Published in final edited form as:

Stat Med. 2006 February 15; 25(3): 459–479.

Bayesian multivariate hierarchical transformation models for ROC analysis

A. James O'Malley^{1,*†} and Kelly H. Zou^{1,2}

¹Department of Health Care Policy, Harvard Medical School, Boston, MA 02115, U.S.A.

²Department of Radiology, Brigham and Women's Hospital, Boston, MA 02115, U.S.A.

SUMMARY

A Bayesian multivariate hierarchical transformation model (BMHTM) is developed for receiver operating characteristic (ROC) curve analysis based on clustered continuous diagnostic outcome data with covariates. Two special features of this model are that it incorporates non-linear monotone transformations of the outcomes and that multiple correlated outcomes may be analysed. The mean, variance, and transformation components are all modelled parametrically, enabling a wide range of inferences. The general framework is illustrated by focusing on two problems: (1) analysis of the diagnostic accuracy of a covariate-dependent univariate test outcome requiring a Box–Cox transformation within each cluster to map the test outcomes to a common family of distributions; (2) development of an optimal composite diagnostic test using multivariate clustered outcome data. In the second problem, the composite test is estimated using discriminant function analysis and compared to the test derived from logistic regression analysis where the gold standard is a binary outcome. The proposed methodology is illustrated on prostate cancer biopsy data from a multi-centre clinical trial.

Keywords

Bayesian methods; hierarchical models; multivariate analysis; receiver operating characteristic (ROC) curve; Box–Cox transformation

1. INTRODUCTION

A popular graphical tool for assessing the accuracy of a diagnostic test is a receiver operating characteristic (ROC) curve which accounts for the inherent trade-off between the sensitivity and specificity of the test. The ROC curve is a plot of all possible pairs of the false positive rate (1-specificity) and true positive rate (sensitivity) of the test [1]. An empirical ROC curve plots the observed pairs of these fractions as the positivity threshold for the test is varied over all possible values. Smooth ROC curves can be easily obtained by fitting a parametric binormal model to the two-sample data [2].

Recently, several regression procedures for ROC analysis have been proposed (e.g. References [3,4]). Rather than add to the growing frequentist methodology for ROC analysis (see [5] for an overview of methods), here we adopt Bayesian parametric regression and consider two important new applications of ROC analysis. Bayesian methodology is now used in many areas

*Correspondence to: A. James O'Malley, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

†E-mail: omalley@hcp.med.harvard.edu

Contract/grant sponsor: U.S.A. National Institutes of Health; contract/grant numbers: R01LM7861, R03HS13234, PO1CA41167, UO1CA45256

of statistical application, including ROC curve analysis [6-9]. However, this method has not yet dealt with clustered data with continuously distributed diagnostic outcomes. An advantage of the Bayesian paradigm is that it provides a logical procedure for resolving complicated inferences. Bayesian methods are particularly well-equipped to contend with problems involving hierarchically clustered data.

The applications of ROC analysis to be considered are: (1) analysis of hierarchical structured outcome data from a diagnostic test that requires cluster-specific monotone transformations; (2) determination of the optimal combination of clustered multivariate test outcomes. A Bayesian multivariate hierarchical transformation model (BMHTM) is developed for both applications. An important aspect of application (1) is whether different transformations of the outcome are needed for each cluster, and the impact of the transformation used on the resulting ROC curves. The second problem is important because a composite test has the potential to yield superior overall performance than a single test.

Multilevel data with clustered responses often arise in diagnostic medicine. Such data consist of the true disease status (i.e. the gold standard), individual-level characteristics (e.g. demographic information and clinical history), and cluster-level covariates (e.g. clinical experience and institutional characteristics). For example, in radiologic studies patients may be nested completely within hospitals, or images might be interpreted independently by different radiologists [7]. The performance of tests that involve repeated measurements over a series of times (e.g. longitudinal biomarkers) may be described using time dependent ROC curves [10-12].

Outcomes with skewed distributions are common with continuous diagnostic data. Ignoring skewness by applying a symmetric bi-distributional model may lead to biased results [13]. To account for clustering and skewness, a monotonically transformed outcome that better satisfies the model assumptions may be fitted using a subject-specific mixed effects regression model [14]. Historically, convenient transformations have often been chosen based on *ad hoc* considerations, or estimated prior to the main ROC analysis, despite the fact that an inappropriate choice of transformation may lead to lack-of-fit and erroneous inferences [15]. Semi-parametric [4,16,17] and non-parametric [18,19] approaches provide alternative ways of accounting for skewness and other distributional difficulties, but they tend to become burdensome when extended to account for clustering.

We develop a model embedding an optimal cluster-specific Box-Cox transformation that allows the uncertainty in the transformation parameters to propagate through all inferences. Using cluster-specific transformation parameters accounts for heterogeneity between clusters in the shape of the outcome distribution. A similar rationale is given in Cole [20] for the use of different transformation parameters across strata (e.g. age groups) of a population. However, in contrast to Cole's model-fitting procedure, we use formal likelihood-based Bayesian methods for parameter estimation and the evaluation of other inferences.

For application (2) we focus on multivariate outcomes with the goal of combining multiple tests into a single (composite) test. Multivariate outcomes present additional analytical challenges, particularly due to the need to account for the possible correlation between the outcomes. However, by analysing multiple outcomes simultaneously, we are able to pool information across tests enabling improved inference about a single test outcome, and greater accuracy in predicting disease status.

A composite test that is optimal under a given model can be constructed by either conditioning on the gold standard and modelling the diagnostic data as the outcomes, or by treating the diagnostic data as covariates and modelling the gold standard (binary in typical ROC analyses). The former approach is discriminant analysis and usually assumes normally distributed

diagnostic data, while the latter typically involves a generalized linear model such as logistic regression. Although the normal discriminant function and logistic regression models are asymptotically equivalent [21], they are different procedures and therefore yield different results when applied to the same data set. The sensitivity of a composite test to these procedures has not been previously investigated. However, Slate and Turnbull [10] used ROC curves to compare different decision rules for deciding if an individual has prostate cancer. A distinctive feature of our problem is that the decision rules being compared are based on different models, whereas Slate and Turnbull's are based on the same model.

In Section 2 we introduce notation and derive the BMHTM for evaluating multivariate clustered diagnostic marker data. The motivating prostate cancer biopsy data are described in Section 3. In Section 4 we derive several summary statistics of diagnostic accuracy for parametric transformations of clustered outcome data. These functions involve complicated functions of the model parameters making inferences difficult to evaluate using classical methods whereas a Bayesian calculation is straight-forward to formulate. Section 5 considers methodology for deriving a single composite test derived from multiple test outcomes, and for comparing the relationship between the discriminant function and logistic regression approaches. In Section 6 results are presented for the prospective multi-institutional prostate cancer biopsy data, and the conclusion is given in Section 7.

