Published in final edited form as:

J Am Stat Assoc. 2015 September 1; 110(511): 923–934. doi:10.1080/01621459.2015.1023806.

An Integrated Bayesian Nonparametric Approach for Stochastic and Variability Orders in ROC Curve Estimation: An Application to Endometriosis Diagnosis

Beom Seuk Hwang [Postdoctoral Fellow] and Zhen Chen [Investigator]

Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892

Beom Seuk Hwang: beomseuk.hwang@nih.gov; Zhen Chen: chenzhe@mail.nih.gov

Abstract

In estimating ROC curves of multiple tests, some *a priori* constraints may exist, either between the healthy and diseased populations within a test or between tests within a population. In this paper, we proposed an integrated modeling approach for ROC curves that jointly accounts for stochastic and variability orders. The stochastic order constrains the distributional centers of the diseased and healthy populations within a test, while the variability order constrains the distributional spreads of the tests within each of the populations. Under a Bayesian nonparametric framework, we used features of the Dirichlet process mixture to incorporate these order constraints in a natural way. We applied the proposed approach to data from the Physician Reliability Study that investigated the accuracy of diagnosing endometriosis using different clinical information. To address the issue of no gold standard in the real data, we used a sensitivity analysis approach that exploited diagnosis from a panel of experts. To demonstrate the performance of the methodology, we conducted simulation studies with varying sample sizes, distributional assumptions and order constraints. Supplementary materials for this article are available online.

Keywords

Area under the curve; Dirichlet process mixture; Gold standard; Order restricted analysis

1. Introduction

The receiver operating characteristic (ROC) curve is a graphical tool to assess the performance of continuous diagnostic tests in distinguishing between diseased and healthy populations. The theoretical ROC curve is a plot of all possible pairs of true positive rate (sensitivity) versus false positive rate (1-specificity) across all threshold values. A useful index of diagnostic accuracy is the area under the ROC curve (AUC), which can be interpreted as the probability that a randomly selected diseased subject has a higher value than a randomly selected healthy subject. Models for ROC curves are abundant, and range from parametric (Hanley, 1996; Pepe, 1998; Faraggi and Reiser, 2002; Sorribas, March and Trujillano, 2002; Pepe, 2003) to nonparametric (Hsieh and Turnbull, 1996; Lloyd, 1998; Pepe, 2000; Pepe, 2003; Peng and Zhou, 2004; Cai and Moskowitz, 2004), to Bayesian (Peng and Hall, 1996; Gu, Ghosal and Roy, 2008; Gu and Ghosal, 2009), and cover both the case with and without a gold standard for the disease (Jafarzadeh et al., 2010; Tang et al.,

2013). More recently, nonparametric Bayesian approaches have also been developed (Branscum, Gardner and Johnson, 2005; Erkanli et al., 2006; Branscum et al., 2008; Hanson, Branscum, and Gardner, 2008; Cheng et al., 2012; De Carvalho et al., 2013; Rodriguez and Martinez, 2014).

In the application of ROC curves, it's often the case that some *a priori* constraints may exist. It has been well established in the literature that when such constraints are available, incorporating them in the estimation can improve the statistical efficiency of the estimates and even result in different estimates that conform with biological theory or expert consensus (Robertson, Wright and Dykstra, 1988). One commonly encountered constraint is that the diseased population tends to have larger test scores on average than the healthy population. Hanson, Kottas and Branscum (2008) modeled this constraint by a stochastic order in the healthy and diseased distributions. A less studied constraint may occur in multiple tests in that one test score tends to have larger variability than another. While these two orders have been studied separately in ROC curves estimation (Gelfand and Kottas, 2001; Kottas and Gelfand, 2002; Dunson and Peddada, 2008; Hanson, Kottas and Branscum, 2008), joint modeling of them does not exist to our best knowledge. This paper proposes an integrated approach to jointly model stochastic and variability orders.

This research was motivated by the Physician Reliability Study (PRS) (Schliep et al., 2012) that examined the diagnostic accuracy of diagnosing endometriosis, a women's health disorder that occurs when cells from the lining of the uterus grow in other parts of the body. With 12 physicians in obstetrics and gynecology (OB/GYN) invited to review clinical information of 156 participating women, the PRS provides the revised American Society for Reproductive Medicine (rASRM) score that ranges from 1 to 159. Higher rASRM score is indicative of higher likelihood of an endometriosis disease. The 12 physicians consist of three groups of obstetricians and gynecologists: the 4 international experts (IEs) are physicians with international reputation in the field and have considerably extensive experience in laparoscopy and endometriosis diagnosis, the 4 regional experts (REs) are physicians affiliated with Utah Medical Center with extensive experience in the field, and the 4 residents (RDs) are new graduates from medical schools conducting guided practice at Utah Medical Center. The reviewing and diagnosis were conducted at four different settings, with setting 1 having clinical information consisting of only the digital intrauterus image taken during a laparoscopy procedure. At setting 2, the clinical information was augmented to include surgeon's notes. MRI and histopathology reports were further successively added to the clinical information pool at settings 3 and 4, respectively. In ordering the four settings this way, the PRS attempts to follow the usual practice in the field: a laparoscopy is usually done and the surgeon provides a diagnosis of endometriosis upon seeing the inside of the uterus (settings 1 and 2). When ambiguity still exists, other tests such as MRI (setting 3) and histopathology (setting 4) are then conducted to aid the diagnosis. One research question of interest is to estimate the different diagnostic accuracy of the group of 4 regional experts at settings 1 and 2. These REs are of interest because they represent everyday physicians in practice; we focus on settings 1 and 2 as they are the most commonly used steps in diagnosing endometriosis and as only a small number of women have MRI and histopathology results.

In trying to estimate the two ROC curves of the REs at settings 1 and 2, two constraints were brought up by our expert collaborators. First, within each setting, the rASRM scores should be larger in the diseased population than in the healthy population. To reflect this feature, a stochastic order constraint is assumed between the two populations. Secondly, with the study design that more clinical information is available at setting 2 than 1, the variability in the rASRM scores within each population should be smaller at setting 2 than 1. According to our collaborators, this variability order constraint is sensible because the physicians are able to make fewer guesses in checking the various components of the rASRM scoring sheet that produced the rASRM score when these physicians have more clinical information. In this paper, we use the variability order notion proposed by Kottas and Gelfand (2002) due to its nice connection to the Bayesian nonparametric framework. Our interest is to account for these two *a priori* constraints simultaneously when we estimate the ROC curves for the group of REs.

