Published in final edited form as:

Biom J. 2005 August ; 47(4): 417–427.

A Bayesian Hierarchical Non-Linear Regression Model in Receiver Operating Characteristic Analysis of Clustered Continuous Diagnostic Data

Kelly H. Zou^{1,2,*} and A. James O'Malley¹

¹ Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA

² Department of Radiology, Brigham & Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA, 02115, USA

Abstract

Receiver operating characteristic (ROC) analysis is a useful evaluative method of diagnostic accuracy. A Bayesian hierarchical nonlinear regression model for ROC analysis was developed. A validation analysis of diagnostic accuracy was conducted using prospective multi-center clinical trial prostate cancer biopsy data collected from three participating centers. The gold standard was based on radical prostatectomy to determine local and advanced disease. To evaluate the diagnostic performance of PSA level at fixed levels of Gleason score, a normality transformation was applied to the outcome data. A hierarchical regression analysis incorporating the effects of cluster (clinical center) and cancer risk (low, intermediate, and high) was performed, and the area under the ROC curve (AUC) was estimated.

Keywords

Sensitivity; Specificity; Receiver operating characteristic analysis; Bayesian hierarchical models

1 Introduction

In 2004 approximately 230110 new cases of prostate cancer were diagnosed in the United States. Prostate cancer ranked the highest among all male malignancy cancers and was the second leading cause of deaths by cancer in American men, with an estimated 29900 deaths. One of the most useful screening tools for prostate cancer is the prostate specific antigen (PSA) test. See American Cancer Society (2004)'s cancer facts and figures (<http://www.cancer.org>). The importance of the prostate cancer markers for radiation therapy has been demonstrated clinically by Pound et al. (1999), Polascik et al. (1999), Stamey et al. (1999), Walsh (1999), D'Amigo et al. (2000), Feneley and Partin (2000), Lehrer et al. (2002), and Lieberfarb et al. (2002). In a prospective multi-center Radiological Diagnostic Oncology Group (RDOG) study reported by Tempany et al. (1994) and Seltzer et al. (1997) on prostate cancer staging sponsored by the National Cancer Institutes of Health, baseline variables were collected, including biopsy results (e.g., PSA level and Gleason score) and subject characteristics (e.g., age, sex, and race).

There are several challenges to analyze diagnostic and radiologic data from such a large and complex multi-center clinical trial: (1) The effect of cluster: Subjects enrolled from different clusters (e.g., clinical centers) may exhibit different characteristics. Therefore, observations

* Corresponding author: e-mail: zou@bwh.harvard.edu.

for subjects in the same cluster may be more similar (correlated) compared to observations from different clusters. A pooled analysis that does not consider the correlation of observations within clusters may be too anti-conservative, while a stratified analysis that does not allow information to be shared between clusters may be too conservative. (2) The effect of covariate: A large number of variables are typically available in a multi-center clinical trial, often necessitating multivariable regression analysis such as in Goldstein (1995). (3) The availability of prior information: Existing data from pilot study or prior knowledge may be utilized and incorporated in the current study, particularly in more costly imaging studies. (4) The need for a data transformation: The receiver operating characteristic analysis is a useful method for evaluating diagnostic data. When dealing with continuous diagnostic outcome data such as PSA, a bi-normal semi-parametric model based on the joint ranks of the diagnostic data was recommended by Metz et al. (1998). This model assumed that the transformed outcome data have two normal distributions with different means and variances. As recommended by Zou and Hall (2000; 2002), it is useful to take a normality transformation before applying the bi-normal model.

We address all of the above issues in analyzing multi-center diagnostic clinical trial data by extending the Bayesian regression approaches described in Spiegelhalter (1994) and Gelman et al. (1995). Bayesian analysis differs from the more conventional frequentist methodology in that model parameters are treated as random variables rather than fixed (unknown) constants. Therefore, Bayesian inferences are expressed as conditional statements about the parameters given the observed data, as opposed to confidence statements or significance-levels about the likelihood of having observed the data if particular values or hypotheses about the parameters were true. Because the model parameters are random variables, external information (e.g. elicited expert opinions, relevant historical information) can be incorporated into the model by constructing a probability distribution that describes the uncertainty in the model parameters (prior to the observing data from the experiment). This distribution is commonly referred to as a prior distribution. The general Bayesian computation centers around the use of Bayes' theorem to update the prior distribution with the information contained in the likelihood function of the experimental data to form a posterior (after data) distribution of the model parameters, or functions thereof, from which inferences are evaluated.