2. BAYESIAN MULTIVARIATE HIERARCHICAL TRANSFORMATION MODEL (BMHTM)

Let $\mathbf{X}_{ij} = (X_{1ij}, \dots, X_{Kij})$ be a vector of K outcomes, and $\mathbf{Z}_{ij}$ an associated vector of covariates, measured for the j th subject in the i th cluster ($i=1, \dots, m; j=1, \dots, n_i$). Suppose each subject belongs to one of two populations, diseased and non-diseased, as specified by a gold standard. The diseased and non-diseased populations are denoted by $T_{ij} = 1$ and $T_{ij} = 0$, respectively. Let $h(\cdot)$, $\mu(\cdot)$, and $\sigma^2(\cdot)$ denote the transformation, mean, and variance functions, respectively.

The following hierarchical model is the basis of the methods developed herein:

$$\text{Level I: } h(X_{kij}, \lambda_{ki}) = \mu(T_{ij}, \mathbf{Z}_{ij}, \boldsymbol{\beta}_{ki}) + \sigma(T_{ij}, \mathbf{Z}_{ij}, \boldsymbol{\theta}_{ki}) \varepsilon_{kij} \quad (1)$$

$$\text{Level II: } \boldsymbol{\beta}_{ki} = \boldsymbol{\beta}_{k0} + \mathbf{b}_{ki}, \quad \boldsymbol{\theta}_{ki} = \boldsymbol{\theta}_{k0} + \mathbf{t}_{ki}, \quad \lambda_{ki} = \lambda_{k0} + l_{ki} \quad (2)$$

where β_{ki} and θ_{ki} are the regression and variance parameters of diagnostic test k ($k = 1, \dots, K$) for hospital i ($i = 1, \dots, m$); $\mathbf{b}_{ki}$ and $\mathbf{t}_{ki}$ are random effects which are independent across the outcomes and hospitals, with $\mathbf{b}_{ki} \sim N(\mathbf{0}, \Sigma_k)$ and $\mathbf{t}_{ki} \sim N(\mathbf{0}, \Omega_k)$, where $N(\cdot)$ denotes the multivariate normal density and $\mathbf{0}$ is a vector of K zeros; λ_{ki} is the transformation parameter for test k in hospital i ; l_{ki} is a random effect independent across (k, i) with $l_{ki} \sim N(0, \tau_k^2)$; and $\varepsilon_{ij} \sim N(\mathbf{0}, \Psi)$ is a vector of K within individual errors independent of $(\mathbf{b}_{ki}, \mathbf{t}_{ki}, l_{ki})$. In ROC curve analysis Ψ typically a non-diagonal matrix.

The above model incorporates random effects within the mean, variance, and transformation components of the model. Furthermore, it allows for correlation between the level II error terms for the mean and variance function parameters, in addition to correlation due to clustering and correlation between $\mathbf{b}_{ki}$ over outcomes k . The model can be extended further by allowing general (i.e. non-normal) distributions for the random effects; however, we do not consider such models in the present application.

Regression analysis of continuous responses often assumes that the variance of the outcomes does not depend on covariates. However, this assumption is an over-simplification in analysing diagnostic data because diseased and non-diseased patients typically have very different

distributions [6]. Therefore, the variance function should at least depend on the gold standard, T_{ij} .

In a special case of the model, the effect of the gold standard varies by cluster, and the remaining covariates ($\mathbf{Z}_{ij}$) impact the mean but not the variance. The mean and variance functions then become

$$\mu(T_{ij}, \mathbf{Z}_{ij}, \boldsymbol{\beta}_{ki}) = \beta_{ki0} + \beta_{ki1} T_{ij} + \boldsymbol{\beta}_k' \mathbf{Z}_{ij} \quad (3)$$

and

$$\sigma^2(T_{ij}, \boldsymbol{\theta}_{ki}) = (\theta_{ki0} + \theta_{ki1} T_{ij})^2 \quad (4)$$

where, consistent with previous notation, $\boldsymbol{\beta}_{ki} = (\beta_{ki0}, \beta_{ki1}, \boldsymbol{\beta}_k')'$.

In the present analysis, the Box–Cox transformation [22] given by

$$X_{kij}^{(\lambda_{ki})} = h(X_{kij}, \lambda_{ki}) = \begin{cases} (X_{kij}^{\lambda_{ki}} - 1) / \lambda_{ki} & \text{if } \lambda_{ki} \neq 0 \\ \log(X_{kij}) & \text{if } \lambda_{ki} = 0 \end{cases} \quad (5)$$

where λ_{ki} is an unknown parameter, defines transformed outcomes $X_{kij}^{(\lambda_{ki})}$ that are assumed to be normally distributed. The purpose of (5) is to facilitate valid comparisons of diagnostic accuracy by standardizing the outcome distributions across clusters. It is important to realize that the same transformation must be applied to diseased and non-diseased subjects within the same cluster. Otherwise the ability to make comparisons between diseased and non-diseased subjects is lost. However, the transformations need not be the same over different clusters; if it is believed that the shape of the outcome distribution varies across clusters, then different transformations may be necessary in order to accommodate the analysis of data from multiple clusters in a single model [20].

The likelihood function for the model defined by (3)–(5) is given by

$$L \propto \prod_{j=1}^m \prod_{k=1}^K \left[\prod_{i=1}^{n_i} \sigma(T_{ij}, \boldsymbol{\theta}_{ki})^{-1} X_{kij}^{\lambda_{ki}-1} \exp \left\{ - \frac{(X_{kij}^{(\lambda_{ki})} - \mu(T_{ij}, \mathbf{Z}_{ij}, \boldsymbol{\beta}_{ki}))^2}{2\sigma^2(T_{ij}, \boldsymbol{\theta}_{ki})} \right\} \right] \cdot \left\{ \det(\Sigma_b \Sigma_t) \tau_k^2 \right\}^{-1/2} \exp \left\{ (\boldsymbol{\beta}_{ki} - \boldsymbol{\beta}_{k0})' \Sigma_k^{-1} (\boldsymbol{\beta}_{ki} - \boldsymbol{\beta}_{k0}) / 2 \right\} \cdot \exp \left\{ - (\boldsymbol{\theta}_{ki} - \boldsymbol{\theta}_{k0})' \Omega_k^{-1} (\boldsymbol{\theta}_{ki} - \boldsymbol{\theta}_{k0}) / 2 - (\lambda_{ki} - \lambda_{k0})^2 / (2\tau_k^2) \right\} \quad (6)$$

In Bayesian analysis the likelihood function is augmented with a prior distribution for β_{k0} , Σ_k , θ_{k0} , Ω_k , λ_{k0} , and τ_k^2 . Because of the non-linearity introduced by the Jacobian for the transformation from X_{kij} to $X_{kij}^{(\lambda_{ki})}$ and the non-constant variance function, closed form solutions to Bayesian inferences based on the likelihood function in (6) are in general intractable. In practice, numerical simulation methods are used instead.

