Journal of Data Science **10**(2012), 37-49

Using Appropriate Functional Forms for Continuous Variables and Improving Predictive Accuracy in Developing the Risk Model of Clostridium Difficile Infection

Yan Yan*, Kimberly A. Reske, Victoria J. Fraser,
Graham A. Colditz and Erik R. Dubberke
Washington University School of Medicine

Abstract: Simple parametric functional forms, if appropriate, are preferred over more complicated functional forms in clinical prediction models. In this paper, we illustrate our practical approach to obtaining the appropriate functional forms for continuous variables in developing a clinical prediction model for risk of Clostridium difficile infection. First, we used a nonparametric regression smoother to establish the reference curve. Then, we used regression spline function-restricted cubic spline (RCS) and simple parametric forms to approximate the reference curve. Based on the shape of the reference curve, the model fit information (AIC), and the formal statistical test (Vuong test), we selected the simple parametric forms to replace the more elaborated RCS functions. Finally, we refined the simple parametric forms in the multiple variable regression model using the Wald test and the likelihood-ratio test. In addition, we compared the calibration and discrimination aspects between the model with appropriate functional forms and the model with simple linear terms. The calibration χ^2 (8.4 versus 10) and calibration plot, the area under ROC curve (0.88 vs 0.84, $p < 0.05$), and integrated discrimination improvement (0.0072, $p < 0.001$) indicated the model with appropriate forms was better calibrated and had higher discrimination ability.

Key words: Clinical prediction model, functional form, predictive accuracy.

1. Introduction

Clinical prediction models, which relate patient characteristics to a certain outcome, are a useful tool for clinicians in their routine clinical practice and clinical research. In routine clinical practice, these models facilitate diagnostic testing, treatment selection, and follow-up planning. In clinical research, these

*Corresponding author.

models facilitate patient stratification into different homogeneous risk groups to better test the treatment effect of new agents or procedures [1].

Clinical prediction models are often obtained by regressing the outcome variable on a set of patient characteristics. The data for common clinical outcomes have several different types − continuous, categorical, and the time to event. Different types of outcome data require different distributional assumptions in a regression model. For a continuous outcome, the normality and constant variance assumptions are often postulated. For a categorical outcome, binomial or multinomial distribution is commonly assumed. For a time-to-event outcome, various distributional assumptions about the event time (for example, exponential, Weibull, and log normal) can be assumed in parametric survival models. In addition to the required distributional assumptions for the outcome variables, a regression model assumes certain shapes of the relationships between the continuous variables and the outcome variable (or transformed outcome variable). The shapes can range from a simple linear relationship to some unspecified arbitrary trends. Many of these shapes can be represented by certain functional forms in the regression model. The most commonly used functional forms include a single linear term, the second or third order polynomials, the logarithm, etc. More elaborate forms may include linear or polynomial regression splines. There is no functional form for a smoother, so there cannot be an equation with a smoother for the purpose of easy prediction. Use of the appropriate functional form for a continuous variable is crucial for valid predictions because the expected value of outcome can be different for the same value of a continuous variable with different functional forms.

Clinical prediction models are developed with intent to apply them to similar patient populations. Therefore, functional forms that represent the underlying patterns as closely as possible, but are not sensibly affected by the idiosyncrasies in the dataset, should be used. In this paper, we describe our approach to selecting appropriate functional forms for continuous variables in developing a risk prediction model for Clostridium difficile infection (CDI), and we compare the predictive accuracy between the model with appropriate functional forms and the model with simple linear terms.

2. Study Subjects and Data

The study population consisted of 35,350 patients admitted to a tertiary-care hospital for at least 48 hours during 2003. There were 329 CDI cases among the study population. Data on patient demographics, medications, and laboratory results were collected electronically from hospital databases. Patients already known to be at high risk for CDI (those with a recent history of CDI or those admitted to leukemia/bone marrow transportation wards) were excluded. The

Washington University Human Research Protection Office approved this study.

3. Methods

3.1 Selecting Appropriate Functional Forms in Univariate Setting

In developing the risk model of Clostridium difficile Infection, we used the logistic regression model with CDI infection (Yes/No) as the outcome variable and the patient characteristics as the explanatory variables. After a series of variable selections, 10 clinically important and statistically significant variables were identified for the prediction model, of which five variables were on a continuous scale: age at admission (AGE), a modified acute physiology score (MOD_{APS}), a modified measurement of colonization pressure based on clinically symptomatic CDAD cases (CP) [2], days on high risk antibiotics (HRABX), and number of admissions within 60 days (ADMIT60D).