Bayesian methods have several advantages: (1) They incorporate complex designs. (2) The parameters in the model or functions of them, such as summary ROC measures, may be simulated directly via methods for the exploration of posterior distributions. (3) External information may easily be incorporated into the models. In radiological studies, prior information about diagnostic tests can often be extracted from pilot studies, meta-analysis of relevant literature, or scientific theory. When prior information and experimental data cohere, inferences based on Bayesian posterior distributions are often more precise than those using non-Bayesian methods. In many situations, a prior distribution that contains virtually no information about the value of the model parameters is used for a Bayesian analysis. Such prior distributions are referred to as non-informative or diffuse priors.

Continuous diagnostic data derived from multi-center clinical trials have been used frequently due to the advancement in laboratory and bioinformatics tools. Thus, the methodology proposed here may address an important question concerning diagnostic testing evaluation. Multivariate methods for non-Bayesian ROC analysis have been developed, including random-effects models by Gatsonis (1995), Tosteson and Begg (1988), Pepe (1997; 1998). Bayesian methods have been used widely in diagnostic medicine, but mainly for ordinal data, e.g. Rutter and Gatsonis (1995; 2001), Peng and Hall (1996), Helmich et al. (1998), Ishwaran and Gatsonis (2000). Recently, O'Malley et al. (2001) applied Bayesian methods to continuous data, and evaluated the sensitivity of the results (e.g. the accuracy of PSA testing) to different priors. However, none of these methods dealt with clustered continuous outcome data.

The remainder of the article is organized as follows. In Section 2, we present the multi-center prospective RDOG trial that motivated this work. In Section 3, we develop a Bayesian hierarchical non-linear regression model for ROC analysis. We perform a secondary analysis of the accuracy of the diagnostic markers, PSA level and Gleason score, for prostate cancer staging, validated against a gold standard based on radical prostatectomy. We focus on the differences and similarities between clusters (clinical centers) and risks for prostate cancer (Gleason score). The results of the prostate biopsy analysis are provided in Section 4, followed by a discussion in Section 5.

2 A Motivating Example: Prostate Cancer Biopsy

2.1 Subject population

Subjects were enrolled from 3 participating centers, here labeled as A to C, as part of a multi-collaborative RDOG trial reported by Tempany et al. (1994). Among 213 subjects with prostate cancer, 180 cases had both PSA level and Gleason score taken at the baseline biopsy. After excluding the missing data in either variable 180 cases were available for our analyses. We treat PSA as the main outcome variable and Gleason score as an additional covariate, to be validated against the gold standard described below. Other baseline patient characteristics (e.g., age, sex, and race) were also collected. However, they did not have a strong influence on diagnostic accuracy, and therefore were omitted from this analysis.

2.2 Gold standard

Radical prostatectomy provided a four-stage measure of cancer stage for all subjects. However, we used the binary gold standard defined as either local (stages A and B) or advanced disease (periprosthetic invasion of tumor and spread of disease to the seminal vesicles and lymph nodes) for ROC analysis.

2.3 Data collection

The standardized forms recording the subject's demographics and medical profiles, the imaging studies results, and the surgery and pathology findings were maintained at the American College of Radiology office (http://www.acr.org/links/rdog_intro.html). These forms were reviewed for consistency before starting the final data analysis by the Statistical Center in the Department of Health Care Policy, Harvard Medical School, Boston, MA, USA.

3 Hierarchical Regression Model in ROC Analysis

To validate the accuracy of PSA level and Gleason score, we developed a new Bayesian hierarchical non-linear regression model. We incorporated the possible correlation between observations in the each of the 3 clusters (clinical centers). Statistical methods described in detail below included: summary statistics, counts and proportions, analysis of variance (ANOVA), hierarchical regression incorporating the effects of cluster (clinical center) and cancer risk (low, intermediate, and high), and area under the ROC curve (AUC).