There are other features of the PRS data that make the analysis challenging. First, the rASRM scores can be non-normal, even after taking the logarithm transformation (see Figure 1). This warrants some flexible models for the scores that can accommodate the skewness and possible bimodality. To handle this feature, we propose to model these scores using a Bayesian nonparametric approach. More specifically, we will utilize the Dirichlet process mixture (DPM) for the rASRM score on the logarithm scale so that it affords the needed flexibility in the distributions. The DPM is also desirable as it enables us to extend the work of Gelfand and Kottas (2001) and Kottas and Gelfand (2002) to jointly model the stochastic and variability orders.

Another challenge in the analysis is that there does not exist a gold standard for endometriosis in the PRS data. To address this issue, we exploit the fact that four of the 12 physicians are international experts (IEs) who tend to have greater experience in diagnosing endometriosis. The diagnostic results of these 4 IEs can be exploited to create an imperfect reference standard. We will propose a sensitivity analysis scheme and study the robustness of this mechanism by varying the sensitivity parameters in a scientifically reasonable way. This is an extension of the sensitivity approach introduced by Zhang, Chen and Albert (2012).

The rest of the paper is organized as follows. In Section 2, we take an initial look at the PRS data. In Section 3, we introduce the ROC curve, order constraints and Dirichlet process mixture models, and propose our integrated model of stochastic and variability orders using nonparametric Bayesian methods. Section 4 specifies priors and presents posterior computations of models developed in Section 3 using Markov chain Monte Carlo (MCMC) methodology. In Section 5, we apply our model to the PRS data and discuss the substantive results. In this section, we also consider the sensitivity analysis approach that exploits IEs' imperfect reference standard to address the issue of not having a gold standard. In Section 6, we conduct simulation studies to illustrate the proposed model and to evaluate their performance. We conclude with a discussion of our proposed approach and future work in Section 7.

2. An Initial Look at the Physician Reliability Study Data

In the PRS, the 12 invited physicians consist of 4 international experts (IEs), 4 regional experts (REs) and 4 residents. In this paper, we focus on the diagnostic accuracy of the REs in diagnosing endometriosis at settings 1 and 2, separately, using their average rASRM scores. Averaging REs' scores is reasonable here because the 4 REs are thought to be exchangeable as they all have similar years of experience and have been affiliated with the same institute. Out of the 156 women participants in the PRS, 113 at setting 1 and 126 at setting 2 have non-missing data in the average REs' rASRM scores and form our working datasets.

Table 1, Figures 1 and 2 provide descriptive statistics of the rASRM scores based on these data. Several observations can be made:

1. The rASRM scores within each setting are generally skewed and possibly have multimodality. The logarithm transformation does not help in achieving symmetry or uni-modality (Figure 1);
2. These scores are still non-symmetric and multi-modal within each setting by disease status (Figure 2), where disease status is estimated by posterior modes from models specified in Section 5;
3. Healthy and diseased populations combined, the scores are lower and have smaller variance in setting 1 than 2 (Table 1);
4. Stratified by disease status, the scores are higher in their means in the diseased than the healthy population in both settings; within each population, the variances are higher in setting 2 than 1 (Table 1).

Although there is no gold standard in the PRS for diagnosing endometriosis, the 4 IEs' diagnosis can be exploited to provide an imperfect reference standard. In Section 5, we propose a sensitivity analysis approach to make use of the IEs' diagnostic results to aid the estimation of the REs' ROC curves. The 4 IEs made 1/0 (presence/absence) endometriosis diagnosis at each setting for each woman and we use their proportion of positive diagnosis in the sensitivity analysis (see Section 5 for details). The relationship between REs' rASRM scores and the IEs' proportion of positive diagnosis is presented in Figure 3 for settings 1 and 2, separately and combined. We observed that in general, higher REs' scores are associated with higher proportion of IEs' positive diagnosis, as expected.

As an initial look, we fit both a binormal model (Pepe, 2003) and a bi-DPM model where scores from the healthy and diseased populations are each modeled using a Dirichlet process mixture (Erkanli et al., 2006). The "gold standard" for this initial analysis is constructed by IEs' proportion of positive diagnosis in settings 1 and 2 combined as follows: a woman is assigned a presence of endometriosis if the proportion is greater than 0.5, and absence if less than 0.5. In the case of a proportion of 0.5, a one-time random draw from *Bernoulli*(0.5) is made and considered as her disease status. The results are demonstrated in Figure 4, where both ROC curves and their corresponding AUCs are presented. From the initial analysis, we saw that setting 1 has a higher AUC than setting 2 in both the binormal and bi-DPM models. AUCs in the bi-DPM models tend to be smaller than their counterparts in the binormal

models, but this is likely due to the different distributional assumptions used to model the rASRM scores.

This initial analysis does not incorporate the constraints and assumes a known gold standard. The following two sections propose methodologies to address these two issues.

3. Methods

3.1 ROC Curve and AUC

Let F_H and F_D denote the cumulative distribution functions (CDFs) of healthy and diseased populations, respectively. The ROC curve is defined by $ROC(u) = 1 - F_D\{F_H^{-1}(1-u)\}$, for $0 \leq u \leq 1$. A useful index of diagnostic accuracy is the area under the ROC curve (AUC), which has the definition of $AUC = \int_0^1 ROC(u) du$. Bamber (1975) has shown that $AUC = Pr(Y^D > Y^H)$, where Y^D and Y^H denote the measurements for diseased and healthy populations, respectively.

A binormal ROC curve is a widely used parametric method, and assumes that test results follow normal distributions for the healthy and diseased populations after an unspecified transformation S : $S(Y^H) \sim N(0, 1)$ and $S(Y^D) \sim N(\mu, \sigma^2)$. Then, the functional form of the binormal ROC curve is $ROC(u) = \Phi(a + b\Phi^{-1}(u))$, where $a = \mu/\sigma$, $b = 1/\sigma$ and Φ denotes the standard normal cumulative distribution function. Also, the AUC for the binormal ROC curve is $AUC = \Phi\left(a/\sqrt{1+b^2}\right)$ (Pepe, 2003). However, this binormal model can be overly restrictive in practice. If the data are not normally distributed, it is desirable to use more flexible class of distributions.