2.1. Model fitting

Our approach is to assume a non-informative prior for the model parameters and to fit the models using a Markov Chain Monte-Carlo (MCMC) algorithm based on the Gibbs sampler and Metropolis–Hastings algorithms. Metropolis–Hastings steps are used to sample from conditional posterior distributions that cannot be determined analytically. Unlike inferences based on Laplace approximation, analytical approximation is not required. Computations can

be easily performed using BUGS [23], while CODA [24] can be used to monitor convergence of the Markov chain.

3. MOTIVATING EXAMPLE

The proposed methodology is motivated by a prospective Radiologic Diagnostic Oncology Group (RDOG) prostate cancer biopsy trial that includes data from three clinical centres [25]. Radical prostatectomy was performed in all patients to provide a binary gold standard, which was based on pathology to classify patients into local (periprostic invasion of tumour and spread of disease to the seminal vesicles and lymph nodes) *versus* advanced (states A and B) prostate cancer.

Each individual's prostate specific antigen (PSA) level was taken. Another important measure is the Gleason score, which is used to classify patients into three risk groups according to the clinical criteria of D' Amico *et al.* [26]: low risk if Gleason score is less than 7, intermediate risk if Gleason score equals 7, and high risk if Gleason score is above 7. Additional covariates such as ethnicity were omitted because they did not have a strong effect on the outcome in this particular example.

Table I shows both the cluster-specific and overall summary statistics for the disease stage, PSA level, and Gleason score. Due to concerns over patient privacy, the three institutions (hospitals) are labelled A, B, and C.

The primary objective of the analysis is to determine the diagnostic efficiency of PSA-level (a continuous outcome variable), over and above the information provided by Gleason score, when the gold standard is the binary stage of disease and data are clustered by hospital. A secondary objective is to optimally combine the PSA-level and Gleason score into a single composite test.

In earlier work, we conducted a pooled over hospitals analysis by transforming all observations using a Box–Cox transformation with power coefficient of 0.33 [6]. However, due to different laboratory techniques, differences in operator experience, differences in the case-mix of patients, the rate of outliers and thus the shape of the outcome distribution may vary between hospitals. Therefore, the analysis might be improved by fitting a model that allows for variable transformations across clusters.

4. ROC FOR THE BMHTM

The BMHTM defined in equations (1) and (2) is used to derive the ROC curve and associated summary measures of accuracy for the test outcomes. The diagnostic accuracy of a test outcome over and above the information in the covariates $\mathbf{Z}_{ij}$ is of interest. In the prostate biopsy example based on equations (3)–(5), X_{1ij} is the PSA level ($K = 1$), T_{ij} is prostate cancer stage (advanced $T_{ij} = 1$ *versus* local $T_{ij} = 0$), and $\mathbf{Z}_{ij}$ is Gleason score. To simplify the presentation we omit subscripts for the remainder of this section, but emphasize that the results are outcome and cluster (hospital in our application) specific.

Under the assumption that a Box–Cox transformation is applied to the outcome, the ROC curve for individuals with $\mathbf{Z} = \mathbf{z}$ is defined by the co-ordinates

$$(1 - \text{Specificity}) : p = p(s; \mathbf{z}, \beta, \theta, \lambda) = \Pr(X^{(\lambda)} \geq s \mid T = 0, \mathbf{z}, \beta, \theta, \lambda)$$

and

$$\text{Sensitivity} : q = q(s; \mathbf{z}, \beta, \theta, \lambda) = \Pr(X^{(\lambda)} \geq s \mid T = 1, \mathbf{z}, \beta, \theta, \lambda)$$

for the positivity threshold $s \in (-\infty, \infty)$. Under the BMHTM model, solving for s and utilizing the symmetry for the normal density function yields

$$s = \mu(0, \mathbf{z}, \beta) + \sigma(0, \mathbf{z}, \theta) \Phi^{-1}(1 - p)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution. Therefore, the ROC curve is defined by the pairs of (1-specificity) and sensitivity, given by

$$\left[p, \Phi\{\eta + v\Phi^{-1}(p)\} \right]$$

where the ROC curve parameters $(\eta; v)$ are given by

$$\eta = \frac{\mu(1, \mathbf{z}, \beta) - \mu(0, \mathbf{z}, \beta)}{\sigma(1, \mathbf{z}, \theta)}$$

and

$$v = \frac{\sigma(0, \mathbf{z}, \theta)}{\sigma(1, \mathbf{z}, \theta)}$$

The fact that the ROC curve is independent of the transformation parameter λ , is a consequence of the general result that ROC curves are invariant to monotone transformations [16]. However, because estimates of η and v depend on the transformation, inferences about the ROC curve still depend on the transformation model.

4.1. Summary measures

Summary measures of diagnostic accuracy considered here include the area under the curve (AUC) [27] and the maximum improvement of sensitivity (MIS) over chance [6]. Both are functions of the ROC parameters (η, v) . The AUC of the λ -transformed outcomes for an individual having covariate $\mathbf{z}$, is given by

$$\text{AUC}(\mathbf{z}) = \Phi\left\{\frac{\eta}{(1 + v^2)^{1/2}}\right\}$$

The MIS is defined as

$$\text{MIS}(\mathbf{z}) = \Phi\left[\eta + v\Phi^{-1}\{1 - p(\mathbf{z})\}\right] - p(\mathbf{z})$$

where

$$p(\mathbf{z}) = \Phi\left[\frac{-\eta v + \{\eta^2 + 2(v^2 - 1)\log(v)\}^{1/2}}{v^2 - 1}\right]$$

for all $v \neq 1$. When $\eta \geq 1$ and $v = 1$, $\tilde{p}(\mathbf{z}) = \Phi(-\eta/2)$; and when $\eta \geq 1$ and $v = 1$, $p(\mathbf{z}) \in \{0, 1\}$.

Notice that $\tilde{p}(\mathbf{z})$ is the value of p at which the function $(1 - p) + \Phi\{\eta + v\Phi^{-1}(p)\}$, i.e. the sum of specificity and sensitivity, is maximized. Therefore, $\tilde{p}(\mathbf{z})$ is the optimal threshold on the probability scale when utility is defined by the L_1 -norm with true positives and negatives weighted equally, and false positives and negatives weighted equally.