3.1 Summary statistics

For each clinical center, univariate summary statistics of the variables were computed. The proportions and counts for binary disease stage (local and advanced disease), and means and standard errors (SE) of PSA, Gleason score and other covariates, were compared across clusters. The Pearson correlation coefficient between PSA and Gleason score was also calculated. An ANOVA was used to test for differences in the mean PSA level and the mean Gleason scores across clusters.

3.2 Normality transformation

Because PSA is a positively valued continuous outcome variable, a non-linear Box-Cox normality transformation was applied to each cluster before applying the bi-normal model (Metz et al., 1998). A transformation coefficient of 0 suggests the natural log transformation should be used; however, if the coefficient is far from 0, then using the log transformation could result in misleading results.

For each cluster, the Box-Cox transformation was used to parameterize an “optimal” non-linear transformation of the positive-valued diagnostic outcome data. This is given by

$$h(Y_{ij}, \lambda_i) = \frac{(Y_{ij}^{\lambda_i} - 1)}{\lambda_i}, \text{ if } \lambda_i \neq 0 \text{ and } h(Y_{ij}, \lambda_i) = \log(Y_{ij}), \text{ if } \lambda_i = 0, \text{ as the natural log (base } e) \text{ is}$$

a special case of the Box-Cox transformation when $\lambda_i = 0$.

We now develop the Bayesian nonlinear hierarchical regression model. For subject j in cluster i ($i = 1, \dots, m$ clusters with $j = 1, \dots, n_i$ cluster-specific subjects), let Y_{ij} be the main outcome variable (e.g., PSA after the Box-Cox Transformation), X_{ij} be the gold standard (e.g., local or advanced prostate cancer) and Z_{ij} be a vector of additional covariates. We assume that $X_{ij} = 1$ indicates disease and $X_{ij} = 0$ indicates non-disease, and that Z_{ij} contains indicator variables for the levels of Gleason score. More covariates (e.g., age) may be included in the model simply by adding additional components to Z_{ij} . The Bayesian hierarchical regression model can be expressed as

$$h(Y_{ij}, \lambda_i) = f(X_{ij}, Z_{ij}, \beta_i) + g(X_{ij}, Z_{ij}, \theta_i) \varepsilon_{ij}$$

where β_i and η_i are vectors of regression and variance function parameters for cluster i ($i = 1, \dots, m$); λ_i is the scalar parameter characterizing the Box-Cox non-linear transformation; and the individual error terms ε_{ij} ($j = 1, \dots, n_i$) follow mutually independent standard normal distributions with mean 0 and variance 1 independent of β_i and θ_i . By the independence of θ_i and ε_{ij} it follows that $E[g(X_{ij}, Z_{ij}, \theta_i) \varepsilon_{ij}] = 0$ at fixed values of X_{ij} and Z_{ij} .

We assumed linear regression equations allowing the variance to differ among all clusters. The expectation and variance of the transformed outcomes are, respectively:

$$f(X_{ij}, Z_{ij}, \beta_i) = E\{h(Y_{ij}, \lambda_i)\} = \beta_{0i} + \beta_{1i}X_{ij} + \beta_{2i}Z_{ij} + \beta_{3i}X_{ij}Z_{ij}$$

$$g^2(X_{ij}, Z_{ij}, \theta_i) = \text{Var}\{h(Y_{ij}, \lambda_i)\} = (\theta_{0i} + \theta_{1i}X_{ij})^2,$$

with $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})$ and $\theta_i = (\theta_{0i}, \theta_{1i})$. An interaction effect between disease status and Gleason score leads to different ROC curves for different levels of Gleason score, whereas if $\beta_{3i} = 0$ the ROC curve is the same across all levels of Gleason score (O'Malley et al., 2001). Because of the limited size of the data set, we place the restrictions $\beta_{2i} = \beta_{2*}$ (indicating constancy across clusters) and $\beta_{3i} = 0$ on the above general model when analyzing the PSA data. Note that an alternate way of modeling the variance is to use an exponential function as opposed to a quadratic function.