For each continuous variable, a locally weighted scatterplot smoother (LOWESS) was used to depict the arbitrary shape of its relationship with the outcome − CDI infection (Yes/No) [3]. For a given value of a continuous variable, the value of the smoother was given by a predicted value from the weighted regression model using data points within a local neighborhood. The weights for the local data points were defined by a tricubic function that weighed closer data points more than further ones [3]. The resultant curve was used as the reference curve. Then, the restricted cubic spline (RCS) functions were used to approximate the reference curve. RCS represented the fit as piecewise cubic polynomials with the first and the last pieces forced to be linear. The pieces were defined by the regions, which were separated by a sequence of breakpoints, called knots. RCS could fit sharply curving shapes, and at the same time could avoid the poor behavior in the fit in the first and the last region [4]. Finally, simple parametric forms (a single linear term, piecewise linear terms, low order polynomials, or logarithm) were compared with RCS. To determine if a simple parametric form or RCS should be used, visual plots and Akaike Information Criterion (AIC) were compared. To formally test the superiority of one model over another, the Vuong test [5] was used. Different from AIC, the Vuong test takes into account the probabilistic nature of the statistical model selection. Under the null hypothesis that the two models are equally adequate for the data, the Vuong statistic is distributed as a standard normal random variable.

3.2 Refining Appropriate Functional Forms in the Multiple Variable Regression Setting

After appropriate simple parametric forms were selected in the univariate

setting, a multiple variable regression model was fit with all five continuous and categorical variables in the model. Using Wald χ^2 for each parameter estimate, the likelihood ratio test, and the partial residual plot, the simple parametric functional forms for the continuous variables was refined further.

3.3 Comparing Predictive Accuracy

After the appropriate forms for the five continuous variables were identified, we fit two prediction models: Model 1, in which the appropriate forms for the five continuous variables were used plus other 5 important categorical variables, and Model 0 in which a simple linear term was used for the five variables in addition to the same categorical variables as used in Model 1.

To compare predictive accuracy between Model 1 and Model 0, the likelihood ratio Chi-square was used to test which model fit the data better. Then the calibration and discrimination aspects of predictive accuracy were compared between the two models. For calibration, the Hosmer-Lemshow (calibration) χ^2, and calibration plots were compared [6]. The calibration χ^2 summarizes the difference between the observed and expected frequencies in several groups (often 10) defined by the predicted probabilities. Given the same number of groups, a larger calibration χ^2 value for a model indicates poorer calibration for the model. A calibration plot displays the relationship between the predicted and the true probabilities, and a departure from the diagonal line indicates poor calibration.

For discrimination, the area under receiver-operator characteristic (ROC) curve (AUC) of the two models was compared using Delong's method [7]. A larger value of AUC indicates a better separation of the patients with and without CDI, with a value of 0.5 being random assignment of the patients into two groups. In addition to AUC, the improvement in discrimination ability was quantified by Integrated Discrimination Improvement (IDI), and the improvement in sensitivity and specificity components of the IDI was assessed [8]. IDI is the sum of the improvement in the average sensitivity and the average specificity. With the assumption of independence between the infected and non-infected groups, the variance of IDI is the sum of the variances of two corresponding components. Under the null hypothesis of IDI = 0, the ratio of the estimated IDI over the square root of its variance is asymptotically distributed as a standard normal variable. To assess the improvement in the sensitivity component, the difference in the average sensitivity between the two models was obtained. The average sensitivity is defined as the integral of sensitivity over all possible cut-off values of predicted probabilities in those with the outcome of interest (disease group), which can be estimated by the mean probability in that group. The one-sample paired t-test was used to test the statistical significance in the average improvement in sensitivity. The improvement in the specificity component was assessed

in the same way among study subjects without the outcome of interest.

The free software R [9] and some functions from Design Package [10] were used for all statistical computation and graphics.

4. Results

4.1 Selecting Appropriate Functional Forms in the Univariate Setting

Figure 1 is a panel of density histograms depicting the distribution of five continuous variables. Except for the variable AGE, all other four variables were right skewed. Especially for CP, HRABX, and ADMIT60D, most observations were located at the low end of the distributions. Figure 2 is a panel of plots comparing the RCS curves with the simple parametric curves for five continuous variables. In each plot, the solid line represented the LOWESS fit (the reference line). The dashed line and dotted line were the RCS and the simple parametric fit, respectively. For both AGE and MOD_APS, the simple parametric form was the second order polynomials (square term), and for CP, HRABX and ADMIT60D, the simple parametric form was the two piece linear terms with one knot. For CP, the knot was located at 0.3. For HRABX and ADMIT60D, the knots were located at 8 and 1, respectively. The second order polynomials for AGE and MOD_APS were used since the reference lines looked like concave curves, and the two piece linear term was used for the other three variables since the reference lines had one clear turning point.