4.2. Optimal threshold on the original scale

In practice, it is helpful to report the optimal threshold on the original rather than transformed measurement scale of a test. Let p^{opt} denote the optimal value of (1-specificity) for a test under some criterion (e.g. MIS). The optimal threshold on the transformed scale is then

$$s^{\text{opt}} = \mu(0, \mathbf{z}, \beta) + \sigma(0, \mathbf{z}, \theta) \Phi^{-1}(1 - p^{\text{opt}})$$

The problem is to transform s^{opt} back to the original scale of the data. This generally relies on the inverse-transformation function, defined for the Box-Cox transformation $h(X; \lambda)$ in (5) as

$$x^{\text{opt}}(\lambda) = h^{-1}(s^{\text{opt}}, \lambda) = \begin{cases} (1 + \lambda s^{\text{opt}})^{1/\lambda} & \text{if } s^{\text{opt}} > -1/\lambda \text{ and } \lambda \neq 0 \\ \exp(s^{\text{opt}}) & \text{if } \lambda = 0 \end{cases} \quad (7)$$

When $s^{\text{opt}} \geq -1/\lambda$ the expression in (7) is undefined, which may occur with an unusual set of parameter values. In this case the minimum (maximum) values of X if λ is positive (negative) could be taken to be the associated variates of the posterior distribution of $x^{\text{opt}}(\lambda)$. If

$s^{\text{opt}} \geq -1/\lambda$ occurs frequently we recommend that the optimal threshold value on the original scale be estimated by evaluating $h^{-1}(\cdot)$ at the posterior mean of $(s^{\text{opt}}, \lambda)$. Alternatively, the prior distribution could be subject to the constraint $s^{\text{opt}} > -1/\lambda$, enabling all of the information in the data to be used for model fitting.

5. OPTIMAL COMBINATION OF TEST SCORES

We now consider the problem of optimally combining multiple correlated outcomes into a single composite test when the outcomes need to be transformed, depend on additional covariates, and form clusters of correlated groups. In our motivating example, PSA and Gleason score are to be combined into a univariate measure in order to predict or classify patients according to prostate cancer stage.

The standard approaches to forming composite tests include discriminant function analysis in which the test outcomes are the dependent variables [28], or using a generalized linear model (e.g. logistic regression) to relate the test outcomes and any other covariates to the gold standard disease status [29]. In this section, we compare the methodologies.

5.1. Discriminant function approach

The objective is to determine whether an arbitrary individual belongs to the diseased or non-diseased population, indicated by $T_{ij} = 1$ and $T_{ij} = 0$, respectively. The prevalence of diseased and non-diseased subjects is given by $\pi_i = \Pr(T_{ij} = 1)$. To simplify the presentation of the following methods, the cluster indicator variables are absorbed in $\mathbf{Z}_{ij}$, and we omit the subscripts for cluster and individual from the remainder of this section.

The decision rule which minimizes the expected probability of misclassification classifies a patient with $\mathbf{X} = \mathbf{x}$ and covariates $\mathbf{Z} = \mathbf{z}$ as diseased if and only if

$$\pi f_1(\mathbf{x} | \mathbf{z}) \geq (1 - \pi) f_0(\mathbf{x} | \mathbf{z}) \quad (8)$$

where $f_t(\mathbf{X} | \mathbf{z})$ is the probability density function of $\mathbf{X}$ given $(T = t; \mathbf{Z} = \mathbf{z})$ [21]. The expression in (8) is Bayes' rule under a 'zero-one' loss function [30].

Let $\mu_{kt}(\mathbf{z}) = \mu(t, \mathbf{z}, \beta_k)$, $\sigma_{kt}(\mathbf{z}) = \sigma(t, \mathbf{z}; \theta_k)$, $\mu_t(\mathbf{z}) = \{\mu_{1t}(\mathbf{z}), \dots, \mu_{Kt}(\mathbf{z})\}$, and $\Sigma_t(\mathbf{z}) = D_t \Psi D_t$, where $D_t = \text{diag}\{\sigma_{1t}(\mathbf{z}), \dots, \sigma_{Kt}(\mathbf{z})\}$. Under the BMHTM defined in (1) and (2), $f_t(\mathbf{X} | \mathbf{z}) = N\{\mu_t(\mathbf{z}), \Sigma_t(\mathbf{z})\}$, and it follows that

$$\Pr(T = 1 | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \frac{\exp\{a(\mathbf{z}) + \delta(\mathbf{z})' \mathbf{x} + \mathbf{x}' \Delta(\mathbf{z}) \mathbf{x}\}}{1 + \exp\{a(\mathbf{z}) + \delta(\mathbf{z})' \mathbf{x} + \mathbf{x}' \Delta(\mathbf{z}) \mathbf{x}\}} \quad (9)$$

where

$$\begin{aligned} a(\mathbf{z}) &= \log\left(\frac{\pi}{1 - \pi}\right) - \frac{1}{2} \log\left\{\frac{\det \Sigma_1(\mathbf{z})}{\det \Sigma_0(\mathbf{z})}\right\} - \left\{\mu_1(\mathbf{z})' \Sigma_1^{-1}(\mathbf{z}) \mu_1(\mathbf{z}) - \mu_0(\mathbf{z})' \Sigma_0^{-1}(\mathbf{z}) \mu_0(\mathbf{z})\right\} / 2 \\ \delta(\mathbf{z}) &= \Sigma_1^{-1}(\mathbf{z}) \mu_1(\mathbf{z}) - \Sigma_0^{-1}(\mathbf{z}) \mu_0(\mathbf{z}) \\ \Delta(\mathbf{z}) &= -\left\{\Sigma_1^{-1}(\mathbf{z}) - \Sigma_0^{-1}(\mathbf{z})\right\} / 2 \end{aligned}$$

At $\mathbf{Z} = \mathbf{z}$, a patient with $\mathbf{X} = \mathbf{x}$ is classified as diseased if and only if $\delta(\mathbf{z})' \mathbf{x} + \mathbf{x}' \Delta(\mathbf{z}) \mathbf{x} \geq s$, for some threshold s , and non-diseased, otherwise. An often-times unsettling feature of the quadratic classification rule is that when $\delta(\mathbf{z}) \approx 0$ it assigns patients with tests scores $\mathbf{X} = \mathbf{x}$ and $\mathbf{X} = -\mathbf{x}$ to the same population. However, for tests with good diagnostic properties, such anomalies are unlikely as $\delta(\mathbf{z})$ tends to dominate the classification rule.

When $\Sigma_1(\mathbf{z}) = \Sigma_0(\mathbf{z})$, $\Delta(\mathbf{z}) = 0$, and so the linear discriminant function $\delta(\mathbf{z})' \mathbf{X}$ is obtained. Su and Liu [28] showed that $\delta(\mathbf{z})' \mathbf{X}$ is the optimal linear combination of test outcomes under homogeneous covariance matrices.