Because pilot data for the RDOG study were not available, in this investigation a diffuse prior was used. The prior distributions for all regression coefficients, the components of the β 's, were assumed to have independent normal $N(0, 10^6)$ distributions (the variance of this distribution is so large that the density is essentially flat). The prior distributions for the variance parameters were specified as $\sigma_{ij}^2 = (\theta_{0i} + \theta_{1i}Z_{ij})^2 \sim \text{IG}(0.001, 0.001)$, where IG denotes an inverse gamma distribution. For certain simple problems diffuse priors yield similar inferences to frequentist methods such as maximum likelihood. However, a key advantage of Bayesian

methods is that it provides a formal procedure for solving inference problems. Furthermore, the results from a diffuse prior analysis such as this may be used for a subsequent study.

We developed C programs to implement Markov-Chain Monte Carlo (MCMC) methods via the Gibbs sampler (see, e.g., Gelman et al., 1995 and Gamerman, 2002 for a description of MCMC methodology). We used a burn-in of 2000 iterations and a main simulation of 10000 iterations. Furthermore, we verified our results using BUGS, a freely available software program.

3.3 ROC analysis

The accuracy of a diagnostic test may be presented in terms of an ROC curve, a plot of sensitivity (or true positive fraction) versus 1-specificity (or false positive fraction) at all possible decision thresholds of a diagnostic test at given values of the predictors. In a regression analysis under the bi-normal model, the ROC parameters may be derived from the regression coefficients. Denoting specificity as $1 - p$, the corresponding point on an ROC curve is $(1 - \text{specificity}, \text{sensitivity})$, or $[p, \Phi\{\mu + v\Phi^{-1}(p)\}]$, where Φ is the cumulative distribution function of a standard normal distribution with mean 0 and variance 1, Φ^{-1} is the quantile function of a standard normal distribution, and μ and v are two parameters specific to the ROC curve. The ROC curve corresponding to the model used to analyze the PSA data in this paper, in which clustering is accounted for but there is no interaction term between X_{ij} and Z_{ij} , is given by

$$\mu_i = \frac{f(1, X_{ij}, Z_{ij}, \beta_i) - f(0, X_{ij}, Z_{ij}, \beta_i)}{g(1, X_{ij}, Z_{ij}, \theta_i)} = \frac{\beta_{1i}}{|\theta_{0i} + \theta_{1i}|}, \quad v_i = \frac{|g(0, X_{ij}, Z_{ij}, \theta_i)|}{|g(1, X_{ij}, Z_{ij}, \theta_i)|} = \frac{|\theta_{0i}|}{|\theta_{0i} + \theta_{1i}|}.$$

Several summary measures of diagnostic accuracy are described in Campbell et al. (1994) and O'Malley et al. (2001). We focus on the area under the curve (AUC), a popular overall measure of diagnostic accuracy given in Hanley and McNeil (1982) and Faraggi and Reiser (2002). The AUC for cluster i at any value of Z_{ij} is given by

$$A_i = \Phi\left(\frac{\mu_i}{\sqrt{1 + v_i^2}}\right).$$

By applying a Box-Cox transformation to induce normality, inferences about quantities such as A_i are likely to be more reliable because the assumption of normality inherent to its calculation is more reasonable. Other summary measures can be derived as functions of μ_i and v_i but are omitted here.

3.4 Effects of cluster and risk on ROC curves

We used Bayesian inference to derive the posterior means and standard deviations of the estimated ROC parameters and the AUC via MCMC computation. The AUC of each risk group was also computed. The risk groups were determined from Gleason score as low (L) risk if Gleason score ≤ 6 , intermediate (I) risk if Gleason score = 7, and high (H) risk if Gleason score ≥ 8 (D'Amico et al., 2000). One often wishes to compare the AUCs for different ROC curves. The Bayesian model derived here enables complicated inferences such as tests of differences in areas between two or more ROC curves to be evaluated simply by recording the proportion of MCMC draws in which the area for one curve is less than that of another curve (Hellmich et al., 1998). Alternatively, Bayesian 95% highest posterior density (HPD) regions may be constructed for the difference in AUCs. Similar comparisons can be made in the frequentist domain by using DeLong et al. (1988)'s non-parametric approach which utilizes fewer

parametric assumptions. Because this frequentist calculation only applies to unclustered data, we do not compare the Bayesian and frequentist methods for comparing areas of ROC curves.