3.2 Dirichlet Process Mixtures

The Dirichlet process (DP) is a commonly used prior in a nonparametric Bayesian framework when the distribution exhibits more flexibility in location and shape, which may not be well captured by parametric models. Suppose a real-valued continuous random variable Y_i , $i = 1, \dots, n$, is sampled from an unknown distribution F . We could place a DP prior on the distribution F (Ferguson, 1973). The DP with precision α and base parametric distribution F_0 , denoted by $F \sim DP(\alpha F_0)$, can be described by the stick-breaking representation (Sethuraman, 1994):

$$F = \sum_{h=1}^{\infty} p_h \delta_{\theta_h^*}, \quad p_h = V_h \prod_{l=1}^{h-1} (1 - V_l), \quad V_h \sim \text{Beta}(1, \alpha), \quad \theta_h^* \sim F_0,$$

where δ_{θ^*} is a point mass at θ^* and all V_h 's and θ_h^* 's are independent. The θ_h^* 's are random atoms which represent all possible values of Y_i . The corresponding probability weights ($p_1, p_2, \dots$) are generated from a stick-breaking procedure using independent Beta distributed V_h ($h = 1, 2, \dots$). The base measure F_0 corresponds to our best guess for unknown distribution F . The precision parameter $\alpha > 0$ controls the variability in this guess for base measure, with F approaching F_0 as $\alpha \rightarrow \infty$. The unknown distribution F can be interpreted as a discrete distribution comprised of a countably infinite number of distinct values according to random

weights. One simple extension to remove the discreteness restriction is a Dirichlet process mixture (DPM) model (Antoniak, 1974). The DP can be employed as a prior for the mixture distribution, e.g., $F(y; F_1) = \int K(y; \theta) dF_1(\theta)$, $F_1 \sim DP(aF_0)$. Then, the discreteness constraint of DP can be solved by choosing a continuous distribution for the kernel, $K(y; \theta)$. In the application of ROC curves, a bi-DPM model can be constructed if we assume the healthy and diseased populations each follow a separate DPM (Erkanli et al., 2006).

3.3 Modeling Stochastic Order

In comparing two populations, let F_1 and F_2 be two population distributions on $\theta \in \mathcal{R}$. The stochastic order of two populations is defined as follows: F_2 is stochastically larger than F_1 if $F_1(c) \geq F_2(c)$ for all c , denoted by $F_1 \leq_{st} F_2$. Gelfand and Kottas (2001) showed how stochastic order can be retained through mixture distributions. Consider a parametric family of distributions $\{K(y; \theta) : \theta \in \mathcal{R}\}$, with $K(y; \theta)$ differentiable and strictly decreasing in θ , and mixture distribution $F(y; F_i) = \int K(y; \theta) dF_i(\theta)$, $i = 1, 2$. Then, if $F_1 \leq_{st} F_2$, $F(y; F_1) \leq_{st} F(y; F_2)$. A convenient choice of F_1 and F_2 satisfying $F_1 \leq_{st} F_2$ is $F_1 = G_1$ and $F_2 = G_1 G_2$, where G_1 and G_2 are distribution functions on $\mathcal{R}$ (Gelfand and Kottas, 2001; Kottas and Gelfand, 2002; Hanson, Kottas and Branscum, 2008). Then, the space $\mathcal{P} = \{(F_1, F_2) : F_1 \leq_{st} F_2\}$ can be induced from the subspace $\mathcal{P}' = \{(F_1, F_2) : F_1 = G_1, F_2 = G_1 G_2\}$. Here, F_1 can be considered as the distribution of θ and F_2 as the distribution of $\max(\theta, \delta)$, where $\theta \sim G_1$ and independently $\delta \sim G_2$. The stochastically ordered mixture distribution is then given as

$$\begin{aligned} F(y; F_1) &= \int K(y; \theta) dG_1(\theta), \\ F(y; F_2) &= \int \int K(y; \max(\theta, \delta)) dG_1(\theta) dG_2(\delta), \end{aligned}$$

with $F(y; F_1) \leq_{st} F(y; F_2)$. Independent DP priors can be used for G_1 and G_2 : $G_1 \sim DP(a_1 G_{10})$, $G_2 \sim DP(a_2 G_{20})$. Thus the DPM model turned out to be a useful framework to account for stochastic order restriction (Gelfand and Kottas, 2001; Hanson, Kottas and Branscum, 2008).

3.4 Modeling Variability Order

A similar approach can be taken to account for ordering in variability (Kottas and Gelfand, 2002). Let X and Y be two random variables with distribution functions F_1 and F_2 . Following the notion of sign changes of a function, the number of sign changes of a function h is given by $S^-(h) = \sup S^-(h(u_1), \dots, h(u_m))$ for $u_1 \leq \dots \leq u_m$, $u_i \in U \subseteq \mathcal{R}$, where m is arbitrary, but finite and $S^-(\cdot)$ is the number of sign changes of the indicated sequence. Then, Y is said to be larger than X in variability order, denoted by $F_1 \leq_{var} F_2$ if F_1 and F_2 have the same location and $S^-(F_2 - F_1) = 1$, the sign sequence being $+$, $-$. To see how variability order can be preserved through the mixture distribution, consider a parametric family of distributions $\{K(y; \theta) : \theta \in \mathcal{R}\}$, with $K(y; \theta)$ differentiable in θ , such that for any $\theta_1 < \theta_2$, $S^-(K(y; \theta_2) - K(y; \theta_1)) = 1$, the sign sequence being $+$, $-$. Then, if $F_1 \leq_{st} F_2$, $F(y; F_1) \leq_{var} F(y; F_2)$, where $F(y; F_i) = \int K(y; \theta) dF_i(\theta)$, $i = 1, 2$. As in stochastic order, we can let $F_1 = G_1$ and $F_2 = G_1 G_2$, where $G_1 \sim DP(\beta_1 G_{10})$, $G_2 \sim DP(\beta_2 G_{20})$, and choose a symmetric scale family kernel with a location parameter μ for the mixtures. It follows

$$\begin{aligned} F(y; \mu, F_1) &= \int K(y; \mu, \theta) dG_1(\theta), \\ F(y; \mu, F_2) &= \int \int K(y; \mu, \max(\theta, \delta)) dG_1(\theta) dG_2(\delta), \end{aligned}$$

with $F(y; \mu, F_1) \leq_{var} F(y; \mu, F_2)$.