5.2. Logistic regression approach

When $\mathbf{X}$ given $(T, \mathbf{Z})$ is multivariate normal, the posterior probability in (9) has the form of a logistic regression probability. That is, logistic regression analysis of the binary variable T on $\mathbf{X}$ (including relevant quadratic terms from $\mathbf{X}\mathbf{X}'$) and $\mathbf{Z}$ is a different way of estimating the discriminant function.

Because the likelihood function is different, inferences determined using logistic regression are not equivalent to those from the normal discriminant model. It is anticipated that the normal discriminant model will yield estimates that are more precise when the transformed data are normally distributed, while the logistic regression model will be more robust to the assumption of normality. Kao and McCabe [21] noted that the binormal and logistic regression estimates of $\delta(\mathbf{z})$ are asymptotically equivalent when the outcomes are normally distributed and the covariance matrix of the outcomes is constant. However, when a nonlinear transformation is applied to the outcomes, the normal discriminant function and logistic regression approaches will yield expressions for discriminating between individuals' disease states that have different functional forms.

6. RESULTS FOR THE MOTIVATING EXAMPLE

We now present results for the two problems introduced in Section 1. Because of the limited data available, the number of terms needed to account for the clustering by hospital and to estimate the transformation, and a desire to keep this illustrative example as simple as possible, no interaction effects between the Gleason score covariate and disease status are included in the present analyses. We refer readers to our previous pooled analysis [6] for an analysis of the Gleason score by disease status interaction, ignoring clustering by hospital.

6.1. Problem 1: hierarchical transformation model (BMHTM)

In our example, X_{1ij} is the PSA level, T_{ij} is cancer stage (advanced *versus* local), and Z_{ij} is the Gleason score. We wish to measure the diagnostic accuracy of PSA over and above Gleason score. We assume that the parameters of the intercepts and slope of the gold standard in (3) are random effects, but treat the Gleason score coefficients, and the variance and transformation parameters, as fixed effects. The model is expressed as

$$\text{Level I: } X_{1ij}^{(\lambda_{1i})} \sim N[\beta_{1i0} + \beta_{1i1} T_{ij} + \beta_{1i2} Z_{ij}, (\theta_{1i0} + \theta_{1i1} T_{ij})^2] \quad (10)$$

where $X_{1ij}^{(\lambda_{1i})}$ is defined as in (5), and

$$\text{Level II: } \beta_{1i} \sim N(\beta_{10}, \Sigma_1) \quad (11)$$

where Σ_1 is a diagonal matrix with l th element denoted ς_l^2 (the between cluster variance of the l th regression effect) for $l = 0, 1, 2$. The ROC curve parameters for the model defined in (10)

and (11) are $\eta_{1ij} = \beta_{1i1}/|\theta_{1i0} + \theta_{1i1}|$ and $v_{1ij} = |\theta_{1i0}|/|\theta_{1i0} + \theta_{1i1}|$, where $|\cdot|$ is the absolute value function.

Because an interaction effect between Gleason score and the gold standard is not included in the model in (10) and (11) the ROC curves and associated summary statistics are identical for all Gleason scores. If (3) contained an interaction effect between Gleason score and disease status this would not be the case (results under this scenario but for a model that does not account for clustering by hospital are provided in O'Malley *et al.* [6]).

To conduct Bayesian analysis, non-informative priors are assumed for θ_{1i} , λ_{1i} , β_{10} , and ζ_I^2 . For MCMC computations in BUGS the priors must be proper probability distributions.

Accordingly, we specify $\sigma_{1i}^2 = (\theta_{1i0} + \theta_{1i1})^2 \sim \text{IG}(10^{-3}, 10^{-3})$, $\theta_{1i0}^2 \sim \text{IG}(10^{-3}, 10^{-3})$, $\lambda_{1i} \sim N(1, 10)$, $\beta_{10} \sim N(0, 10^6)$, and $\zeta_I^2 \sim \text{IG}(10^{-3}, 10^{-3})$, where IG denotes an inverse-gamma density. We found that changing the variance parameter of the normal prior, or the parameters of the inverse-gamma distribution, by a factor of 10 had miniscule impact on the results; hence, the specified priors may be considered to be non-informative. To maintain identifiability of θ_{1i} we apply the constraints $\theta_{1i0} \geq 0$ and $\theta_{1i1} \geq -\theta_{1i0}$.

6.1.1. Preliminary analysis: missing data.—Of the total of 213 observations there are 32 observations with missing values for PSA or Gleason score. All but one of the observations with missing Gleason score also had missing stage of disease, whereas no observations with missing PSA had missing stage of disease. To retain consistency with previous analyses of these data [6,25], our main results are based on the 181 observations with complete data. However, we checked the results from our complete case analysis against those from a multiple imputation analysis [31] of all 213 observations. The results were very similar to those for the complete data analysis, especially for hospitals B and C which comprise the greatest number of patients. For example, the posterior means (standard deviations) of the AUC of the ROC curve under the heterogeneous transformation model for hospitals A, B, C were 0.571 (0.106), 0.703 (0.055), 0.760 (0.061) and 0.626 (0.065), 0.704 (0.050), 0.746 (0.058) for the multiple imputation and complete case analyses, respectively. Therefore, omitting observations with missing values has a non-trivial impact on the results for hospital A only. The robustness of the results for other hospitals is not surprising given the small amount of data missing.

6.1.2. Transformations to normality.—Because PSA is clustered and skewed even after standardization to a common mean and variance within each hospital and disease status (see upper plot in Figure 1), it must be transformed before normality can be assumed. To investigate the effect of assumptions concerning the Box–Cox transformation parameters λ_{1i} on the analysis, we consider three computations: fixing $\lambda_{1i} = 0.33$ (as in Reference [6]), setting $\lambda_{1i} = \lambda_{10}$ (homogeneous), and leaving λ_{1i} unrestricted for $i = 1, 2, 3$. The transformed PSA density that results from the unrestricted model is displayed in the lower part of Figure 1. The assumption of normality is clearly much more reasonable after transformation.

The posterior distributions of the transformation parameters may be used to gauge the level of heterogeneity in the shape of the outcome distributions for different hospitals. Small differences between hospitals are apparent (Figure 2). The posterior means under the unrestricted Box–Cox transformation are 0.388, 0.404, and 0.304 for hospitals A, B, and C, respectively, in contrast to the mean of 0.372 under the homogeneous transformation model and the overall MLE of 0.33 used previously [6]. The posterior distribution of λ_{1A} is more dispersed than λ_{1B} and λ_{1C} (posterior standard deviation of 0.121 compared to 0.078 and 0.092 for hospitals B and C, respectively). This is not surprising as hospital A had the smallest number of patients.

