



From the SelectedWorks of Dunlei Cheng

January 2012

Sample Size Calculations for ROC Studies: Parametric Robustness and Bayesian Nonparametrics

Contact
Author

Start Your Own
SelectedWorks

Notify Me
of New Work

Available at: http://works.bepress.com/dunlei_cheng/4

Sample Size Calculations for ROC Studies: Parametric Robustness and Bayesian Nonparametrics

DUNLEI CHENG¹, ADAM J. BRANSCUM², and WESLEY O. JOHNSON³

¹*Institute for Health Care Research and Improvement, Baylor Health Care System,
Dallas, TX 75206, USA*

²*Department of Public Health, Oregon State University,
Corvallis, OR 97331, USA*

³*Department of Statistics, University of California, Irvine,
Irvine, CA 92697, USA*

Correspondence to: Dunlei Cheng, Institute for Health Care Research and Improvement, Baylor Health Care System, 8080 N. Central Expressway, Suite 500, Dallas, TX 75206, Email: dunleic@baylorhealth.edu

Abstract

Methods for sample size calculations in ROC studies often assume independent normal distributions for test scores among the diseased and non-diseased populations. We consider sample size requirements under the default two-group normal model when the data distribution for the diseased population is either skewed or multimodal. For these two common scenarios we investigate the potential for robustness of calculated sample sizes under the mis-specified normal model and we compare to sample sizes calculated under a more flexible nonparametric Dirichlet process mixture model. We also highlight the utility of flexible models for ROC data analysis and their importance to study design. When non-standard distributional shapes are anticipated, our Bayesian nonparametric approach allows investigators to determine a sample size based on the use of more appropriate distributional assumptions than are generally applied. The method also provides researchers a tool to conduct a sensitivity analysis to sample size calculations that are based on a two-group normal model. We extend the proposed approach to comparative studies involving two continuous tests. Our simulation-based procedure is implemented using the WinBUGS and R software packages and example code is made available.

Key Words: AUC, Bayesian design, Dirichlet process mixtures, ROC curve, simulation

1. Introduction

Careful planning is important in the design of studies to evaluate the performance of medical tests. The ability of a continuous test or, more generally, a classification procedure, to distinguish between two groups (e.g., disease positive, D , and disease negative, \bar{D}) is characterized by the tests' receiver operating characteristics (ROC) curve and the corresponding area under the curve (AUC). Let y denote a test score and let c denote a cutoff threshold to categorize subjects as positive or negative ($T+$ or $T-$). We adhere to the convention that higher test scores are indicative of disease presence, so $y > c$ corresponds to $T+$. For all c , the ROC curve plots a tests' true positive probability (sensitivity), denoted $\eta(c) = \Pr(y > c | D)$, against its false positive probability ($1 -$ specificity), where the specificity at cutoff c is $\theta(c) = \Pr(y < c | \bar{D})$. When planning a

test-accuracy study, a crucial element of the design is proper determination of a sample size that will enable conclusions to be drawn about AUC.

Frequentist and Bayesian sample size procedures for ROC studies have often assumed normal distributions for test scores within the D and \bar{D} populations [e.g., 1, 2]. Modeling ROC data as normal when they are skewed or multimodal can result in biased inference for AUC when there is overlap in the distributions of test scores for the D and \bar{D} populations. A multimodal distribution for the population D will occur, for instance, when subgroups of individuals are in different stages of disease; individuals in higher stages of disease will tend to have higher test scores than individuals who are newly infected. For example, Bayesian nonparametric estimates of ROC curves for two assays designed to detect Johne's disease in dairy cattle have been used as an alternative to two-group normal inference in order to account for multiple disease stages [3]. A right skewed distribution of test scores in D may occur when only a small proportion of individuals survive long enough to experience disease progression. Although a normal distribution will often accurately model population \bar{D} , it can fail when certain subgroups are more likely to have false positives due to, for instance, cross-reactions of the test to a related infectious agent. We focus on one-test and two-test scenarios with continuous data; in contrast, a hierarchical Bayesian model for sample size determination in ROC analysis of multiple ordinal tests can be found in [4].

The importance of performing sample size and power calculations using the same model that is planned to be used for data analysis is well recognized. If a nonparametric analysis of ROC data is planned, then sample size calculations would ideally use the same nonparametric model. With this motivation, the goals of the current study are: (i)

to develop a flexible Bayesian approach to sample size analysis in ROC studies using Dirichlet process mixture (DPM) models and (ii) to investigate the robustness of power calculated under the two-group normal when that model does not hold due to skewness or multimodality. The DPM model used in this paper is operationally a mixture of normal distributions with a large number of mixing components. Hence, DPMs have the flexibility to identify unanticipated multimodality, clumping, and skewness.

Throughout this paper the terms power and predictive power are used to describe the following approach to sample size determination in ROC studies. The criterion we adopt involves sampling multiple data sets from a marginal predictive distribution. Marginalization occurs with respect to a (sampling) prior distribution for all unknown model parameters. For each simulated data set, we compute the posterior probability that AUC exceeds some threshold (e.g., the AUC of a competitor test). The calculated predictive power is then a Monte Carlo average of those posterior probabilities. Similar usage of the term power in Bayesian sample size studies appears in [2]. Other methods used for sample size determination in ROC studies include the average length criterion [7] and the average variance criterion [9].

There has been much recent activity in the development of nonparametric approaches to ROC data analysis. Frequentist methods have used empirical likelihood [10], semiparametric generalized linear models [11], and kernel density estimation [12, 13] to estimate ROC curves. Bayesian DP mixtures of normals were used by [14] and [15]. A comparison of Bayesian DPMs to a two-group normal ROC analysis by [14] found that, in a particular example with sample sizes of about 1500 in each group, the parametric approach overestimated AUC when both groups have multimodal

distributions. ROC analysis using Bayesian mixtures of Polya tree models were developed by [15] and [16], but we focus solely on DPMs and do not explore Polya tree models here.

In the next section we describe the DPM model used in this study, including prior specification, posterior approximation, and AUC estimation and testing. Section 3 presents the general computational algorithm for predictive power estimation under the DPM and two-group normal procedures. Section 4 reports on a simulation study to compare sample sizes under the two procedures using data sets where normality is satisfied or not. While most of our presentation focuses on studies involving one diagnostic test, in Section 4.5 we detail sample size methods when the goal is to directly compare the accuracy of two tests. A summary of our research is provided in Section 5.