3.5 Joint Modeling of Stochastic and Variability Orders

Let $\mathbf{Y}_{ij} = (Y_{ij1}, \dots, Y_{ijn_{ij}})^T$ be a $(n_{ij} \times 1)$ response vector measured on i th population, $i = 1$ (healthy), 2 (diseased) at j th setting, $j = 1, 2$. We develop scale and location mixture models for the distribution functions of each population, $F_{ij}(y) \equiv \int \int K(y; \theta, \sigma) dH_i(\theta) dG_j(\sigma)$, $i = 1, 2$ and $j = 1, 2$. For simplicity in illustration, we assume for now that there exists a gold standard for disease and will relax this assumption in Section 5. It is possible that $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$, $i = 1, 2$ are correlated; direct modeling of this correlation is not straightforward in the Bayesian nonparametric framework. For this reason, we chose not to explicitly build a correlation structure, although the constraints across settings will induce some dependence. We account for two order constraints simultaneously: stochastic order and variability order. First, for the j th setting, $j = 1, 2$, the response vector $(\mathbf{Y}_{1j}, \mathbf{Y}_{2j})$ has a stochastic order between the healthy and diseased populations. That is, the diseased population is more likely to have higher values of $\mathbf{Y}$ than the healthy population. Second, for the i th population, $i = 1, 2$, the response vector $(\mathbf{Y}_{i1}, \mathbf{Y}_{i2})$ has a variability order between two settings. It implies that setting 2 is more likely to have smaller variability of $\mathbf{Y}$ than setting 1. The following result illustrates how stochastic and variability orders can be preserved through scale and location mixture distributions, which extends the work of Gelfand and Kottas (2001) and Kottas and Gelfand (2002) that have separately considered these two constraints. The proof of this result can be found in the supplementary materials.

Result 1—Consider a parametric family of distributions $\{K(y; \theta, \sigma) : \theta \in \mathcal{R}, \sigma \in \mathcal{R}^+\}$, where θ is a location parameter, σ is a scale parameter and $\mathcal{R}^+$ is the positive real line. Suppose $K(y; \theta, \sigma)$ is differentiable and strictly decreasing in θ and also differentiable in σ , such that for any $\sigma_1 < \sigma_2$, $S^-(K(y; \theta, \sigma_2) - K(y; \theta, \sigma_1)) = 1$, the sign sequence being $+$, $-$, where the crossing point y_0 of $K(y; \theta, \sigma_2)$ and $K(y; \theta, \sigma_1)$ is the same for all σ_1 and σ_2 . For distributions H_1 and H_2 on $\mathcal{R}$, and distributions G_1 and G_2 on $\mathcal{R}^+$, define the mixtures $F_{ij}(y; H_i, G_j) \equiv \int_{\mathcal{R}^+} \int_{\mathcal{R}} K(y; \theta, \sigma) dH_i(\theta) dG_j(\sigma)$, $i, j = 1, 2$. Then if $H_1 \leq_{st} H_2$, $F_{1j}(y; H_1, G_j) \leq_{st} F_{2j}(y; H_2, G_j)$ for $j = 1, 2$, and if $G_1 \leq_{st} G_2$, $F_{i1}(y; H_i, G_1) \leq_{var} F_{i2}(y; H_i, G_2)$ for $i = 1, 2$.

In practice, we use $K(y; \theta, \sigma) = \Phi(y; \theta, \sigma)$ as it satisfies the conditions in Result 1. We let $H_1 = P_1$, $H_2 = P_1 P_2$ for the location parameter and $G_1 = Q_1 Q_2$, $G_2 = Q_1$ for the scale parameter, and introduce latent variables, $\theta \sim P_1$, $\delta \sim P_2$ and $\sigma \sim Q_1$, $\psi \sim Q_2$. The likelihood of the proposed model under jointly stochastic and variability order constraints is given by

$$L = \prod_{i=1}^2 \prod_{j=1}^2 \prod_{k=1}^{n_{ij}} F_{ij}(y_{ijk}; H_i, G_j), \quad (1)$$

where the corresponding mixture distributions are as follows:

$$\begin{aligned}
F_{11}(y; H_1, G_1) &= \int \int \Phi(y; \theta, \max(\sigma, \psi)) dP_1(\theta) dQ_1(\sigma) dQ_2(\psi); \\
F_{21}(y; H_2, G_1) &= \int \int \int \Phi(y; \max(\theta, \delta), \max(\sigma, \psi)) dP_1(\theta) dP_2(\delta) dQ_1(\sigma) dQ_2(\psi); \\
F_{12}(y; H_1, G_2) &= \int \int \Phi(y; \theta, \sigma) dP_1(\theta) dQ_1(\sigma); \\
F_{22}(y; H_2, G_2) &= \int \int \Phi(y; \max(\theta, \delta), \sigma) dP_1(\theta) dP_2(\delta) dQ_1(\sigma).
\end{aligned}$$

4. Priors and Posterior Computations

4.1 Dirichlet Process Priors

We take independent DP priors with normal base distributions for P and inverse-gamma base distributions for Q to obtain the following hierarchical model:

$$\begin{aligned}
Y_{11k_1} | \cdot &\stackrel{\text{ind}}{\sim} N(\theta_{k_1}, \max(\sigma_{n_{12}+n_{22}+k_1}, \psi_{k_1})), & k_1=1, \dots, n_{11}, \\
Y_{21k_2} | \cdot &\stackrel{\text{ind}}{\sim} N(\max(\theta_{n_{11}+k_2}, \delta_{k_2}), \max(\sigma_{n_{11}+n_{12}+n_{22}+k_2}, \psi_{n_{11}+k_2})), & k_2=1, \dots, n_{21}, \\
Y_{12k_3} | \cdot &\stackrel{\text{ind}}{\sim} N(\theta_{n_{11}+n_{21}+k_3}, \sigma_{k_3}), & k_3=1, \dots, n_{12}, \\
Y_{22k_4} | \cdot &\stackrel{\text{ind}}{\sim} N(\max(\theta_{n_{11}+n_{21}+n_{12}+k_4}, \delta_{n_{21}+k_4}), \sigma_{n_{12}+k_4}), & k_4=1, \dots, n_{22},
\end{aligned} \tag{2}$$

where

$$\begin{aligned}
\theta_{k_1} &\stackrel{\text{iid}}{\sim} P_1, & P_1 &\sim DP(\alpha_1 P_{10}), & P_{10} &= N(\mu_1, \tau_1^2), & k_1=1, \dots, (n_{11}+n_{21}+n_{12}+n_{22}); \\
\theta_{k_2} &\stackrel{\text{iid}}{\sim} P_2, & P_2 &\sim DP(\alpha_2 P_{20}), & P_{20} &= N(\mu_2, \tau_2^2), & k_2=1, \dots, (n_{21}+n_{22}); \\
\sigma_{k_2} &\stackrel{\text{iid}}{\sim} Q_1, & Q_1 &\sim DP(\beta_1 Q_{10}), & Q_{10} &= IG(a_1, b_1), & k_3=1, \dots, (n_{11}+n_{21}+n_{12}+n_{22}); \\
\psi_{k_4} &\stackrel{\text{iid}}{\sim} Q_2, & Q_2 &\sim DP(\beta_2 Q_{20}), & Q_{20} &= IG(a_2, b_2), & k_4=1, \dots, (n_{11}+n_{21}).
\end{aligned}$$

