d=0}^{L-1} P^{(t)}(d) = 1$, $l_1(\boldsymbol{\alpha}_d^{(t)})$ is a proper log likelihood function. Additionally, $l_1(\boldsymbol{\alpha}_d^{(t)})$ is concave, thus its maximization can be carried out by borrowing a polytomous regression routine that exists in many statistical softwares. Our computation utilizes the R function "multinom" in the package "nnet".

The second part of the maximization function, $l_2(\lambda_k^{(t)}, \boldsymbol{\beta}_{kd}^{(t)}, \nu_k^{(t)})$, given in equation (7.3), is also a proper log likelihood function. In fact, it is the log likelihood function of a transformation regression model with $N \times L \times K$ observations, weights $P_i^{(t)}(d)$, transformation $H_k^{(t)}(\cdot)$ and error distribution $G_k^{(t)}(\cdot)$. In fact, the original $N \times K$ observations can be considered replicated L copies with 1 copy for each disease severity group. This log likelihood function takes the total $N \times L \times K$ observations, but weights them according to the probability of a subject belonging to a disease severity group $P_i^{(t)}(d)$, computed in the E step. This is clearer with matrix notation. Assume that $\mathbb{X}(D)$ is the design matrix defined by the measurement model $P(T_k \mid D, \mathbf{X})$, including all covariates $\mathbf{X}$ and their interactions with D. Create a stacked design matrix $\mathbb{X}_{st}$ with $N \times L$ rows, $\mathbb{X}_{st} = [\mathbb{X}(D = 0)', \ldots, \mathbb{X}(D = L - 1)']'$, where superscript $'$ denotes matrix transpose. Similarly, created a stacked outcome vector $Y_{k,st}^{(t)}$ with $N \times L$ elements, $Y_{k,st}^{(t)} = (Y_k^{(t)'}, \ldots, Y_k^{(t)'})'$, where $Y_k^{(t)}$ is a vector with N elements of the transformed outcomes $H_k^{(t)}(t_{ik})$ at the tth EM iteration. In addition, let $\mathbb{W}$ be a $N \times L$ by $N \times L$ diagonal matrix $\mathbb{W}^{(t)} = diag\{P^{(t)}(0), \ldots, P^{(t)}(L-1)\}$, where $P^{(t)}(d) = (P_1^{(t)}(d), \ldots, P_N^{(t)}(d))$, $d = 0, \ldots, L-1$. Then equation (7.3) can be rewritten as follows:

$$
(7.4) \quad l_2 = \sum_{k=1}^{K} \mathbb{W}^{(t)} \log g_k^{(t)}(Y_{k,st}^{(t)} - \mathbb{X}_{st} \boldsymbol{\beta}_{kd}^{(t)}) + \sum_{k=1}^{K} \log J(t_{ik}; \lambda_k^{(t)}).
$$

Maximization can be performed with these stacked matrices as if D is known.

The above formulation with a design matrix also makes it easy to impose constraints on the covariate effect $\boldsymbol{\beta}_{kd}$. For example, if some of the covariates are assumed to have the same effect across disease groups, one can simply remove the corresponding interaction terms between D and these covariates in $\mathbb{X}_{st}$. Similarly, if the design matrix and outcome vectors for each k are stacked, and the interaction terms between the covariate effects and the biomarkers are included, constraints on the covariate effects across different biomarkers can easily be imposed. Here, we do not further complicate the design matrix since the constraints on covariate effects across different biomarkers are used less commonly.

APPENDIX B

TABLE 6

Mean and standard error (in parentheses) estimates based on 1000 simulations for biomarkers with different performances.

$N = 500$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.502 (0.21)	1.007 (0.30)	-1.005 (0.20)	1.512 (0.26)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 1$	$\beta_{k3} = 2.5$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.127 (0.95)	-1.045 (0.29)	1.049 (0.31)	2.641 (0.81)	0.519 (0.15)	1.575 (0.46)
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 0.5$	$\beta_{k3} = 1$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 2$	5.075 (0.85)	-1.032 (0.28)	0.513 (0.15)	1.034 (0.28)	0.516 (0.15)	1.550 (0.43)
$k = 3$	5.020 (0.38)	-1.008 (0.15)	0.511 (0.10)	1.017 (0.16)	0.500 (0.10)	1.514 (0.24)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.004 (0.16)	0.499 (0.11)	0.000 (0.03)	0.520 (0.16)	0.512 (0.15)	0.502 (0.08)

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(2):427–435.