2. Parametric and Nonparametric Models

We consider designs that involve a single (imperfect) continuous diagnostic test. We assume that true disease status has been ascertained by some other means for every sampled individual. Let y_{D_i} ($i = 1, \dots, n_D$) and $y_{\bar{D}_j}$ ($j = 1, \dots, n_{\bar{D}}$) denote (possibly transformed) test scores from a random sample of individuals within the D and \bar{D} populations, respectively. The standard two-group normal model assumes independent normal distributions for test scores:

$$y_{D_i} \mid \mu_D, \sigma_D^2 \sim N(\mu_D, \sigma_D^2) \text{ and } y_{\bar{D}_j} \mid \mu_{\bar{D}}, \sigma_{\bar{D}}^2 \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}}^2).$$

For sample size determination, it is common to use a diffuse proper prior for $(\mu_D, \mu_{\bar{D}}, \sigma_D^2, \sigma_{\bar{D}}^2)$, e.g., independent normals with large variances for the mean parameters and

independent gamma or inverse gamma priors for the variances or standard deviations.

The modeled AUC is given by

$$AUC = \Phi \left(\frac{\mu_D - \mu_{\bar{D}}}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}} \right), \quad (1)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

A flexible Bayesian nonparametric alternative is the DPM normal model with a countable and infinite number of mixture components. Here, test scores are modeled as independent $N(\mu_{D_i}, \sigma_D^2)$ or $N(\mu_{\bar{D}_j}, \sigma_{\bar{D}}^2)$, where the μ_{D_i} 's and $\mu_{\bar{D}_j}$'s are treated as latent variables sampled from unspecified probability distributions G_D and $G_{\bar{D}}$, respectively. The independent prior distributions for G_D and $G_{\bar{D}}$ are taken as Dirichlet processes with normal base distributions and weight parameters α_D and $\alpha_{\bar{D}}$. Hierarchical models for the two-group data are given by:

$$\begin{aligned} y_{D_i} \mid \mu_{D_i}, \sigma_D^2 &\sim N(\mu_{D_i}, \sigma_D^2) & y_{\bar{D}_j} \mid \mu_{\bar{D}_j}, \sigma_{\bar{D}}^2 &\sim N(\mu_{\bar{D}_j}, \sigma_{\bar{D}}^2) \\ \mu_{D_i} \mid G_D &\sim G_D & \mu_{\bar{D}_j} \mid G_{\bar{D}} &\sim G_{\bar{D}} \\ G_D &\sim \text{DP}(N(\mu_{G_D}, \sigma_{G_D}^2), \alpha_D) & G_{\bar{D}} &\sim \text{DP}(N(\mu_{G_{\bar{D}}}, \sigma_{G_{\bar{D}}}^2), \alpha_{\bar{D}}). \end{aligned}$$

A popular approach to construct a Dirichlet process is the stick-breaking procedure developed by [17]. Ishwaran and James [18] connected the finite stick-breaking process with the finite Dirichlet distribution to approximate the DP using a Dirichlet-Multinomial allocation (DMA) [see also 19]. It is well-known that the DP is discrete with probability one, and thus samples from a DP will cluster. The DMA approximation to a DPM involves clustering as well. Let w_i (\bar{w}_j) denote the sampled

group corresponding to individual i in population D (individual j in population \bar{D}). Under the DMA process, individual test scores from the D (or \bar{D}) population are assigned to a cluster that is defined by a parameter $\mu_D[w_i]$ (or $\mu_{\bar{D}}[\bar{w}_j]$), where w_i (or \bar{w}_j) indexes the cluster that contains y_{Di} (or $y_{\bar{D}j}$). Denote the maximum number of mixture components by M_D or $M_{\bar{D}}$. Test scores with the same value of $\mu_D[w_i]$ (or $\mu_{\bar{D}}[\bar{w}_j]$) belong to the same mixture component. The vectors of mixture probabilities, p_D and $p_{\bar{D}}$, are modeled with Dirichlet prior distributions. Letting μ_D and $\mu_{\bar{D}}$ denote vectors of distinct mean parameters for the mixture components in populations D and \bar{D} , respectively, the DPM model in this context has the form:

$$\begin{aligned}
y_{Di} \mid w_i, \mu_D, \sigma_D^2 &\sim N(\mu_D[w_i], \sigma_D^2) & y_{\bar{D}j} \mid \bar{w}_j, \mu_{\bar{D}}, \sigma_{\bar{D}}^2 &\sim N(\mu_{\bar{D}}[\bar{w}_j], \sigma_{\bar{D}}^2) \\
\mu_D[w_i] &\sim N(\mu_{G_D}, \sigma_{G_D}^2) & \mu_{\bar{D}}[\bar{w}_j] &\sim N(\mu_{G_{\bar{D}}}, \sigma_{G_{\bar{D}}}^2) \\
w_i \mid p_D &\sim \text{Multinom}(1; p_D) & \bar{w}_j \mid p_{\bar{D}} &\sim \text{Multinom}(1; p_{\bar{D}}) \\
p_D &\sim \text{Dirichlet}\left(\frac{\alpha_D}{M_D}, \dots, \frac{\alpha_D}{M_D}\right) & p_{\bar{D}} &\sim \text{Dirichlet}\left(\frac{\alpha_{\bar{D}}}{M_{\bar{D}}}, \dots, \frac{\alpha_{\bar{D}}}{M_{\bar{D}}}\right)
\end{aligned}$$

Independent normal-gamma priors can be placed on $(\mu_{G_D}, \sigma_{G_D}^2)$ and $(\mu_{G_{\bar{D}}}, \sigma_{G_{\bar{D}}}^2)$, and independent gamma priors can be used for σ_D^2 and $\sigma_{\bar{D}}^2$. Various approaches have been used to handle α_D and $\alpha_{\bar{D}}$. Erkanli et al. [14] and [15] assigned gamma priors to the weight parameters, while [19] and [20] used a fixed value, such as 1, in conjunction with a sensitivity analysis for this choice. The choice of the maximum number of mixture components in the DMA process is also important. In their ROC study, Erkanli et al. [14] used a maximum of 10 for both M_D and $M_{\bar{D}}$.

The AUC comparing the separation of the mixture components corresponding to, say, clusters w and \bar{w} is given by:

$$AUC(w, \bar{w}) = \Phi \left(\frac{\mu_D[w] - \mu_{\bar{D}}[\bar{w}]}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}} \right), \quad w = 1, \dots, M_D; \quad \bar{w} = 1, \dots, M_{\bar{D}}. \quad (2)$$

Erkanli et al. [14] show that the overall AUC is a weighted average of the components in (2), where the weights are determined by the mixture proportions in the vectors p_D and $p_{\bar{D}}$, namely

$$AUC = \sum_{w=1}^{M_D} \sum_{\bar{w}=1}^{M_{\bar{D}}} p_D[w] p_{\bar{D}}[\bar{w}] AUC(w, \bar{w}). \quad (3)$$

The Bayesian model is fitted by generating a Monte Carlo approximation to the posterior distribution using Gibbs sampling. Estimates of ROC curves and AUC are determined by posterior means and outer percentiles. History plots, autocorrelations, and running multiple chains from a variety of starting values for a random selection of data sets are used to evaluate mixing and convergence of the Markov chain sampling procedure.