3.5 Comparison with Non-Bayesian Approach

For comparison with the Bayesian estimates of the transformation parameters and summary ROC measures, we obtained maximum likelihood estimates of the model parameters using a stratified multiple linear regression model. This model differs from the model for the Bayesian analysis by not allowing information to be shared between clusters, enabling the effect of pooling information to be evaluated.

4 Results

4.1 Summary statistics

Of the 180 subjects, 66 (37%) subjects had local disease, and 114 (63%) had advanced disease. There were 25 (14%), 93 (52%) and 62 (34%) subjects enrolled from Centers A to C. The counts of the 4 tumor stages by the binary gold standard were presented in Table 1.

The PSA level (outcome variable) of these subjects ranged from 0.1 to 58.0 (mean = 1.66 and standard deviation = 9.37). Figure 1 displays the nonparametric density estimates of the PSA level by the gold standard, suggesting that the PSA data were quite skewed, requiring normality transformations before regression analysis. Based on Gleason scores, 88 subjects (49%) were classified as low risk, 51 (28%) as intermediate risk, and 41 (23%) as high risk. Table 2 reports the summary statistics of PSA, Gleason score, and age.

To justify our regression approach, the overall cluster-to-cluster variability in both mean PSA and Gleason scores was analyzed using one-way ANOVAs across the three centers, yielding $p = 0.04$ for comparing mean PSA's and $p = 0.14$ for comparing mean Gleason scores, respectively. Age was not a significant predictor. Summary statistics of these variables by cluster and binary gold standard are presented in Table 2.

4.2 Normality transformation

The posterior distributions of the estimated Box-Cox coefficient (the power transformation parameter) are displayed in Figure 2. The posterior means and standard errors (SE) of the transformation parameters (the λ 's) are 0.39 (SE = 0.12), 0.40 (SE = 0.08), and 0.30 (SE = 0.09) for Centers A, B and C, respectively. The clinical interpretation of this result is that the PSA measurements at center C is further from normal than those at centers A and B. Under a pooled analysis, an overall transformation of 0.37 (SE = 0.08) was obtained. In practice, an ad-hoc log transformation of positive data is often employed. Our results suggest that simply taking a log transformation of positive-valued marker data such as PSA is not appropriate before assuming a bi-normal model.

4.3 Effects of cluster and risk on ROC curves

The ROC curves showed that diagnostic accuracy differed among the clinical centers (Table 3 and Figure 3). Diagnostic accuracy was the highest for Center C, AUC = 0.75, but lowest for Center A, AUC = 0.63 (Table 4). The cluster-level differences might reflect differences in the case-mix of subjects across centers (Tables 1 and 2). Unfortunately, the sample size for Center A was much lower ($n = 25$) than for the remaining two centers. To thoroughly explore the differences in subject populations among center, it would have been better to have a larger study with equal sized clusters. The ROC analyses also showed that PSA was most accurate for diagnosing staging among intermediate risk subjects with Gleason scores equal to 8, AUC = 0.77, but least accurate for low risk subjects with Gleason scores no greater than 6, AUC = 0.67 (Table 4 and Figure 4).

4.4 Comparison with non-Bayesian regression

In Table 5, we list the non-Bayesian and Bayesian results by center. The estimated power coefficient of the Box-Cox transformations were: 0.26, 0.39 and 0.18 for non-Bayesian vs. 0.39, 0.40 and 0.30 for Bayesian methods. The AUC's were 0.53, 0.71, and 0.77 for non-Bayesian vs. 0.63, 0.70, and 0.75 for Bayesian methods. It is not surprising that the parameter estimates differed the most for clinical center A as it only has a small number of subjects. The Bayesian model "borrowed strength" from all three centers, whereas the classical methods partitioned the data for each analysis and thus is unreliable when sample sizes are very small (Table 5).