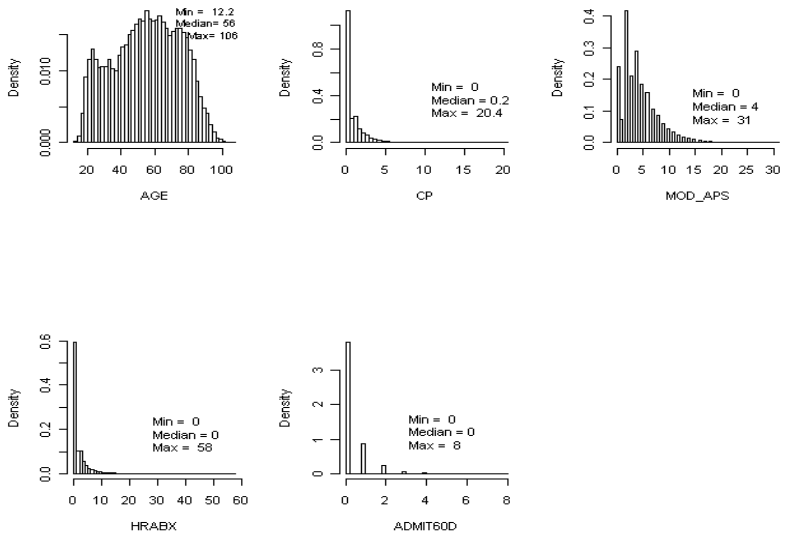

Figure 1: Histogram of density plot of 5 continuous variables

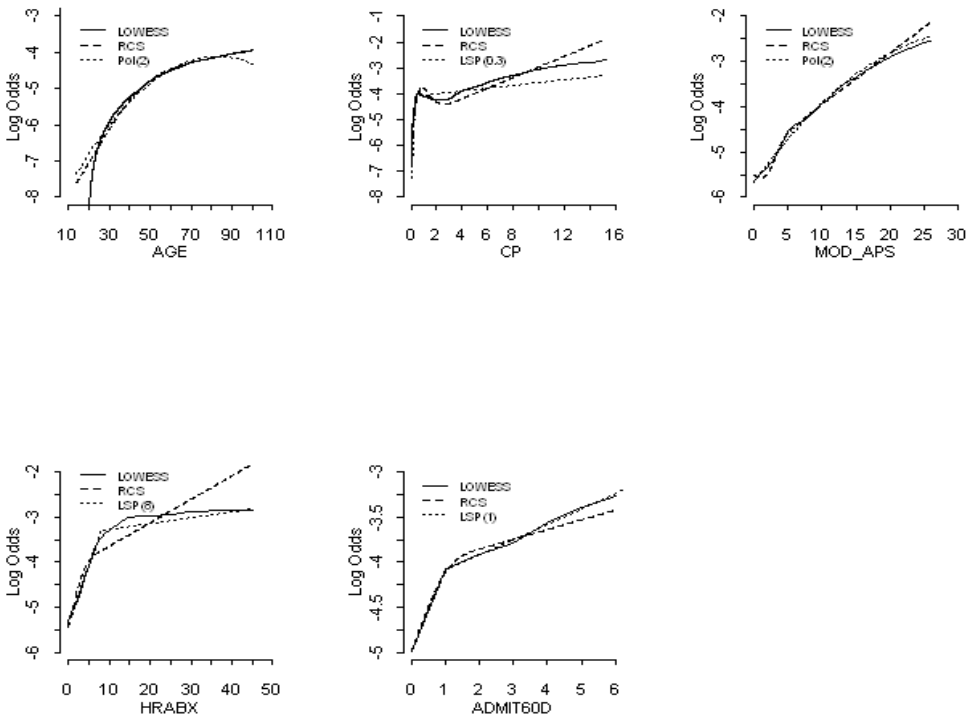

Figure 2: Comparison of selected simple parametric forms with RCS function. Solid line: LOWESS fit, dashed line: RCS fit. Dotted line: simple parametric fit - LSP(k), two piece linear terms with the knot at k; Pol(k), kth order polynomials

Comparison of the RCS curves with simple parametric forms revealed that the simple parametric forms tracked the reference curves more closely than the RCS for the variables MOD_APS, HRABX, and ADMIT60D. For AGE, the RCS and simple parametric curves were very close before 90 years old, and then the simple parametric curve departed from the reference curve. For CP, the RCS curve tracked the reference curve reasonably well when CP was less than 8, after that it departed from the reference curve drastically as the value of CP increased. In contrast, the two piecewise linear curve continued to track the reference curve when the value of CP was beyond 8, although it went underneath the reference curve. Table 1 presents the comparison of the model fit between the RCS and the simple parametric forms. For all five variables, AIC was smaller in the simple parametric fit, especially for the variable CP and HRABX. The results from the Vuong test indicated that the superiority of the simple parametric model fit for CP and HRABX over the RCS fit could not be explained by the chance alone, with p-value < 0.001 and p-value $= 0.002$, respectively.