3. Power Criterion and Simulation Algorithm

We determine a sample size combination $(n_D, n_{\bar{D}})$ that yields a high predictive power of concluding that $AUC > k$ when in fact the tests' AUC exceeds k . The value of k can be determined based on the AUC of a competitor's test or it can be chosen according to a desired level of accuracy, e.g. 0.70, 0.80 or 0.90. The minimum value is $k = 0.50$, which would be used if the aim of the study is simply to establish that the test performs better than chance.

We employ a sample size criterion that is based on predictive power. Our simulation-based approach iterates between two steps: for a given sample size combination, first simulate multiple data sets under a model of interest from populations D and \bar{D} , and then fit a DPM (or two-group normal) model to each simulated data set using Gibbs sampling and calculate the posterior probability that $AUC > k$. The average of the posterior probabilities across multiple simulated data sets is the predictive power.

The procedure requires two types of prior distributions that are aligned with the two steps. *Sampling* priors are used to generate model parameters needed to simulate data sets, and *fitting* priors are used in the analysis of the generated data [21]. The purpose of sampling priors is to incorporate uncertainty about the otherwise assumed “true” values of input parameters. These distributions (e.g., normal or uniform) are usually narrowly concentrated around a presumed true value. On the other hand, fitting priors usually have large variability so as to not overly influence the power analysis.

For each simulated data set, calculate the posterior probability that $AUC > k$, namely $\Pr(AUC > k \mid y_D^f, y_{\bar{D}}^f)$, where y_D^f and $y_{\bar{D}}^f$ denote vectors of (future) test scores that are simulated from the D and \bar{D} populations, respectively. Sample sizes n_D and $n_{\bar{D}}$ can be used for the future study if, averaged over the diseased and non-diseased populations, the posterior probability is sufficiently high:

$$E[\Pr(AUC > k \mid y_D^f, y_{\bar{D}}^f)] \geq 1 - \beta. \quad (4)$$

The specific value of β depends on the nature of the problem, but, in general, reasonable choices include 0.10 and 0.20. Based on (4), we select the current sample size combination if the average posterior probability that AUC is above the benchmark k is at least as high as $1 - \beta$.

Under the two-group normal model, we require prior distributions for μ_D , $\mu_{\bar{D}}$, σ_D^2 , and $\sigma_{\bar{D}}^2$. Since the two-group normal is the base distribution of the DPM model, the same prior distributions are used for μ_{G_D} , $\mu_{G_{\bar{D}}}$, $\sigma_{G_D}^2$, and $\sigma_{G_{\bar{D}}}^2$. For instance, we can assign a sampling prior of Uniform(2.5, 3.5) to the parameter μ_D , which reflects some uncertainty about a prior guess of 3. Fitting priors for these parameters are diffuse normal / inverse gamma joint distributions.

The computing algorithm below results in estimated predictive power across a range of sample size combinations. For every selected sample size pair, we simulate B data sets. The subscript $t \in \{1, \dots, B\}$ refers to the iteration in the simulation. The following steps are used for a single sample size combination $(n_D, n_{\bar{D}})$ and are repeated across a set of sample size pairs.

1. Generate B parameter sets from sampling priors and then, for $i = 1, \dots, n_D$ and $j = 1, \dots, n_{\bar{D}}$, simulate test scores y_{Dit} and $y_{\bar{D}jt}$, $t = 1, \dots, B$, from their presumed true population distributions.
2. Fit the DPM using a fitting prior for $(\mu_{G_D}, \sigma_{G_D}^2, \mu_{G_{\bar{D}}}, \sigma_{G_{\bar{D}}}^2)$. The DMA process involves α_D and $\alpha_{\bar{D}}$, which can be set equal to constants or assigned fitting priors. Specify M_D and $M_{\bar{D}}$, and specify priors for p_D and $p_{\bar{D}}$.
3. Numerically approximate the posterior distribution of AUC_t as defined in equation (3), by first computing the individual components, $AUC(w_i, \bar{w}_i)$, in (2).
4. Numerically approximate the posterior probability that AUC_t exceeds k . A Monte Carlo approximation to this posterior probability is computed as the

proportion of iterates for which AUC_t is greater than k , namely calculate

$$r_t = \frac{1}{m} \sum_{s=1}^m I(AUC_t^s > k),$$

where $I(\cdot)$ denotes the indicator function and AUC_t^s

denotes iterate s out of m simulated values from the posterior distribution of AUC_t .

5. Calculate the average of the r_t 's, $\sum_{t=1}^B r_t / B$, which approximates the predictive power at the sample size combination $(n_D, n_{\bar{D}})$. Finally, repeat steps 1 – 5 over a range of sample size combinations to determine a sample size that achieves the desired predictive power.

To calculate predictive power under the two-group normal instead of the DPM, we use methods from [2]. Specifically, in step 2 we fit the two-group normal model and in step 3 the approximation of the posterior distribution of AUC_t follows from formula (1) instead of (3).

We implemented the algorithm for the DPM model using a combination of the software packages R and WinBUGS. The parameters and data sets are generated in R, while posterior distributions are approximated using WinBUGS. A program to perform our procedure will be posted to the website http://works.bepress.com/dunlei_cheng/4.

4. Illustrations

We first illustrate that the DPM model does not overfit normal data and gives accurate predictive power estimates when normality holds. Then we investigate the robustness, or lack thereof, of sample size requirements calculated under the two-group normal when the true distribution of test scores in population D is skewed and when it is

bimodal; both occur in applications [e.g., 3]. Finally, we consider the two-test scenario where, in population D , scores for one of the tests are right skewed and for the other test they have a mixture distribution. All simulations in this section used $B=1000$ data sets with posterior approximations based on $m=5000$ Monte Carlo iterations after a 1000 iteration burn-in. We did not identify any difficulties with convergence.

4.1. Two-group normal data

Data are generated under the two-group normal model. Suppose our best guesses for the means and variances, μ_D , $\mu_{\bar{D}}$, σ_D^2 , and $\sigma_{\bar{D}}^2$, are 3.0, 0, 2.0, and 1.0. Using formula (1), our best guess for AUC is therefore 0.96. Sample size combinations of $(n_D, n_{\bar{D}}) = (10, 10), (20, 20), \dots, (100, 100)$ are used with $k = 0.90$.