To check the validity of the assumption that the transformed PSA is normal given hospital, disease state, and gleason score we evaluated Bayesian predictive p -values of the skewness (third moment) and kurtosis (fourth moment) coefficients [32] for the residuals of the transformed PSA values. Because both coefficients are 0 for normal densities, and the variance of PSA varies across hospitals and disease groups, we measure departures from normality using the following weighted averaged distance measures:

$$S_1(\mathbf{X}_1, \boldsymbol{\beta}_1) = \sum_{i,t} w_{it} |SK(\mathbf{X}_{1i}, \boldsymbol{\beta}_{1i})|$$

and

$$S_2(\mathbf{X}_1, \boldsymbol{\beta}_1) = \sum_{i,t} w_{it} |KUR(\mathbf{X}_{1i}, \boldsymbol{\beta}_{1i})|$$

where $\mathbf{X}_1$ is a vector of all the transformed PSA values, w_{it} is the proportion of subjects from hospital i and with disease state $T_{ij} = t$, and $SK(\mathbf{X}_{1i}, \boldsymbol{\beta}_{1i})$ and $KUR(\mathbf{X}_{1i}, \boldsymbol{\beta}_{1i})$ are the skewness (SK) and kurtosis (KUR) coefficients for the distribution of the residuals

$X_{1ij}^{(\lambda_{1i})} - \beta_{1i0} - \beta_{1i1}T_{ij} - \beta_{1i2}Z_{ij}$. Bayesian predictive p -values are obtained by simulating replicate values, $\mathbf{X}_1^{\text{rep}}$, of the transformed outcomes from the model and evaluating the proportion of times that $S_m(\mathbf{X}_1^{\text{rep}}, \boldsymbol{\beta}_1) > S_m(\mathbf{X}_1, \boldsymbol{\beta}_1)$ for $m = 1, 2$. The larger the p -values the more confident we can be that the data are normal whereas small p -values (e.g. less than 0.05) cast doubt on whether the data are normal.

The Bayesian predictive p -values for (S_1, S_2) were (0.480, 0.926) and (0.484, 0.941) under the homogeneous and hospital-specific transformation models, respectively. Therefore, we conclude that the assumption of normality is reasonable.

Figure 3 displays the posterior mean ROC curves for each hospital when λ_{1i} is unrestricted. The ROC curve indicates that hospital A performs the worst, while hospital C outperforms hospital B whenever specificity is greater than 0.38 (approximately where the ROC curves cross) but performs worse at low specificity.

A detailed comparison of the ROC curves in Figure 3 with those for the homogeneous ($\lambda_{1i} = \lambda_{10}$) case revealed that the largest differences occurred when specificity is close to 0.5 for hospital A and close to 0 or 1 for hospitals B and C (see Figure 4). However, the differences are fairly small (a maximum difference in sensitivity of 0.03) as the population ROC curve for the BMHTM is invariant to monotonic transformations of the outcome distributions, and it is only through the estimation of the model parameters that the transformation model has an influence. The lack of conformity at the extremes in hospitals B and C is important when one evaluates the sensitivity of a test at a high value of specificity.

6.1.3. Summary measures of diagnostic accuracy.—Comparisons of AUC and MIS across hospitals in Table II reveal that hospital C performs the best while hospital A performs the worst. The ROC curve summary statistics in Table II reveals little difference in the overall measures of model accuracy between the three transformation models.

6.1.4. Optimal threshold analysis.—The values of $x^{\text{opt}}(\lambda_{1i})$ vary much more than the summary measures of diagnostic accuracy. This is particularly true for hospital B. The posterior standard deviations are the smallest when $\lambda_{1i} = 0.33$ is assumed, and largest for hospitals A and C when λ_{1i} is unrestricted. The posterior distributions of the optimal thresholds on the original scale of the data are presented in Figure 5. A few values (0.18 per cent) generated from the posterior distribution of S_{1A}^{opt} (the optimal threshold for hospital A) lay outside the domain

of the inverse Box–Cox transformation, and hence could not be inverted. Non-invertible values were assigned 0 according to the discussion following equation (7). Hospital A's optimal threshold is the least precisely estimated of the hospitals (its standard deviation is larger by a factor of approximately 3) because it has the smallest number of patients.

6.2. Problem 2: optimal combinations on PSA data

For this application, X_{1ij} denotes PSA and X_{2ij} denotes Gleason score for the j th patient in hospital i . The actual Gleason scores are used as normality is more tenable than for the tri-chotomization of risk levels used when Gleason score functioned as a covariate in problem 1. To compare the discriminant function and logistic regression approaches in the simplest scenario we set $\lambda_{1i} = 0.33$ and $\lambda_{2i} = 0$, ignore the clustering of observations by hospitals, and assume that the outcomes have the same covariance for each disease group. This defines the following models:

$$\text{BMHTM} : \begin{pmatrix} X_{1ij}^{0.33} \\ \log(X_{2ij}) \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \beta_{1i0} + \beta_{1i1} T_{ij} \\ \beta_{2i0} + \beta_{2i1} T_{ij} \end{pmatrix}, \Sigma \right]$$

and

$$\text{Logistic} : T_{ij} \sim \text{Bernoulli} \left[\text{logit}^{-1} \{ a_{1i} + \gamma_1 X_{1ij}^{0.33} + \gamma_2 \log(X_{2ij}) \} \right]$$

where $\beta_{kil} \sim N(\beta_{k0l}, \varsigma_{k0l}^2)$ and $a_{1i} \sim N(a_0, \omega_0^2)$ ($k \in \{1, 2\}; l \in \{0, 1\}$). Non-informative prior $N(0, 10^6)$ densities for the mean function parameters (β_{k0l} , a_0 , and γ_1 for $k \in \{1, 2\}; l \in \{0, 1\}$), $IG(10^{-3}, 10^{-3})$ densities for ς_{k0l}^2 and ω_0^2 , and an inverse-Wishart prior Σ , complete Bayesian specifications of both models. The inverse-Wishart density was assigned two degrees of freedom, to make the prior specification non-informative, and its scale parameter was set equal to the sample covariance matrix across all clusters for PSA and Gleason score.

Note that the Box–Cox transformation parameters are pre-specified in order to maximize the correspondence between the normal discriminant function and logistic regression approaches. In general, λ_{ki} should be included in the modelling process for both analyses, and one should account for the clustering of observations by hospital. If additional covariates are available these may simply be added to the mean and variance functions.

Because π_i has no bearing on the form of the classification rule (it only affects the constant term in equation (9)), in the discriminant function approach we set π_i equal to the proportion of cases of advanced cancer in each cluster. For hospitals A, B, and C, the assumed proportion of disease cases are $\pi_A = 0.56$, $\pi_B = 0.55$, and $\pi_C = 0.79$ (Table I). If instead we incorporated prior information about π_i then because the normal discriminant function and the logistic regression models are different, the prior would affect the optimal classification differently, and thus make the two approaches less comparable.