The hyperparameters $\mu_i, \tau_i^2, a_i, b_i, i = 1, 2$ are all assumed fixed such that μ_i is roughly located in the middle of the data and the rest exhibit large variance. The precision parameters $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ of the DP play an important role in deciding the number of distinct values in our model. For instance, for relatively small α_1 , each random weight is likely to be near one and as a result, many θ_{k_1} 's will share the same distinct value. Ideally, we would like to assign a hyperprior to α and β . However, for simpler implementation, we conduct a sensitivity analysis for fixed α and β . If the results differ substantially with α and β , choice of α and β could be determined by comparing predictive and empirical distributions. As a result of the hierarchical presentation in (2), the likelihood function (1) can be rewritten as:

$$\begin{aligned}
L &= \prod_{k_1=1}^{n_{11}} N(y_{11k_1}; \theta_{k_1}, \max(\sigma_{n_{12}+n_{22}+k_1}, \psi_{k_1})) \\
&\quad \times \prod_{k_2=1}^{n_{21}} N(y_{21k_2}; \max(\theta_{n_{11}+k_2}, \delta_{k_2}), \max(\sigma_{n_{11}+n_{12}+n_{22}+k_2}, \psi_{n_{11}+k_2})) \\
&\quad \times \prod_{k_3=1}^{n_{12}} N(y_{12k_3}; \theta_{n_{11}+n_{21}+k_3}, \sigma_{k_3}) \times \prod_{k_4=1}^{n_{22}} N(y_{22k_4}; \max(\theta_{n_{11}+n_{21}+n_{12}+k_4}, \delta_{n_{21}+k_4}), \sigma_{n_{12}+k_4})
\end{aligned} \tag{3}$$

4.2 Posterior Computations

Posterior computation proceeds with a hybrid MCMC algorithm consisting of Gibbs sampling (Casella and George, 1992) and Metropolis-Hastings steps (Chib and Greenberg, 1995). We employ a truncation approximation for the stick-breaking representation of the DP priors for P_1, P_2, Q_1 and Q_2 , proposed by Ishwaran and Zarepour (2000). This method samples from the posteriors using MCMC algorithms, resulting in rapid mixing of the Markov chain and is easy to apply to normal mean mixture models. We let

$$P_1 = \sum_{h_1=1}^M p_{h_1} \delta_{\theta_{h_1}^*}, \quad P_2 = \sum_{h_2=1}^M p_{h_2} \delta_{\delta_{h_2}^*}, \\ Q_1 = \sum_{l_1=1}^L q_{l_1} \delta_{\sigma_{l_1}^*}, \quad Q_2 = \sum_{l_2=1}^L q_{l_2} \delta_{\psi_{l_2}^*}.$$

Appropriate truncation levels M and L can be determined by examining the moments of the random weights. As a simpler approximation, we recommend using a moderately large numbers (e.g., $M=30$) and sensitivity to the approximation could be then assessed by using a larger M and comparing results.

The predictive density of $Y_{ijk}, i, j = 1, 2$ can be obtained at each iteration of the MCMC:

$$p(Y_{11k_1} | \cdot) \propto \sum_{h_1=1}^M \sum_{l_1=1}^L \sum_{l_2=1}^L p_{h_1} q_{l_1} q_{l_2} N(Y_{11k_1}; \theta_{h_1}^*, \max(\sigma_{l_1}^*, \psi_{l_2}^*)), \\ p(Y_{21k_2} | \cdot) \propto \sum_{h_1=1}^M \sum_{h_2=1}^M \sum_{l_1=1}^L \sum_{l_2=1}^L p_{h_1} p_{h_2} q_{l_1} q_{l_2} N(Y_{21k_2}; \max(\theta_{h_1}^*, \delta_{h_2}^*), \max(\sigma_{l_1}^*, \psi_{l_2}^*)), \\ p(Y_{12k_3} | \cdot) \propto \sum_{h_1=1}^M \sum_{l_1=1}^L p_{h_1} q_{l_1} N(Y_{12k_3}; \theta_{h_1}^*, \sigma_{l_1}^*), \\ p(Y_{22k_4} | \cdot) \propto \sum_{h_1=1}^M \sum_{h_2=1}^M \sum_{l_1=1}^L p_{h_1} p_{h_2} q_{l_1} N(Y_{22k_4}; \max(\theta_{h_1}^*, \delta_{h_2}^*), \sigma_{l_1}^*).$$

More details regarding our sampling algorithm including full conditional posterior distributions can be found in the supplementary materials.

4.3 ROC curve and AUC Estimations

The estimation of ROC curves is obtained using predictive distribution function of $\mathbf{Y}_{ij}, i, j = 1, 2$. That is, $\hat{ROC}(u) = 1 - \hat{F}_D\{\hat{F}_H^{-1}(1-u)\}$, for $0 \leq u \leq 1$, where $\hat{F}_H(\cdot)$ and $\hat{F}_D(\cdot)$ are posterior distributions of Y in healthy and diseased populations, respectively, and s is a threshold value. To assess diagnostic accuracy, AUC is estimated by the posterior mean of the AUCs obtained using $AUC = Pr(Y^D > Y^H)$ at each iteration of the MCMC, where Y^H and Y^D are sampled from predictive densities of the healthy and diseased populations, respectively at each iteration (Bamber, 1975).

5. A Revisit of the Physician Reliability Study

In this section, we build on the initial analysis in Section 2 and re-analyze the PRS data with two objectives in mind. First, we incorporate constraints in the estimation and second, we address the issue of no gold standard.

5.1 Order Constraints

As described earlier, the design of PRS involves successively augmented clinical information across settings for the physicians to use to diagnose endometriosis. In particular, the physicians were only provided with the women's intrauterus digital images at setting 1. At setting 2, surgeons' notes were added to the clinical file that was then made available to the physicians. As a result, the physicians were seeing a larger collection of clinical information for diagnosing endometriosis at setting 2 than at setting 1. In our collaborations with OB/GYN experts two order constraints were suggested:

- **C1:** For a given physician at a given setting, the rASRM score should have a stochastic order between the diseased and healthy populations. It means that the diseased population is more likely to have higher rASRM score than the healthy population.
- **C2:** For a given woman and a given physician, the rASRM score in setting 2 is more likely to have smaller variability than setting 1.