We assume that on average, test scores in the D group exceed those in the \bar{D} group, which is reflected in non-overlapping sampling priors $\mu_D \sim \text{Uniform}(2.5, 3.5)$ and $\mu_{\bar{D}} \sim \text{Uniform}(-0.5, 0.5)$. Generally, test scores tend to be more variable in the D group, so we used $\text{Uniform}(1.8, 2.2)$ and $\text{Uniform}(0.8, 1.2)$ as the sampling priors for σ_D^2 and $\sigma_{\bar{D}}^2$, respectively. Under the two-group normal model, we used a diffuse proper fitting prior that approximates the Jeffreys prior, namely $N(0, 1000)$ for means and $\text{IG}(0.001, 0.001)$ for variances. These distributions were also used in the DPM model as fitting priors for μ_{G_D} and $\mu_{G_{\bar{D}}}$, and for $\sigma_{G_D}^2$ and $\sigma_{G_{\bar{D}}}^2$. In the DPM model we set the maximum number of mixing components, M_D and $M_{\bar{D}}$, at 10 and the concentration parameters, α_D and $\alpha_{\bar{D}}$, at 1. Thus, the parameters in the Dirichlet priors on p_D and $p_{\bar{D}}$ are each 0.1.

The calculated predictive power under the parametric and DPM methods was similar across the range of sample sizes considered; the largest difference in predictive power between the two methods did not exceed 0.022. Figure 1 presents a cubic polynomial spline for each approach across $n = n_D = n_{\bar{D}}$. To achieve 80% predictive power, both methods require about 38 total subjects, while 90% predictive power is attained with 6 fewer subjects per cohort under the two-group normal than the DPM. With 100 subjects sampled from each cohort, the predictive power under the parametric model is 0.934 and it is 0.933 under the DPM.

Figure 1 about here.

Using 1000 simulated data sets, we also investigated the variability in the computed value of predictive power under the two-group normal and DPM models. For a total sample size of 20, the variation under the two-group normal was 0.068; a similar value was seen for the DPM (variance of 0.063). With larger sample sizes of $n_D = n_{\bar{D}} = 100$, the DPM and normal models showed similar variability (0.048 vs. 0.040).

We investigated the influence on predictive power of a larger degree of overlap between the distributions of test scores for the diseased and non-diseased cohorts. Specifically, we changed the sampling prior mean of $\mu_{\bar{D}}$ from 0 to 2, so our best guess for the AUC is approximately 0.72. We found that, across sample size combinations ranging from 10 per cohort to 100 per cohort with $k = 0.6$ and $k = 0.65$, predictive power for the two-group normal analysis is only about 0.004 larger than predictive power computed under the DPM (Table 1).

Table 1 about here.

The selection of the maximum number of mixture components (M_D and $M_{\bar{D}}$) in the DPM model was also explored in this setting. We let $\alpha_D = \alpha_{\bar{D}} = 1$ and $M_D = M_{\bar{D}} = 5, 10, \text{ or } 20$. Therefore, p_D or $p_{\bar{D}}$ was changed from 0.2 to 0.1 to 0.05. The calculated predictive power was very similar for these three choices (Table 2). The computational time under $M_D = M_{\bar{D}} = 10$ was about $2/3$ of what it was with $M_D = M_{\bar{D}} = 20$. Thus, using 5 or 10 may be a reasonable initial choice, and a single run with a larger value can be used to assess how results change when more structure is allowed in the model.

Table 2 about here.

4.2. Skewed data

We compared the predictive power calculated under the two-group normal and DPM models when the D group has test scores that vary according to a right skewed distribution. We are particularly interested in the potential for robustness of the two-group normal model in terms of predictive power, without resorting to the use of a data transformation. The Box-Cox power transformation is a popular method to induce normality, but it is difficult to determine its transformation parameter in sample size problems without access to a sufficient amount of pilot data. We note, however, that in the rare case that an appropriate normal-generating transformation was known ahead of seeing the data, then it would be known that the normal-normal model would be

appropriate for the future data, in which case it would be unnecessary to consider our nonparametric approach.

The y_{d_i} 's were sampled from an exponential(θ) distribution with θ drawn from a Gamma(20, $20\sqrt{2}$) sampling prior. We therefore have as a best guess that diseased subjects have a mean test score of $\sqrt{2}$ with a variance of 2 and allow for uncertainty about these values. Since a normal distribution will often be appropriate for the \bar{D} group, the $y_{\bar{d}_j}$'s were generated using the same sampling model and priors as in Section 4.1. Therefore, the "true" AUC is in a neighborhood of 0.79, which we used as our best guess. We used the same reference fitting priors and we compared the predictive power achieved with a total of 20 subjects up to 200 total subjects. Two comparison sizes were used, namely $k = 0.70$ and $k = 0.75$.

For both sizes, the predictive power estimated from the DPM is larger compared to the power from the mis-specified two-group normal model, as demonstrated in Table 3. When $k = 0.70$, the predictive power under the parametric and DPM models is similar except at small sample sizes. However, with the higher effect size, the difference in predictive power between the two approaches increases. For $k = 0.75$, the predictive power under the DPM is on average about 6.3% higher than for the parametric method. For 70% power with an effect size of 0.75, 27 diseased subjects are needed according to the DPM model whereas at least 50 diseased subjects are needed according to the two-group normal.

Table 3 about here.

We noted that when $k = 0.70$ and the sample sizes are large for both the D and \bar{D} groups, the predictive power under the two-group normal and DPM approaches is similar. This occurs, in part, because at large sample sizes the two-group normal appears to overestimate and underestimate different portions of the ROC curve. Figure 2 presents estimates of the ROC curve using data generated as the 1/101, 2/101, ..., 100/101 quantiles from the exponential distribution with mean $\sqrt{2}$ for the D group and standard normal for the \bar{D} group. The DPM model yields an estimated ROC curve that more closely tracks the empirical ROC curve. The two-group normal model appears to drastically overestimate and underestimate different portions of the ROC curve. This realistic example illustrates the utility of flexible models for both ROC data analysis and sample size determination.

Figure 2 about here.

4.3. Bimodal data

Here we compare the sample size requirements under the two-group normal and DPM models for bimodal data. We assume population \bar{D} follows the normal distribution with the same sampling and fitting priors as in Section 4.1, but population D is bimodal. Specifically, the sampling model of y_{Dj} is the mixture model $\pi_{D1}N(\mu_{D1}, \sigma_{D1}^2) + \pi_{D2}N(\mu_{D2}, \sigma_{D2}^2)$, where $\pi_{D1} \sim \text{Beta}(20, 30)$ and $\pi_{D2} = 1 - \pi_{D1}$. The parameters μ_{D2} and σ_{D2}^2 were given the same sampling priors as μ_D and σ_D^2 in Section 4.1, i.e., $\text{Uniform}(2.5, 3.5)$ and $\text{Uniform}(1.8, 2.2)$, respectively. Using a sampling prior for σ_{D1}^2 of $\text{Uniform}(7.8, 8.2)$, the mixture data had a variance of 2. We considered two

sets of sampling priors for μ_{D1} , one with mean 0 the other with mean 1. Under these two sampling priors, the corresponding AUCs are about 0.85 and 0.90, respectively. In both cases we used the same diffuse fitting priors as in Section 4.1. Predictive power was calculated to detect $\text{AUC} > 0.80$. Of particular interest was whether the difference in predictive power between the two-group normal and DPM model changes with decreased test accuracy, i.e. with more overlap between the data distributions for the D and \bar{D} populations.