Albert, P. S., McShane, L. M., and Shih, J. H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, 57(2):610–619.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Bateman, R. J., Xiong, C., Benzinger, T. L., Fagan, A. M., Goate, A., Fox, N. C., Marcus, D. S., Cairns, N. J., Xie, X., Blazey, T. M., et al. (2012). Clinical and biomarker changes in dominantly inherited alzheimer's disease. *New England Journal of Medicine*, 367(9):795–804.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009). An em-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Branscum, A. J., Johnson, W. O., Hanson, T. E., and Baron, A. T. (2015). Flexible regression models for roc and risk analysis, with or without a gold standard. *Statistics in medicine*, 34(30):3997–4015.

Branscum, A. J., Johnson, W. O., Hanson, T. E., and Gardner, I. A. (2008). Bayesian semiparametric roc curve estimation and disease diagnosis. *Statistics in Medicine*, 27(13):2474–2496.

Cheng, R. and Traylor, L. (1995). Non-regular maximum likelihood problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–44.

Collins, J. and Huynh, M. (2014). Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in medicine*, 33(24):4141–4169.

Cook, R. J., Ng, E., and Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent markov models. *Biometrics*, 56(4):1109–1117.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Hebert, L. E., Weuve, J., Scherr, P. A., and Evans, D. A. (2013). Alzheimer disease in the united states (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783.

Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, pages 167–171.

Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128.

Janes, H. and Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96(2):371–382.

Jones, G., Johnson, W., Vink, W., and French, N. (2012). 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*, 68(2):371–379.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. New York: Houghton Mifflin.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*, volume 31. Springer.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, 21(4):331–347.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Pfeiffer, R. M., Carroll, R. J., Wheeler, W., Whitby, D., and Mbulaiteye, S. (2008). Combining assays for estimating prevalence of human herpesvirus 8 infection using multivariate mixture models. *Biostatistics*, 9(1):137–151.

Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, pages 797–810.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239.

Selkoe, D. J. (1991). The molecular pathology of alzheimer's disease. *Neuron*, 6(4):487–498.

Storandt, M., Head, D., Fagan, A. M., Holtzman, D. M., and Morris, J. C. (2012). Toward a multifactorial model of alzheimer disease. *Neurobiology of aging*, 33(10):2262–2271.

van Smeden, M., Naaktgeboren, C. A., Reitsma, J. B., Moons, K. G., and de Groot, J. A. (2013). Latent class models in diagnostic studies when there is no reference standarda systematic review. *American journal of epidemiology*, page kwt286.

Wang, Z. (2013). *Latent Class and Latent Profile Analysis in Medical Diagnosis and Prognosis*. PhD thesis, University of Washington.

Wang, Z. and Zhou, X.-H. (2012). Random effects models for assessing diagnostic accuracy of traditional chinese doctors in absence of a gold standard. *Statistics in medicine*, 31(7):661–671.

Wang, Z. and Zhou, X.-H. (2014). Nonparametric identifiability of finite mixture models with covariates for estimating error rate without a gold standard. *UW Biostatistics Working Paper Series*. Working Paper 403.

Wang, Z., Zhou, X.-H., and Wang, M. (2011). Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard. *Biostatistics*, 12(3):567–581.

Wu, Z., Deloria-Knoll, M., Hammitt, L. L., and Zeger, S. L. (2016). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):97–114.

Zhou, X.-H., Castelluccio, P., and Zhou, C. (2005). Nonparametric estimation of roc curves in the absence of a gold standard. *Biometrics*, 61(2):600–609.

Division of Biostatistics and Bioinformatics
Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University, Baltimore, MD
Department of Biostatistics
Johns Hopkins University, Baltimore, MD
E-mail: wangzy@jhu.edu

Department of Biostatistics
University of Washington, Seattle, WA
Northwest HSR&D Center of Excellence
VA Puget Sound Health Care System, Seattle, WA
E-mail: azhou@uw.edu