While the stochastic order is consistent with what was observed in the data, the actual rASRM score has larger variance in setting 2 than 1 (Table 1). It is of interest to see how our method models the data under the two order constraints, one of which has the opposite direction to the real data.

5.2 Sensitivity Analysis for No Gold Standard

In the initial analysis and model development sections, we have assumed a known gold standard for the disease. To handle the situation that no gold standard exists in the PRS, we construct an imperfect reference standard based on the diagnoses of the four IEs. Let D_k be the true disease status of the k th subject, $k = 1, \dots, N$. We denote T_{jk} as the sum of the IEs' positive diagnoses and J_{jk} as the number of non-missing IEs' diagnoses on k th subject at j th setting, $k = 1, \dots, N, j = 1, 2$. We model D_k as a binary latent variable by using the following logit model,

$$P(D_k=1|T_k, J_k) = \frac{\exp(\gamma_0 + \gamma_1 [1 - \frac{T_k}{J_k}])}{1 + \exp(\gamma_0 + \gamma_1 [1 - \frac{T_k}{J_k}])},$$

where $T_k = T_{1k} + T_{2k}$ and $J_k = J_{1k} + J_{2k}$. Ideally, we would like to take hyperpriors for γ_0 and γ_1 in the model, but that will result in weak identifiability. We instead conducted a sensitivity analysis in which we vary γ_0 and γ_1 in a scientifically reasonable way and assess robustness of the results. In particular, we chose $\gamma_0 = 4.5$ and $\gamma_1 = -9$ so that $P(D_k = 1|T_k/J_k = 1)$ is close to 1 and $P(D_k = 1|T_k/J_k = 0)$ is close to 0. We examine model robustness using

various γ_0 and γ_1 in the PRS data. With the inclusion of the sensitivity analysis component, the likelihood function (3) now has the following form:

$$L = \prod_{k=1}^N [N(y_{1k}; \theta_k, \max(\sigma_{N+k}, \psi_k)) N(y_{2k}; \theta_{N+k}, \sigma_k)]^{1-D_k} \\ \times [N(y_{1k}; \max(\theta_k, \delta_k), \max(\sigma_{N+k}, \psi_k)) N(y_{2k}; \max(\theta_{N+k}, \delta_{N+k}), \sigma_k)]^{D_k} \\ \times \left[1 - \frac{\exp(\gamma_0 + \gamma_1 [1 - \frac{T_k}{J_k}])}{1 + \exp(\gamma_0 + \gamma_1 [1 - \frac{T_k}{J_k}])} \right]^{1-D_k} \left[\frac{\exp(\gamma_0 + \gamma_1 [1 - \frac{T_k}{J_k}])}{1 + \exp(\gamma_0 + \gamma_1 [1 - \frac{T_k}{J_k}])} \right]^{D_k}.$$

5.3 Model Specifications

In the re-analysis, we fit a sequence of models, from no constraint to all constraints, using DPM for each population to allow distributional flexibility and the sensitivity approach in Section 5.2 to address no gold standard problem. These models are:

- i. **No-constraint (NO) model:** No constraint is imposed on the 4 distributions of the response variable (Y), except each follows a different DPM. This amounts to separate ROC estimation between the 2 settings;
- ii. **Stochastic order constraint only (SO) model:** On top of (i), we impose the stochastic order constraint (C1) on the distributions within each setting, while leaving the variability unconstrained across settings;
- iii. **Variability order constraint only (VO) model:** On top of (i), we impose the variability order constraint (C2) on the distributions within each population, while letting the distributional center unconstrained across populations;
- iv. **Joint (JO) model:** On top of (i), we impose the stochastic order (C1) within each setting and variability order (C2) within each population.

More details regarding the four model specifications can be found in the supplementary materials. To make the comparison between the four models fair, the DPM in models (i) to (iii) is implemented with respect to both the location and scale parameters, same as in model (iv). We chose normal mixture kernels $K(y; \theta, \sigma) = \Phi(y; \theta, \sigma)$ for all models, and took the DP priors with normal base distribution $N(0, 10^2)$ for the mean and inverse-Gamma base distribution $IG(2, 1)$ for the variance. The precision parameters of DP priors were fixed at 1, i.e., $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$, since we noticed the parameters were not very sensitive after comparing predicted and empirical distributions. For the truncation approximation of DP, we used $M = L = 30$ as a truncation level, which was large enough based on our empirical examination. We obtained results by running the MCMC algorithm for 50,000 iterations with 25,000 iteration burn-ins. Particularly, the multi-chain MCMC algorithm with various starting points was implemented in the joint model to produce satisfactory convergence results while saving on running times.

5.4 Results

Figure 5 contains the estimated predictive densities (and histograms) of log rASRM scores from models (i) – (iv). This figure suggests that the nonparametric feature of our models identifies very well the non-standard shape of the distributions, such as skewness in the

healthy populations and bimodality in the diseased populations. This may not be possible if parametric models were taken to model these scores.

The main results of applying the proposed methods to the PRS data are provided in Figure 6 and Table 2, where the estimated ROC curves and corresponding posterior estimates of AUCs and their 95% credible intervals are presented for each of the four models at each of the two settings. While we observe that the REs are generally good at diagnosing endometriosis at both setting 1 (AUC ranges from 0.80 to 0.84) and setting 2 (AUC ranges from 0.77 to 0.85), we also notice the impact of the order constraints. The AUC estimates hardly changed from the no-constraint (NO) model to the SO model; this is expected as the PRS data already agree with the stochastic order constraint. Going from the NO to the VO model reveals some subtle changes. First, the direction of the estimated AUCs changed. While the AUC at setting 1 is larger than setting 2 in the NO model, it is smaller once the variability order is considered. This illustrates that although the data may suggest one directional order of the variability, it is the order constraint that dominates the inference when it is in disagreement with the data. We also notice the shrinkage of the 95% CI width in the VO model compared to the NO model, an indication of efficiency gain, although marginally. These two observations are more pronounced once the stochastic and variability orders are jointly considered in the model. As in the VO model, the AUC in setting 2 is larger than setting 1 in the JO model; however, the efficiency gain over the NO model is more substantial, with the width of 95% CIs at 0.11 for the JO model versus 0.23 for the NO model in setting 1 and 0.11 versus 0.19 in setting 2, respectively. This large efficiency gain in AUC estimates is likely a result of the interplay of the stochastic and variability orders.