Figure 3 shows that, across the range of sample sizes considered, the amount of separation between distributions of test scores in this setting does not have a practical impact on the power difference between the DPM and two-group normal model. For instance, with AUC of 0.85, the predictive power under the DPM, averaged over the sample sizes considered, is only 0.001 larger than for the two-group normal. The difference in predictive power increases slightly to 0.003 when the “true” AUC is raised to 0.90. However, an interesting secondary result in this setting is that the DPM method generates increasing power functions whereas, especially for AUC 0.85, the predictive power function under the two-group normal exhibits a zig-zag pattern and decreases for sample sizes of 60 to 80, 100 to 120, and 180 to 200.

Figure 3 about here.

The “true” (hypothetical) AUC in the case of a bimodal distribution for individuals who are D and a single normal distribution for population \bar{D} can often be well-approximated by an AUC that is obtained using a single normal distribution in place

of the true bimodal density, where the mean is the mixture mean and the variance is the true variance of the mixture. We calculated the "true" and "approximate" AUC values over a range of possible inputs and found near equivalence. On the other hand, it was also easy to construct scenarios where this was not the case. For example, with a 0.5 mixture of $N(-1,1)$ and $N(3,4)$ distributions for the population D , and a $N(0,1)$ for the population \bar{D} , the "true" and "approximate" AUCs are 0.576 and 0.644, respectively. We found many other such instances as well.

Thus, if the normal-normal model is used, the estimate of the population D density will be a single normal and consequently the estimated AUC in this setting will be an estimate of the "approximate" AUC. So in the above instance (0.576 versus 0.644), the sample size expected to detect a larger (false) value will be different than it would be when using the more appropriate nonparametric approach, which would be targeting the true value.

4.4. Serologic data example

Hanson et al. [15] analyzed serologic data for Johne's disease in dairy cattle using DPM models. Test scores from the non-infected cows had an approximate normal distribution with mean value of 2 and standard deviation of 0.57. However, test values from the infected cows were bimodal with modes located at 2 and 5, respectively. We used the posterior information from [15] to simulate data with a best guess of AUC at 0.89, and compare the predictive power at cutoff $k = 0.80$ from the two-group normal and the DPM model. The parametric method gave predictive power of 0.801 and 0.817 with a total sample size of 80 and 100, respectively, and the DPM approach led to similar

values of 0.803 and 0.813. Unlike with skewed data, in this example with normal and bimodal data, the two methods again perform similarly.

4.5. Two diagnostic tests

We extend our one-sample approach to studies that directly compare two tests. In the previous sections, we have assumed that a standard test has been in use for some time and that its AUC is well established. When the operating characteristics of the “standard” test are not established, a paired design is more appropriate. In this scenario two tests are applied to each individual, where for ease of exposition we continue to assume that one is a standard (S) test and the other test is a newly (N) developed competitor. Additionally, in this section we consider a sample size criterion that targets precise estimation of the difference in AUCs between two tests. The average length criterion used here finds the minimum sample size such that 95% posterior intervals for the difference in AUCs have a specified average width.

The study goal is to establish the superiority of the new test relative to the standard test. The pairs (S, N) of test scores for non-diseased individuals, $(y_{\bar{D}j \bullet S}, y_{\bar{D}j \bullet N})$, are assumed to be correlated (ρ is about 0.8) and to follow a bivariate normal distribution with mean vector $(0, 0)$ and variance vector $(1, 1)$. For diseased individuals, we assume the $y_{Di \bullet S}$'s have the same exponential distribution used in Section 4.2 (with mean of $\sqrt{2}$ and variance of 2), and that $y_{Di \bullet N}$ is a linear shift of $y_{Di \bullet S}$, specifically through the regression $y_{Di \bullet N} = 2.3 + 0.5 y_{Di \bullet S} + \varepsilon_i$, where ε_i is normal(0, 1.5). Therefore, the difference in AUCs between the N and S tests is approximately 0.15, given that AUC_N is about 0.96 and AUC_S is about 0.81.

We compare both average length and predictive power when bivariate normal and DPM models are fitted to simulated data sets, in both cases using diffuse priors. We investigate the predictive power for testing $AUC_N - AUC_S > 0.05$ and set the desired average length at 0.15. Results in Table 4 show that, in this scenario, the average lengths are comparable under both methods, with a sample size of 50 from each group achieving the desired average length under both approaches. However, predictive power under the DPM is consistently lower than that under the parametric model at the sample sizes considered, reflecting that the two methods lead to different AUC estimates. We found that, across the sample size combinations considered, the two-group normal model consistently underestimated AUC_S whereas the DPM model produced estimates that were close to the best guess of 0.81, especially as sample size increased. Meanwhile, both methods provided similarly accurate estimates for AUC_N , because the data distribution under the new test for the \bar{D} population is normal and it is approximately normal for the D population. As a result, the posterior means of $AUC_N - AUC_S$ under the two-group normal method were on average 8.1% higher across simulated data sets than those under the DPM method. This led to the illusion of higher predictive power at each sample size combination under the normal model compared to the DPM method. This example illustrates the importance of monitoring the individual AUCs in addition to their difference when using predictive power for sample size determination in paired designs.

Table 4 about here.

5. Conclusions

We developed a simulation-based approach that uses nonparametric Dirichlet process mixtures for sample size and predictive power calculations in ROC studies designed to compare a continuous medical test to a known standard test. An advantage of the DPM model for ROC data analysis is its capability to accommodate non-standard features such as skewness and multimodality that a default two-group normal analysis will fail to identify. We showed that predictive power computed under the traditional parametric approach may not be robust to model mis-specification when data are right skewed, but can be fairly robust to bimodal data within the diseased population. For bimodal data in population D and univariate normal data in population \bar{D} , the similarity in predictive power under the DPM and two-group normal approaches is not surprising since it is common in this setting for the region of distributional overlap to mimic the shape of the two-group normal. A further finding based on the average length criterion in paired designs when the study goal is to estimate the difference between two AUCs was that, although the DPM is a much more highly structured model than the normal-normal, the precision in interval estimates was similar for the two approaches in a particular simulation study.