5 Discussion

Multi-center clinical studies or quality of care studies call for complex and efficient study designs and analyses. As derived from the RDOG study, health care utilization and outcome data often exhibit a multi-level structure. The data structure includes individual subjects, physicians, clinical centers, and geographic regions, thus allowing for several levels in a hierarchical analysis. Cluster sample size may vary substantially at each level. In addition, both individual- and cluster-level covariates are often available for measuring, for example, disease severity and co-morbidity for individual subjects, and the location, size, and organizational characteristics of a particular clinical center.

We developed a Bayesian hierarchical non-linear regression model for ROC analysis, and applied it to baseline diagnostic data from the multi-center prospective RDOG trial. We validated the accuracy of PSA level and Gleason score in prostate cancer staging based on the AUC. Specifically, we conducted Bayesian regression analysis to incorporate a center effect in a multi-center clinical trial. We applied optimal normality transformation before assuming regression errors have normal distributions. The proposed Bayesian regression framework provides new methodology for analyzing clustered data, which requires a transformation before a family of outcome distribution can be assumed. Our method is advantageous in a multiple cluster setting because it allows information from different clusters to be pooled, enabling more efficient parameter estimation.

The Bayesian regression model proposed herein may be generalized and extended to the analysis of diagnostic markers in many multi-center prospective and retrospective clinical trials or observational studies. An important feature of this model is the ability to combine information across units. Because the motivating example contains only three centers (clusters) and we had minimal prior information, a frequentist-style fixed effects model would be expected to yield similar results to those presented here. However, non-trivial differences in results were observed for parameters that are nonlinear such as the transformation parameters for whom the posterior distributions are not perfectly symmetric, or for quantities for which there was little information such as the AUC for center A.

Finally, we recapitulate that we did not observe differences based on other variables such as sex and race in our preliminary analyses of the RDOG data. However, race has been reported to be useful in the literature, e.g. Moul et al. (1995) reported that African-American men with newly diagnosed prostate cancer have higher PSA values at initial diagnosis than white men.

Acknowledgements

This work was supported in part by Grants R01LM007861-01A1, R03HS13234-01, and PO1CA41167, awarded by the National Institutes of Health, Department of Health and Human Services, USA. We thank Drs. Barbara J. McNeil, MD, PhD, Clare M. C. Tempany, MD and Daryl J. Caudry, MS of Harvard Medical School and Brigham and Women's Hospital for helpful suggestions and assistance in preparing the data.

References

- American Cancer Society (2004). *Cancer Facts & Figures*¹⁶
- Campbell G. General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* 1994;13:499–508.
- D'Amico AV, Schultz D, Loffredo M, Dugal R, Hurwitz M, Kaplan I, Beard CJ, Renshaw AA, Kantoff PW. Biochemical outcome following external beam radiation therapy with or without androgen suppression therapy for clinically localized prostate cancer. *JAMA* 2000;284:1280–1283. [PubMed: 10979115]
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–845. [PubMed: 3203132]
- Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Statistics in Medicine* 2002;21:3093–3106. [PubMed: 12369084]
- Feneley MR, Partin AW. Diagnosis of localized prostate cancer: 10 years of progress. *Current Opinion in Urology* 2000;10:319–327. [PubMed: 10918970]
- Gamerman, D. (2002). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC: Boca Raton, FL.
- Gatsonis CA. Random-effects models for diagnostic accuracy data. *Academic Radiology* 1995;2:S14–S21.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis* Chapman & Hall: London.
- Goldstein, H. (1995). *Multilevel Statistical Models* Halstead Press: New York.
- Hanley JA, McNeil BJ. The meaning and use of the area under a ROC curve. *Radiology* 1982;143:27–36.
- 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]
- Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canadian Journal of Statistics* 2000;28:731–775.
- Lehrer S, Diamond EJ, Stagger S, Stone NN, Stock RG. Serum insulin level, disease stage, prostate specific antigen (PSA) and Gleason score in prostate cancer. *British Journal of Cancer* 2002;87:726–728. [PubMed: 12232754]
- Lieberfarb ME, Schultz D, Whittington R, Malkowicz B, Tomaszewski JE, Weinstein M, Wein A, Richie JP, D'Amico AV. Using PSA, biopsy Gleason score, clinical stage, and the percentage of positive biopsies to identify optimal candidates for prostate-only radiation therapy. *International Journal of Radiation Oncology, Biology, Physics* 2002;53:898–903.
- Lindley, D. V. (1982). Bayesian Inference. *International Encyclopedia of Statistics* (Kotz, S. and Johnson, N. L., Eds.) 1197–224.
- 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]
- Moul JW, Sesterhenn IA, Connelly RR, Douglas T, Srivastava S, Mostofi FK, McLeod DG. Prostate-specific antigen values at the time of prostate cancer diagnosis in African-American men. *JAMA* 1995;274:1277–1281. [PubMed: 7563532]
- 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* 2001;8:713–725. [PubMed: 11508750]
- Peng F, Hall WJ. Bayesian analysis of ROC curves using Markov-Chain Monte Carlo methods. *Medical Decision Making* 1996;16:404–411. [PubMed: 8912302]
- Pepe MS. A regression modeling framework for ROC curves in medical diagnostic testing. *Biometrika* 1997;84:595–608.
- Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* 1998;54:124–135. [PubMed: 9544511]