In conducting the sensitivity analysis, we varied γ_0 and γ_1 over a wide range of reasonable values: $\gamma_0 = 2, 3, 4.5, 6, 7$ and $\gamma_1 = -15, -12, -9, -8, -7$. When we used different values of γ_0 's with $\gamma_1 = -9$ and different values of γ_1 's with $\gamma_0 = 4.5$, respectively, we found the estimates of ROC curves and AUCs in all scenarios to be very close to those from our primary analysis.

6. Simulation Studies

We conduct simulation studies under four scenarios to assess the performance of our proposed approach. These scenarios consider different specifications in distributions, variability order directions, and sample sizes. We fit the same four models as in Section 5.3 to the simulated datasets and evaluate their performances in terms of bias and efficiency. The results of the simulation studies corroborate most of what was observed in the real data analyses:

- a. Our approach can adequately identify the distributions of the scores;
- b. There are no discernable differences in biases between the two sample sizes and the two distribution assumptions;
- c. There is efficiency gain when constraints are considered;
- d. Having the variability order in the opposite direction of the data results in estimated AUCs with reversed direction.

Detailed simulation set up and results are reported in Section 5 of the supplementary materials.

7. Concluding Remarks

In this paper, we have proposed an integrated framework for jointly incorporating order constraints, one on the center and the other on the variability of the distributions, and applied it to estimating multiple ROCs. The work advances early literature which addressed these constraints separately. This integrated framework was made possible by utilizing a Bayesian nonparametric technique, i.e., the Dirichlet process mixture, which also provided the needed distributional flexibility in the test scores.

Substantively, we found that a regional OB/GYN physician can better diagnose endometriosis using both the intrauterus image and surgeon's notes (setting 2) than using the image only (setting 1) after both order constraints are incorporated. This represents a reversal of the direction of AUCs between settings 1 and 2 when the order constraints were not considered. A closer look reveals that the change of direction comes as a result of the variability order, not of the stochastic order. That the stochastic order is not related to the AUC reversal is not surprising, as the PRS data already agree with this order constraint.

In modeling the REs' ROCs, we took the arithmetic average of the 4 individual rASRM scores in the RE group. This step is reasonable as we assumed that the 4 REs are exchangeable and as the research goal is on the group-specific, not physician-specific diagnostic accuracy. While alternative ways exist (e.g., obtain physician-specific ROCs and then average them), our approach represents our first step of studying REs' diagnostic accuracy. We feel that if the accuracy obtained from the average score is not great, then that from individual scores will be worse. Using the average scores also enabled us to retain those women with some scores missing, which in turn allowed us to have reasonable sample sizes.

An alternative to our imperfect reference standard approach in the PRS data is to model the diagnostic data with neither a gold standard nor an imperfect reference standard. Such an approach essentially amounts to a deconvolution of a two-component mixture; see Branscum et al. (2008). However, as others and Branscum et al. have pointed out, this deconvolution approach inherently lacks identifiability unless strong prior information is used. When aided with no useful prior information, such an approach can result in slow mixing and poor convergence in the Markov chain Monte Carlo iterations. Indeed, when we ran a simulation study using no imperfect reference standard, we observed that under weak prior, the MCMC chain has poor mixing and converges to an AUC that is far away from the truth (see supplementary materials).

While it is common to estimate ROCs and the associated AUCs through modeling the distributions of the observed scores, alternatives have been suggested in the literature that directly models the ROCs. Of particular interest is the idea of using placement values (see e.g., Pepe (2000) and Alonzo and Pepe (2002)). In this framework, *a priori* constraints on the ROCs, whether in the entire or partial range of specificity, can then be incorporated. This is a current research interest of the authors.

In our proposed models, we have specified independent mixture models for the location (θ) and scale (σ) of the DPM kernel and used the idea of product measures to induce order constraints, i.e., $H_1 = P_1$, $H_2 = P_1 P_2$ for stochastic order and $G_1 = Q_1 Q_2$, $G_2 = Q_1$ for variability order (see the last paragraph of Section 3.5). However, the independent mixture can potentially lead to unsubstantiated clusters. Alternatively, one can specify a dependent mixture: for $i = 1, 2, j = 1, 2$ and $k = 1, \dots, n_{ij}$,

$$\begin{aligned} Y_{ijk} &\sim N(\theta_{ik}, \sigma_{jk}), \\ (\theta_{ik}, \sigma_{jk}) &\sim H_{ij}, \end{aligned}$$

where $H_{ij} \sim DP(aH_{ij0})$, say, with H_{ij0} a bivariate distribution of θ and σ . Under this dependent mixture specification, it is not clear how the product measure idea can be used to induce the desired constraints. One possibility is to work with the multivariate distribution of the vector $\xi_k = (\{\theta_{ik}\}, \{\sigma_{jk}\})$ and follow Hoff (2003) to use partially ordered latent observations to model stochastically ordered marginals, e.g., $F(\theta_{1k})$ and $F(\theta_{2k})$ or $F(\sigma_{1k})$ and $F(\sigma_{2k})$ (see also Dunson and Peddada (2008)). This is an active research project of the authors. To gauge the impact of the two mixture specifications, we implemented the dependent mixture idea in the no-constraint model, and found quite similar results (see supplementary materials).

While it is convenient to use the product measures, i.e., P_1 , P_2 and Q_1 , Q_2 in Section 3, to induce the ordering, such a construction also results in a reduced space for the desired order constraints. It is therefore of interest to examine the impact of such a limitation and, if needed, to explore more general ways of incorporating these constraints in a Bayesian nonparametric framework. It is also of interest to consider other order constraints, such as the stochastic precedence on the distributional centers (Chen and Dunson, 2004).

In the PRS, the REs also gave ordinal diagnosis on the participants. Methodologically, it is of interest to explore possible extension of the current framework to the case with ordinal data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (<http://biowulf.nih.gov>). The authors thank Athanasios Kottas, Paul S. Albert, the editor, associate editor, and two anonymous reviewers for insightful and helpful comments.