We emphasize that a reason estimated AUCs can be the same in both skewed and bimodal situations is because the normal approximation for population D is often adequate in terms of estimating $\text{AUC}=\Pr(X > Y)$, where random variable X corresponds to test scores from population D and random variable Y corresponds to test scores from population \bar{D} , while at the same time the corresponding estimate of $\text{ROC}(t)$ is biased high for small t followed by biased low for large t , resulting in cancellation when

calculating AUC. Therefore, the subsequent data analysis may result in a biased estimate of an ROC curve (and AUC) when the two-group normal model is applied to data that are skewed or multimodal. Thus, we recommend using a nonparametric power specification and a subsequent nonparametric data analysis. The DPM method is useful for sample size determination for ROC studies since it is partly immune to model mis-specification. An example of software code to implement the DPM and two-group normal methods that is easily modifiable can be obtained on our website or by emailing the corresponding author.

Sample size calculations are often performed using simpler models for both the data generating distribution and the planned model to be used to analyze the data. Our approach, on the other hand, allows for the use of more complex data generating distributions since our nonparametric modeling of the data in the sample size calculation will adapt appropriately to complexity of the data. Primary motivations in this report are thus to advertise the use of nonparametric methods for ROC data analysis and to provide methods for corresponding sample size and predictive power calculations, subsequently matching sample size analysis and data modeling. Since a major goal is the estimation of the ROC curve itself, and since parametric models for data are often overly restrictive, the DPM model seems ripe for use in making inferences. It only makes sense then to perform predictive power calculations using the same flexible nonparametric family of models since it should provide a more accurate quantification of sample sizes needed in subsequent analysis.

The question arises: when would one decide to select sample size based on our hypothesis testing criterion versus the estimation criterion? It is common in medical

testing to have a standard “reference” marker that has been in use over some period of time by a variety of researchers and practitioners using different cutoffs due to different sensitivity-specificity requirements. Higher cutoffs usually lead to higher specificity and lower sensitivity. Moreover, it is also common for new markers to be developed with the goal of providing higher sensitivity-specificity pairings across cutoffs than would be achieved with the standard marker, which amounts to having a corresponding ROC curve that dominates the curve for the standard test. If the new marker can be shown to have an AUC that is appreciably greater than that for the standard marker, a step will have been taken in this direction. Thus, our hypothesis testing criterion for selecting sample size applies directly when it is of interest to make a decision about the preference for one marker over the other. On the other hand, the estimation criterion would be used when it is desired to establish, precisely, the value of an AUC for a single marker or the difference in AUCs for two markers. In either case, the goal is to find a sample size that will give sufficient precision for inference.

Throughout this paper we have assumed that true disease status could be determined by some other means than the test(s) under study. When this “gold-standard” setting fails to hold and true infection status of sampled individuals is unknown, additional structure is needed to model and impute the latent status. Branscum et al [16] developed a semiparametric approach for joint analysis with no gold-standard data. By ignoring the added uncertainty attached to unknown infection status, sample size estimates could fall far below what is actually required for a test accuracy study. Further research is needed to extend the gold-standard methods presented in the current study to a nonparametric framework for sample size determination in designs with unidentified

disease status. For instance, the parametric methods in [2] for continuous tests without a gold-standard could potentially be extended nonparametrically, or the nonparametric model in [16] could be applied directly to the task of sample size determination. Both approaches require input about disease prevalence in the source population, and informative priors become mandatory in the presence of lack of model identifiability. Similar research for binary tests appears in [7].

Acknowledgements

The authors are grateful to two anonymous reviewers for their comments, which helped improve the paper.

References

1. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving two-group normal ROC curve indices. *Statistics in Medicine* 1997; **16**: 1529-1542.
2. Cheng D, Branscum AJ, Stamey JD. A Bayesian approach to sample size determination for studies designed to evaluate continuous medical tests. *Computational Statistics and Data Analysis* 2010; **54**: 298-307.
3. Hanson TE, Branscum, AJ, Gardner IA. Multivariate mixtures of Polya trees for modeling ROC data. *Statistical Modeling* 2008; **8**: 81-96.
4. Wang F, Gatsonis CA. Hierarchical models for ROC curve summary measure: design and analysis of multi-reader, multi-modality studies of medical test. *Statistics in Medicine* 2008; **27**: 243-256.
5. Cheng D, Stamey JD, Branscum AJ. Bayesian approach to average power calculation for binary regression with misclassified outcomes. *Statistics in Medicine* 2009; **28**: 848-863.
6. Cheng D, Branscum AJ, Stamey JD. Accounting for response misclassification and covariate measurement error improves power and reduces bias in epidemiologic studies. *Annals of Epidemiology* 2010; **20**: 562-567.
7. Dendukuri N, Rahme E, Bélisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics* 2004; **60**: 388-387.
8. Cheng D, Stamey JD, Branscum AJ. A general approach to sample size determination for prevalence surveys that use dual test protocols. *Biometrical Journal* 2007; **49**: 694-706.
9. Stamey JD, Seaman JW, Young DM. Bayesian sample-size determination for inference on two binomial populations with no gold standard classifier. *Statistics in Medicine* 2005; **24**: 2963-2976.

10. Qin G, Zhou X-H. Empirical likelihood inference for the area under the ROC curve. *Biometrics* 2006; **62**: 613-622.
11. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press. 2003.
12. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997; **16**: 2143-2156.
13. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 1998; **93**: 1356-1364.
14. Erkanli A, Sung M, Costello EJ, Angold A. Bayesian semi-parametric ROC analysis. *Statistics in Medicine* 2006; **25**: 3905-3928.
15. Hanson TE, Kottas A, Branscum AJ. Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian nonparametric approaches. *Applied Statistics* 2008; **57**: 207-225.
16. Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine* 2008; **27**: 2474-2496.
17. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**: 639-650.
18. Ishwaran H James, LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; **96**: 161-173.
19. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000; **9**: 249-265.
20. Hanson TE, Branscum AJ, Johnson WO. Bayesian nonparametric modeling and data analysis: an introduction. In Handbook of Statistics, volume 25, ed. Dey DK, Rao CR, 245-278. Elsevier. 2005.
21. Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* 2002; **17**: 193-208.