- Polascik TJ, Oesterling JE, Partin AW. Prostate specific antigen: a decade of discovery – what we have learned and where we are going. *Journal of Urology* 1999;162:293–306. [PubMed: 10411025]
- Pound CR, Partin AW, Eisenberger MA, Chan DW, Pearson JD, Walsh PC. Natural history of progression after PSA elevation following radical prostatectomy. *JAMA* 1999;281:1591–1597. [PubMed: 10235151]
- Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. *Academic Radiology* 1995;2(Suppl 1):S48–56. [PubMed: 9419705]
- Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001;20:2865–2884. [PubMed: 11568945]
- Seltzer SE, Getty DJ, Tempany CM, Pickett RM, Schnall MD, McNeil BJ, Swets JA. Staging prostate cancer with MR imaging: a combined radiologist-computer system. *Radiology* 1997;202:219–226. [PubMed: 8988214]
- Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. *Journal of Royal Statistical Society, Series A* 1994;157:357–416.
- Stamey TA, McNeal JE, Yemoto CM, Sigal BM, Johnstone IM. Biological determinants of cancer progression in men with prostate cancer. *JAMA* 1999;281:1395–1440. [PubMed: 10217055]
- Tempany CM, Zhou X, Zerhouni EA, Rifkin MD, Quint LE, Piccoli CW, Elis 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]
- Tosteson ANA, Begg CB. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1988;8:204–215. [PubMed: 3294553]
- Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* 2000;5:621–631.
- Zou KH, Hall WJ. Semiparametric and parametric transformation models for comparing diagnostic markers with paired design. *Journal of Applied Statistics* 2002;29:803–816.

Figure 1.
Nonparametric probability density estimates of PSA level by the gold standard.

Figure 2.
Posterior distributions of the optimal non-linear Box-Cox transformation by clinical center.

Figure 3.
Estimated ROC curves by clinical center.

Figure 4.
Estimated ROC curves by risk group.

Table 1

Frequency counts of the cancer stage by cluster (clinical center) and the gold standard (local vs. advanced prostate cancer).

Clinical Center	Gold Standard	Sample Size (%)	Cancer Stage			
			I	II	III	IV
A (<i>n</i> = 25)	Local	11 (44%)	11 (100%)	0 (0%)	0 (0%)	0 (0%)
	Advanced	14 (56%)	0 (0%)	11 (79%)	3 (21%)	0 (0%)
B (<i>n</i> = 93)	Local	42 (45%)	42 (100%)	0 (0%)	0 (0%)	0 (0%)
	Advanced	51 (55%)	0 (0%)	39 (76%)	7 (14%)	5 (10%)
C (<i>n</i> = 62)	Local	13 (21%)	13 (100%)	0 (0%)	0 (0%)	0 (0%)
	Advanced	49 (79%)	0 (0%)	42 (86%)	6 (12%)	1 (2%)

Table 2

Summary statistics of the sample sizes, means, and standard errors of the biopsy variables (age, PSA, Gleason score), by cluster (clinical center) and the gold standard (local vs. advanced prostate cancer), along with the correlation coefficients of PSA and Gleason scores.