References

- Alonzo TA, Pepe MS. Distribution-free ROC analysis using binary regression techniques. *Biostatistics*. 2002; 3:421–432. [PubMed: 12933607]

- Antoniak CE. Mixtures of Dirichlet processes with applications to nonparametric problems. *The Annals of Statistics*. 1974; 2:1152–1174.
- Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975; 12:387–415.
- Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine*. 2005; 68:145–163. [PubMed: 15820113]
- Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine*. 2008; 27:2474–2496. [PubMed: 18300333]
- Cai T, Moskowitz CS. Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics*. 2004; 5:573–586. [PubMed: 15475420]
- Casella G, George EI. Explaining the Gibbs sampler. *The American Statistician*. 1992; 46:167–174.
- Chen Z, Dunson DB. Bayesian estimation of survival functions under stochastic precedence. *Lifetime Data Analysis*. 2004; 10:159–173. [PubMed: 15293630]
- Cheng D, Branscum AJ, Johnson WO. Sample size calculations for ROC studies: parametric robustness and Bayesian nonparametrics. *Statistics in Medicine*. 2012; 31:131–142. [PubMed: 22139729]
- Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. *The American Statistician*. 1995; 49:327–335.
- De Carvalho VI, Jara A, Hanson TE, De Carvalho M. Bayesian nonparametric ROC regression modeling. *Bayesian Analysis*. 2013; 8:623–646.
- Dunson DB, Peddada SD. Bayesian nonparametric inference on stochastic ordering. *Biometrika*. 2008; 95:859–874.
- Erkanli A, Sung M, Costello EJ, Angold A. Bayesian semi-parametric ROC analysis. *Statistics in Medicine*. 2006; 25:3905–3928. [PubMed: 16416403]
- Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Statistics in Medicine*. 2002; 21:3093–3106. [PubMed: 12369084]
- Ferguson TS. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*. 1973; 1:209–230.
- Gelfand AE, Kottas A. Nonparametric Bayesian modeling for stochastic order. *Annals of the Institute of Statistical Mathematics*. 2001; 53:865–876.
- Gu J, Ghosal S. Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference*. 2009; 139:2076–2083.
- Gu J, Ghosal S, Roy A. Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine*. 2008; 27:5407–5420. [PubMed: 18613217]
- Hanley JA. The use of the ‘binormal’ model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine*. 1996; 15:1575–1585. [PubMed: 8855483]
- Hanson TE, Branscum AJ, Gardner IA. Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling*. 2008; 8:81–96.
- Hanson TE, Kottas A, Branscum AJ. Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches. *Journal of the Royal Statistical Society: Series C*. 2008; 57:207–225.
- Hoff PD. Bayesian methods for partial stochastic orderings. *Biometrika*. 2003; 90:303–317.
- Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*. 1996; 24:25–40.
- Ishwaran H, Zarepour M. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*. 2000; 87:371–390.
- Jafarzadeh SR, Johnson WO, Utts JM, Gardner IA. Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. *Statistics in Medicine*. 2010; 29:2090–2106. [PubMed: 20603894]
- Kottas A, Gelfand AE. Modeling variability order: A semiparametric Bayesian approach. *Methodology and Computing in Applied Probability*. 2002; 3:427–442.
- Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*. 1998; 93:1356–1364.

- Peng F, Hall WJ. Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods. *Medical Decision Making*. 1996; 16:404–411. [PubMed: 8912302]
- Peng L, Zhou XH. Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference*. 2004; 118:129–143.
- Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*. 1998; 54:124–135. [PubMed: 9544511]
- Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*. 2000; 56:352–359. [PubMed: 10877289]
- Pepe, MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press; New York: 2003.
- Robertson, T.; Wright, FT.; Dykstra, RL. *Order restricted statistical inference*. John Wiley & Sons; New Jersey: 1988.
- Rodriguez A, Martinez JC. Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics*. 2014; 15:353–369. [PubMed: 24174579]
- Schliep KC, Stanford JB, Chen Z, Zhang B, Dorais JK, Johnstone EB, Hammoud AO, Varner MW, Buck Louis GM, Peterson CM. Interrater and intrarater reliability in the diagnosis and staging of endometriosis. *Obstetrics & Gynecology*. 2012; 120:104–112. [PubMed: 22914398]
- Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica*. 1994; 4:639–650.
- Sorribas A, March J, Trujillano J. A new parametric method based on *S*-distributions for computing receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine*. 2002; 21:1213–1235. [PubMed: 12111875]
- Tang LL, Liu A, Chen Z, Schisterman EF, Zhang B, Miao Z. Nonparametric ROC summary statistics for correlated diagnostic marker data. *Statistics in Medicine*. 2013; 32:2209–2220. [PubMed: 23055248]
- Zhang B, Chen Z, Albert PS. Estimating diagnostic accuracy of raters without a gold standard by exploiting a group of experts. *Biometric*. 2012; 68:1294–1302.

Figure 1. Histograms of rASRM scores in the PRS by scale and setting. The scores are not separated by disease status.

Figure 2. Histograms of log rASRM scores in the PRS by disease status and setting. The disease status is from posterior estimates in Section 5.

Figure 3. Scatter plots of REs' rASRM scores and IEs' diagnosis in the PRS. The fitted lines are all significant with $p < 0.01$.

Figure 4.
Estimated ROC curves and AUCs (95% CI) from the initial analysis by setting and model.

Figure 5.

The estimated predictive densities (and histograms) of log rASRM scores in the PRS by disease status and setting. The posterior mean of the density from the JO model is denoted by solid line for each population at each setting. For comparison, densities from NO (dotted line), SO (dashed line) and VO (dash-dotted line) models are provided.

Figure 6. The estimated ROC curve of log rASRM scores in the PRS by setting. The posterior mean of the ROC curve from the JO model is denoted by solid line at each setting. For comparison, the estimated ROC curves from NO (dotted line), SO (dashed line) and VO (dash-dotted line) models are provided.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Descriptive statistics of the log rASRM score in the PRS data. The disease status is from posterior estimates in Section 5.

Score	Setting 1				Setting 2			
	N	Mean	Median	SD	N	Mean	Median	SD
Combined	113	1.81	1.39	1.22	126	1.89	1.39	1.31
Healthy	22	0.78	0.69	0.76	27	0.84	0.69	0.84
Diseased	91	2.06	1.66	1.18	99	2.17	1.71	1.27

Table 2

Posterior mean (95% CI) of AUCs of the PRS data. NO model assumes no constraint on the 4 distributions; SO model assumes the stochastic order constraint within each setting; VO model assumes the variability order constraint within each population; and JO model assumes both the stochastic order within each setting and variability order within each population.

Model	Setting 1	Setting 2
NO model	0.84 (0.73, 0.96)	0.78 (0.68, 0.87)
SO model	0.80 (0.70, 0.91)	0.77 (0.66, 0.85)
VO model	0.81 (0.72, 0.90)	0.83 (0.75, 0.91)
JO model	0.81 (0.75, 0.86)	0.85 (0.79, 0.90)