The posterior means and standard deviations of the parameters in the BMHTM and logistic regression models are provided in Table III, and the associated posterior probabilities of disease as a function of PSA are displayed in Figure 6. It is clear from Table III that PSA and Gleason score both make significant contributions to the model, implying that a combined test will out-perform both of the individual tests. The regression coefficients in the logistic regression model are larger for both PSA and Gleason score than for the normal discriminant function model, suggesting that it is a more sensitive classification rule (especially with respect to Gleason score in Hospitals A and C). This claim is supported by Figure 7 where the contours of constant posterior probability for the logistic regression model are closer together than for the linear discriminant function.

In Figure 6 the posterior probabilities of advanced disease are shown as a function of PSA with Gleason score fixed at its mean, and as a function of Gleason score with PSA fixed at its mean, for each hospital. The posterior probability plot based on logistic regression is slightly more sensitive to PSA and Gleason score than that for the normal discriminant model. The substantial difference between the plotted curves when Gleason score is less than 4 for hospital C reflects the fact that predictions over this range are based on limited data.

Figure 7 reveals that for hospitals A and C higher values of PSA and Gleason score are needed under the logistic regression model than under the normal discriminant function model to attain probabilities of disease in the range [0.1, 0.8]. The opposite is true for hospital B. Note that the non-linearity in the contours arises because they are plotted in probability space and PSA and Gleason score undergo transformations in both models.

7. SUMMARY AND DISCUSSION

To address both problems introduced in Section 1 we have investigated the following issues: multivariate multi-level modelling in the context of ROC analysis, application of nonlinear monotone cluster-specific transformations, optimal threshold analysis and exact inferences for summary ROC measures, and estimation of the optimal combination of test scores. Each of these analyses was performed within the BMHTM framework. This framework offers a wide-range of validative and predictive applications to the analysis of diagnostic data.

Bayesian modelling, coupled with MCMC for model fitting, avoids the need for analytical approximations, and thus in principle allows for exact evaluation of inferences for ROC analysis. Inferences about summary measures of an ROC curve, along with the estimation of optimal combinations of test outcomes, are straight-forward to evaluate even with complex hierarchically clustered data. In addition, information external to an experiment may be incorporated via exible prior distributions. Although nearly non-informative prior distributions were assumed for our Bayesian analyses, informative prior distributions can be applied analogously (see O'Malley *et al.* [6]).

In addition to variance heterogeneity and non-linear regression equations, the BMHTM allows data to undergo different cluster-specific transformations. The insensitivity of the PSA data analysis to cluster-specific transformation parameters might indicate that there is little variation between hospitals in the procedures used or in the overall skill of the operators. If the operators were used as clusters (these were not available to us) we could have expected more heterogeneity in the transformation parameters. Other applications where there might be extensive between operator heterogeneity in the transformation parameter include a radiologist's score of a mammogram, the estimate of bone density from a bone densitometer to diagnose osteoporosis, and self-administered tests (e.g. pregnancy test, prothombin time or blood clot test, and glucose meter).

An important limitation of the methodology used to transform optimal thresholds back to the original scale of the data is the use of *ad hoc* values (e.g. those lying on the boundary of the parameter space) when the inverse transformation is undefined. Although, this had little impact on our numerical results, one would do better (assuming the specified model is correct) by constraining the parameters so that invertability was assured across the range of the data.

An important finding from this research is the non-trivial difference between the normal discriminant function and logistic regression models for combining test scores. Although these models may be equivalent asymptotically, they can lead to different conclusions on the same data set. Therefore, our work suggests that a formal study to compare the estimated optimal combination of test outcomes between the normal discriminant function and logistic regression models under a range of different circumstances is needed. In the prostate biopsy example we

only considered linear combinations of test outcomes. Non-linear combinations could also be considered as a future extension.

Another topic worth future investigation concerns Bayesian model averaging [33–35]. Rather than using standard model selection techniques to choose a single model for inference, we could use Bayesian model averaging to account for model uncertainty. Two ways of combining the BMHTM based normal discriminant function and the logistic regression models into a single test have been outlined by O'Malley and Zou [36].

ACKNOWLEDGEMENTS

This work was supported by the U.S.A. National Institutes of Health Grants R01LM7861, R03HS13234, PO1CA41167 and UO1CA45256. We acknowledge Drs Barbara J. McNeil and Clare M. C. Tempany for providing the RDOG data, and Daryl Caudry for data management. We also thank two anonymous referees for many useful comments on previous versions of the paper.

REFERENCES

1. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* 1994;13:499–508. [PubMed: 8023031]
2. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 1998;17:1033–1053. [PubMed: 9612889]
3. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* 2000;56:352–359. [PubMed: 10877289]
4. Cai T, Pepe MS. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* 2002;97:1099–1107.
5. Zhou, XH.; McClish, DK.; Obuchowski, NA. *Statistical Methods in Diagnostic Medicine*. Wiley; New York: 2002.
6. O'Malley AJ, Zou KH, Fielding JR, Tempany CMC. Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: prostate biopsy and spiral CT of ureteral stones. *Academic Radiology* 8:713–725. [PubMed: 11508750]
7. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *The Canadian Journal of Statistics* 2000;28:731–750.
8. Gatsonis CA. Random-effects models for diagnostic accuracy data. *Academic Radiology* 1995;(2nd Suppl):S12–S21.
9. Hellmich M, Abrams KR, Jones DR, Lambert PC. A Bayesian approach to a general regression model for ROC curves. *Medical Decision Making* 1998;18:436–443. [PubMed: 10372587]
10. Slate EH, Turnbull BW. Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine* 2000;19:617–637. [PubMed: 10694740]
11. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000;56:337–344. [PubMed: 10877287]
12. Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making* 1999;19:242–251. [PubMed: 10424831]
13. Tu W, Zhou X-H. A Wald test comparing medical costs based on log-normal distributions with zero valued costs. *Statistics in Medicine* 1999;18:2749–2761. [PubMed: 10521864]
14. Oberg A, Davidian M. Estimating data transformations in nonlinear mixed effects models. *Biometrics* 2000;56:65–72. [PubMed: 10783778]
15. Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Statistics in Medicine* 2002;21:3093–3106. [PubMed: 12369084]
16. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* 1996;24:25–40.
17. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* 2000;27:621–631.

18. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989;76:585–592.
19. 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. [PubMed: 9330425]
20. Cole TJ. The LMS method for constructing normalized growth standards. *European Journal of Clinical Nutrition* 1990;44:45–60. [PubMed: 2354692]
21. Kao T, McCabe GP. Optimal sample allocation for normal discrimination and logistic regression under stratified sampling. *Journal of the American Statistical Association* 1991;86:432–436.
22. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 1964;26:211–246.
23. Spiegelhalter, DJ.; Thomas, A.; Best, NG. WinBUGS Version 1.2. MRC Biostatistics Unit; Cambridge, U.K.: 1999.
24. Best, N.; Cowles, MK.; Vines, K. Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output. MRC Biostatistics Unit; Cambridge, U.K.: 1995.
25. Tempany CMC, Zhou X-H, Zerhouni EA, Rifkin MD, Quint LE, Piccoli CW, Ellis JH, McNeil BJ. Staging of prostate cancer: results of radiology diagnostic oncology group project comparison of three MR imaging techniques. *Radiology* 1994;192:47–54. [PubMed: 8208963]
26. D'Amico AV, Whittington R, Malkowicz SB, Schultz D, Blank K, Broderick GA, Tomaszewski JE, Renshaw AA, Kaplan I, Beard CJ, Wein A. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *Journal of the American Medical Association* 1998;280:969–974. [PubMed: 9749478]
27. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36. [PubMed: 7063747]
28. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* 1993;88:1350–1355.
29. Marshall RJ. The predictive value of simple rules for combining two diagnostic tests. *Biometrics* 1989;45:1213–1222.
30. Berger, JO. *Statistical Decision Theory and Bayesian Analysis*. 2nd. Springer; New York: 1985.
31. Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley; Chichester, New York: 1987.
32. Mood, AM.; Graybill, FA.; Boes, DC. *Introduction to the Theory of Statistics*. 3rd. McGraw-Hill; Singapore: 1974. p. 76
33. Draper D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B* 1995;57:45–97.
34. Raftery, AE. Bayesian model selection in social research (with discussion). In: Marsden, PV., editor. *Sociological Methodology*. Blackwells Publishers; Cambridge, MA: 1995. p. 111-195.
35. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical Science* 1999;14:382–401.
36. O'Malley AJ, Zou KH. Bayesian hierarchical transformation models for receiver operating characteristic curve analysis. *ASA Proceedings of the Health Policy Statistics Section* 2002;2541–2550.

Figure 1. Non-parametric estimate of probability distribution functions of PSA and transformed PSA for patients with advanced and local prostate cancer. Both raw and transformed PSA are standardized to have mean 0 and variance 1 within each hospital prior to density estimation.

Figure 2. Posterior distribution of transformation parameter under the homogeneous model (overall) and the hospital specific transformations under the unrestricted model. The corresponding posterior expectations are: $E[\lambda_{\text{Overall}} \mid \text{data}] = 0.372(0.082)$, $E[\lambda_1 \mid \text{data}] = 0.388(0.121)$, $E[\lambda_2 \mid \text{data}] = 0.404(0.078)$, $E[\lambda_3 \mid \text{data}] = 0.304(0.092)$.

Figure 3.
ROC curves for the 3 clusters (hospitals) using cluster-specific transformations.

Figure 4.
Difference of the ROC curves based on the cluster-specific and homogeneous transformation models for the 3 clusters (hospitals).

Figure 5. Posterior distribution of the optimal thresholds on the original PSA measurement scale, $x^{\text{opt}}(\lambda_{1i})$ for $i = A, B, C$. The corresponding posterior expectations are: $E[x^{\text{opt}}(\lambda_{1A})|\text{data}] = 11.5$ (4.92), $E[x^{\text{opt}}(\lambda_{1B})|\text{data}] = 8.74$ (1.43), $E[x^{\text{opt}}(\lambda_{1C})|\text{data}] = 11.1$ (1.92).

Figure 6. Posterior probabilities of advanced disease as a function of PSA and Gleason score based on the linear discriminant function and linear logistic regression models.

Figure 7.
Contours in probability space of the classification rule for the linear discriminant function and linear logistic regression models.

Table I

Summary statistics of RDOG prostate trial data.

Hospital	Advanced disease	PSA score			Gleason score		
	Mean (Proportion)	Mean	SD	Range	Mean	SD	Range
A	0.56	9.49	9.37	[0.4, 44]	6.36	1.08	[5, 9]
B	0.55	9.43	7.94	[0.1, 47]	6.66	1.25	[3, 9]
C	0.79	13.1	11.0	[1.3, 58]	6.21	1.67	[2, 9]
Overall	0.63	10.7	9.39	[0.1, 58]	6.46	1.40	[2, 9]

Table II

Posterior means (standard deviations) of characteristics of the ROC Curve for different transformation models.

Model	Hospital	Summary measures of diagnostic accuracy			
		$AUC_1(z_{ij})$	$\tilde{p}_1(z_{ij})$	$MIS_1(z_{ij})$	$x^{opt}(\lambda_{1i})$
Fixed transformation; $\lambda_{1i} = 0.33$	A	0.63 (0.06)	0.27 (0.15)	0.24 (0.09)	11.2 (4.22)
	B	0.70 (0.05)	0.34 (0.09)	0.30 (0.08)	8.25 (1.38)
	C	0.75 (0.06)	0.16 (0.06)	0.42 (0.09)	10.7 (1.58)
Homogeneous transformation; $\lambda_{1i} = \lambda_{10}$	A	0.63 (0.06)	0.25 (0.15)	0.24 (0.09)	11.5 (4.15)
	B	0.70 (0.05)	0.31 (0.10)	0.30 (0.08)	8.59 (1.54)
	C	0.74 (0.06)	0.15 (0.06)	0.42 (0.09)	10.8 (1.58)
Unrestricted transformation; λ_{1i} free	A	0.63 (0.07)	0.25 (0.16)	0.23 (0.09)	11.5 (4.92)
	B	0.70 (0.05)	0.29 (0.09)	0.31 (0.07)	8.74 (1.43)
	C	0.75 (0.06)	0.17 (0.07)	0.42 (0.10)	11.1 (1.92)

AUC = Area under the ROC curve; $\tilde{p}$ = 1-specificity at which MIS (the maximum improvement of sensitivity over chance) occurs; x^{opt} = optimal threshold on the original scale of the outcome variable.

Table III

Posterior mean (standard deviation) of the discrimination function parameters for the normal linear discriminant function and logistic regression models.

Model	Hospital	Discriminant function parameters		
		Constant	PSA score	Gleason score
Normal discriminant function	A	-4.21 (2.19)	0.36 (0.17)	1.89 (1.26)
	B	-6.20 (1.72)	0.40 (0.13)	2.82 (0.95)
	C	-1.48 (1.91)	0.54 (0.14)	0.69 (1.16)
Logistic regression	A	-6.34 (1.76)	0.36 (0.26)	3.03 (0.95)
	B	-7.38 (2.00)	0.67 (0.21)	3.07 (1.07)
	C	-6.21 (1.86)	0.98 (0.29)	2.82 (0.95)