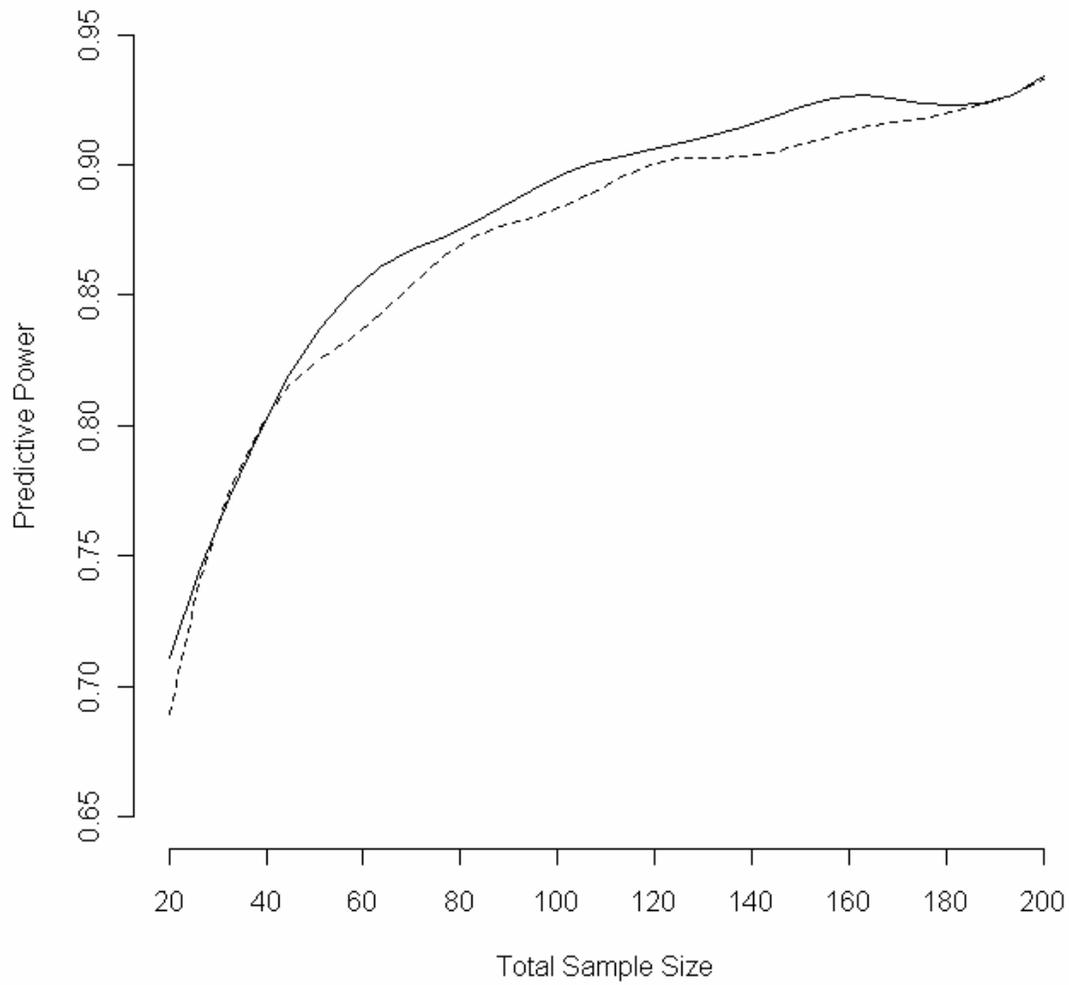


Figure 1. Predictive power with $k = 0.90$ for the two-group normal (solid line) and DPM (dashed line) models when normal distributions are used to generate the data from the diseased and non-diseased populations.

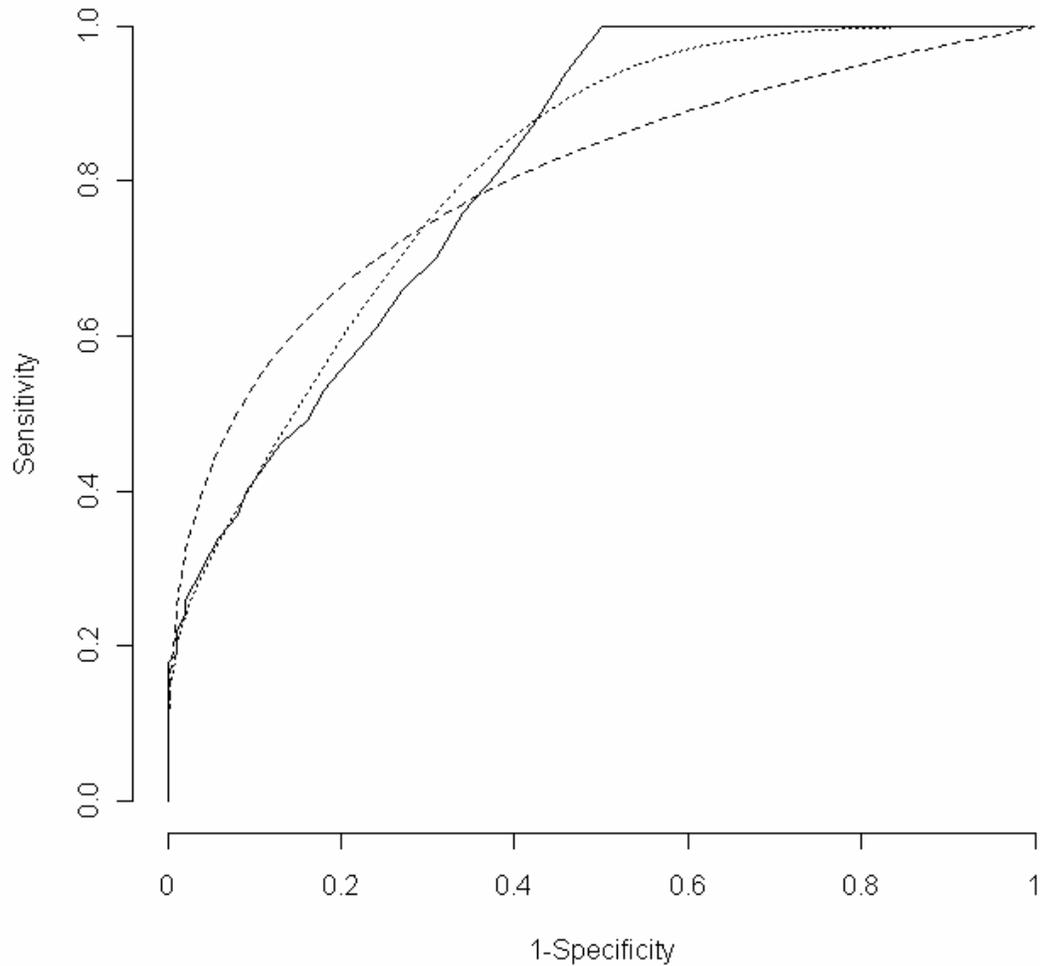


Figure 2. Three estimates of the ROC curve when test scores of diseased subjects follow an exponential distribution with mean $\sqrt{2}$ and scores from non-diseased subjects are distributed standard normal ($n_D = n_{\bar{D}} = 100$). The empirical estimate is plotted as a solid line, the Bayesian parametric estimate as a dashed line, and the DPM estimate as a dotted line.