Clinical Center	Gold Standard	Sample Size (%)	Age Mean (SE)	PSA Mean (SE)	Gleason Mean (SE)	Correlation r (PSA, Gleason)
A	Local	11 (44%)	62.91 (2.27)	7.79 (1.60)	6.00 (0.30)	0.38
$n = 25$	Advanced	14 (56%)	64.43 (1.34)	1.67 (3.13)	6.64 (0.29)	0.03
B	Local	42 (45%)	59.62 (0.82)	5.60 (0.54)	5.81 (0.15)	0.33
$n = 93$	Advanced	51 (55%)	61.06 (0.87)	12.59 (1.28)	7.35 (0.14)	0.49
C	Local	13 (21%)	63.23 (0.51)	5.26 (0.84)	5.54 (0.31)	-0.12
$n = 62$	Advanced	49 (79%)	63.76 (0.79)	15.20 (1.62)	6.39 (0.25)	0.35

Note: Standard Error = Standard Deviation/(Sample Size)^{1/2}, i.e., SE = SD/ $n^{1/2}$.

Table 3

Frequency distribution of the risk group (trichotomized Gleason scores) by clinical center.

Clinical Center	Gleason Score (Risk Group)							
	2 Low (L)	3	4	5	6	7 Intermediate (I)	8 High (H)	9
A (<i>n</i> = 25)	0 (12)	0	0	7	5	11 (11)	1 (2)	1
B (<i>n</i> = 93)	0 (43)	1	1	15	26	25 (25)	19 (25)	6
C (<i>n</i> = 62)	2 (33)	0	6	16	9	15 (15)	8 (14)	6

Note: The numbers in parentheses are totals over the range of values in the line above.

Table 4
Areas under the ROC curves by cluster and risk group.

Variable		Sample Size (%)	AUC (SE)
Clinical Center	A	25 (14%)	0.63 (0.07)
	B	93 (52%)	0.70 (0.05)
	C	62 (34%)	0.75 (0.06)
Risk Group	Low (L)	88 (49%)	0.67 (0.06)
	Intermediate (I)	51 (28%)	0.77 (0.07)
	High (H)	41 (23%)	0.70 (0.13)

Note: Standard Error = Standard Deviation/(Sample Size)^{1/2}, i.e., SE = SD/n^{1/2}.

Table 5

Bayesian vs. non-Bayesian estimates of the regression parameters and corresponding AUC's.

Methods and Center	Sample Size (#Loc, #Adv)	Transformation λ	β_0	β_1	β_2	β_3	σ_1^2	σ_2^2	AUC
(1) non-Bayesian									
A	11 + 14	0.26	2.27	0.18	0.23	0.59	2.13	3.16	0.53
B	42 + 51	0.39	2.04	1.19	0.40	1.34	1.95	2.50	0.71
C	13 + 49	0.18	1.62	1.02	0.89	0.98	0.66	1.27	0.77
Pooled	66 + 114	0.33	1.95	1.13	0.63	1.21	1.71	2.62	0.73
(2) Bayesian									
A	11 + 14	0.39	2.07	1.06	0.62	1.34	1.96	2.65	0.63
B	42 + 51	0.40	2.04	1.22	0.62	1.34	1.47	1.71	0.70
C	13 + 49	0.30	1.97	1.31	0.62	1.34	1.02	1.64	0.75
Pooled	66 + 114	0.37	1.96	1.15	0.57	1.17	1.77	2.70	0.71

Note: β_0 = regression coefficient of the intercept; β_1 = regression coefficient of the disease status; β_2 = regression coefficient of low risk; β_3 = regression coefficient of high risk. σ_1^2 = variance of the local disease stage; σ_2^2 = variance of the advanced disease stage.