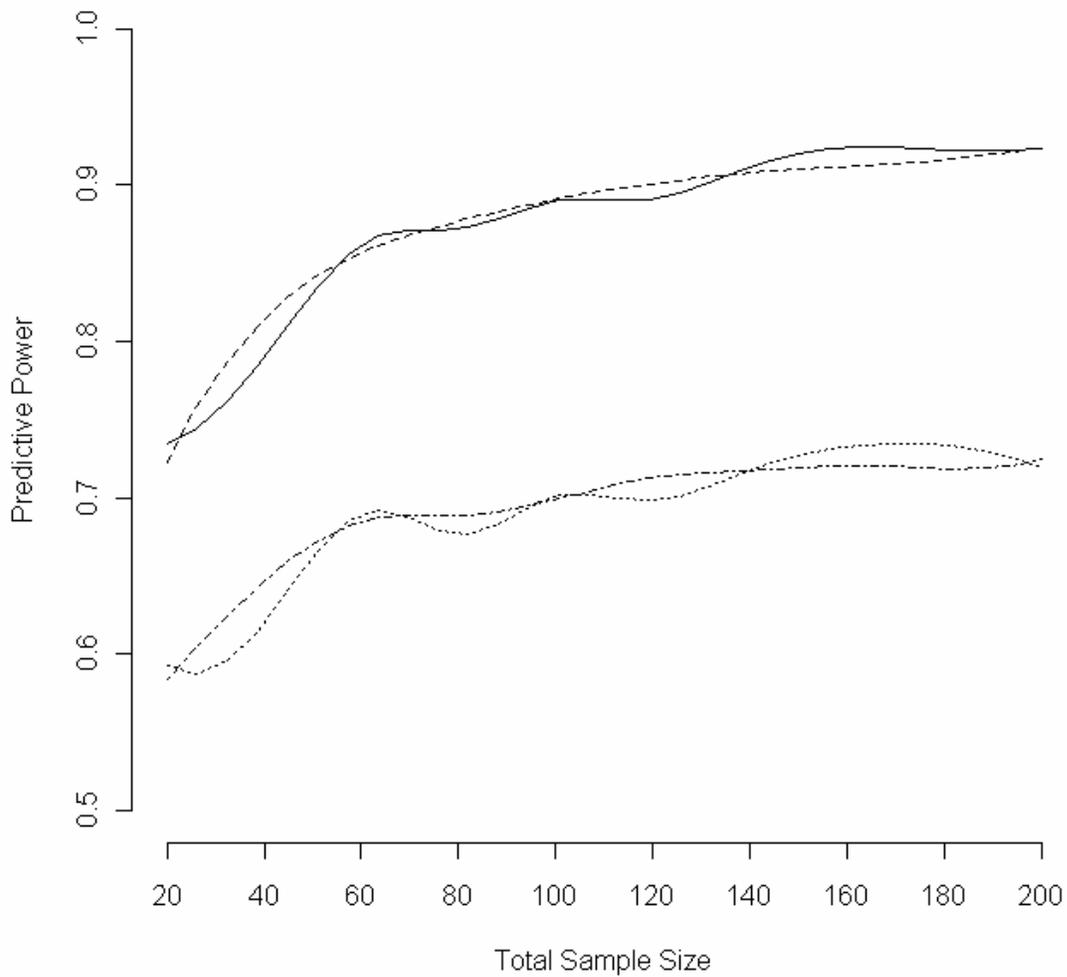


Figure 3. Predictive power curves when bimodal distributions generate data from the diseased group, with a univariate normal distribution for the non-diseased group, using $k = 0.80$. The solid line is for the two-group normal model and the dashed line is for the DPM analysis when AUC = 0.90. The dotted line is for the two-group normal and the dashed-dotted line is for the DPM analysis when AUC = 0.85.

Table 1: Predictive power comparison when the disease and disease-free groups follow a normal distribution: (1) $k = 0.6$ for the two-group normal model; (2) $k = 0.6$ for the DPM model; (3) $k = 0.65$ for the two-group normal; (4) $k = 0.65$ for the DPM model.

Sample Size	Power 1	Power 2	Power 3	Power 4
(10, 10)	0.712	0.702	0.614	0.604
(20, 20)	0.787	0.789	0.676	0.678
(30, 30)	0.818	0.804	0.702	0.687
(40, 40)	0.832	0.831	0.708	0.706
(50, 50)	0.848	0.840	0.724	0.722
(60, 60)	0.854	0.850	0.720	0.724
(70, 70)	0.861	0.848	0.735	0.723
(80, 80)	0.873	0.860	0.733	0.732
(90, 90)	0.863	0.863	0.747	0.736
(100, 100)	0.876	0.880	0.753	0.760

Table 2: Predictive power comparison for three different maximum number of DPM

mixture components, with $\alpha_D = \alpha_{\bar{D}} = 1$ and $k = 0.9$: (1) $M_D = M_{\bar{D}} = 5$; (2)

$M_D = M_{\bar{D}} = 10$; (3) $M_D = M_{\bar{D}} = 20$.

Sample Size	Power 1	Power 2	Power 3
(20, 20)	0.802	0.804	0.803
(40, 40)	0.870	0.870	0.869
(60, 60)	0.900	0.901	0.900
(80, 80)	0.913	0.913	0.913
(100, 100)	0.932	0.933	0.932

Table 3: Predictive power comparison when the D group has a skewed data distribution:

(1) $k = 0.7$ for the two-group normal model; (2) $k = 0.7$ for the DPM model; (3)

$k = 0.75$ for the two-group normal; (4) $k = 0.75$ for the DPM model.

Sample Size	Power 1	Power 2	Power 3	Power 4
(10, 10)	0.710	0.734	0.591	0.624
(20, 20)	0.769	0.806	0.618	0.687
(30, 30)	0.831	0.832	0.676	0.705
(40, 40)	0.841	0.854	0.681	0.732
(50, 50)	0.868	0.876	0.704	0.743
(60, 60)	0.869	0.878	0.706	0.754
(70, 70)	0.875	0.886	0.709	0.756
(80, 80)	0.890	0.890	0.724	0.758
(90, 90)	0.891	0.892	0.730	0.761
(100, 100)	0.885	0.885	0.712	0.758

Table 4: Average length and predictive power comparisons for the two-test scenario computed under the bivariate normal model (1) and the DPM model (2).

Sample Size	Length 1	Length 2	Power 1	Power 2
(10, 10)	0.363	0.400	0.809	0.756
(20, 20)	0.244	0.254	0.887	0.850
(30, 30)	0.194	0.201	0.926	0.887
(40, 40)	0.168	0.170	0.944	0.891
(50, 50)	0.152	0.150	0.959	0.908