Table 1: Comparison of model fit: restricted cubic spline and simple parametric form

Akaike Information Criterion (AIC)			
Variable	Restricted Cubic Spline	Simple Parametric Form	Vuong test
AGE	3622.8	3619.7	$p = 0.362$
CP	3441.2	3396.4	$p < 0.001$
MOD_APS	3585.6	3585.4	$p = 0.173$
HRABX	3560.9	3538.5	$p = 0.002$
ADMIT60D	3664.7	3664.0	$p = 0.241$

Simple parametric form: second-order polynomials for AGE and MOD_APS, and two piecewise linear for CP, HRABX, and ADMIT60D.

4.2 Refining Simple Parametric Forms in the Multiple Variable Regression Setting

Table 2 presents the results for comparison of the model with selected appropriate functional forms in the univariate setting (selected model) versus the model with refined appropriate functional forms in the multiple variable regression setting (refined model). In the refined model, the square terms for variable AGE and MOD_APS were dropped from the selected model. The likelihood ratio test of the refined model (LR $\chi^2 = 744.49$, DF = 13) against the selected model (LR $\chi^2 = 745.69$, DF = 15) yielded a LR χ^2 of 1.2 with 2 degrees of freedom (P = 0.548), indicating adequacy of the refined model in place of the selected model. The two models had the same value for generalized R sqaure, c-index, and Brier score. The partial residual plots for AGE and MO_APS do not indicate the inadequacy of a simple linear term for these two variables.

4.3 Comparison of Predictive Accuracy

Model fit: The model χ^2 for Model 1 was 744.9, with 13 degrees of freedom, and for Model 0 the model χ^2 was 577.1 with 10 degrees of freedom. The likelihood ratio test comparing the two models gave the χ^2 value of 167.8 with 3 degrees of freedom, yielding a p-value < 0.001. The generalized R square for Model 1 was 21%, compared to 16.2% for Model 0. These results indicated that Model 1 fit the data better than Model 0.

Calibration: Model 1 had a smaller calibration χ^2 value than Model 0 (8.4 versus 10). The calibration plot in Figure 3 shows that the predictions from Model 0 departed further from the perfect calibration line than the predictions from Model 1. When the predicted probabilities were less than 0.075, the predictions

Table 2: Comparison of Selected Model versus Refined Model

	Selected Model			Refined Model		
	Coefficient	SE	p-value	Coefficient	SE	p-value
Intercept	-11.2000	0.7762	0.0000	-10.5570	0.4098	0.0000
AGE	0.0509	0.0236	0.0308	0.0268	0.0036	0.0000
AGE^2	-0.0002	0.0002	0.3008			
CP	8.3110	1.0082	0.0000	8.3184	1.0081	0.0000
CP'	-8.2550	1.0152	0.0000	-8.2623	1.0150	0.0000
ADMIT60D	0.7076	0.1259	0.0000	0.7110	0.1258	0.0000
ADMIT60D'	-0.5340	0.1952	0.0062	-0.5366	0.1950	0.0059
MOD_APS	0.0249	0.0345	0.4711	0.0345	0.0131	0.0084
MOD_APS^2	0.0005	0.0017	0.7553			
HRABX	0.1547	0.0205	0.0000	0.1547	0.0205	0.0000
HRABX'	-0.1807	0.0316	0.0000	-0.1805	0.0316	0.0000
ICUPT	0.4625	0.1342	0.0006	0.4688	0.1341	0.0005
LAX	0.2650	0.1239	0.0325	0.2663	0.1239	0.0316
GAS	0.7387	0.1846	0.0001	0.7495	0.1844	0.0000
MOTIL	0.7100	0.1187	0.0000	0.7143	0.1187	0.0000
ALBUMIN	0.5181	0.1206	0.0000	0.5160	0.1204	0.0000
Model LR		745.69			744.49	
Df		15			13	
R^2		0.208			0.208	
C-index		0.879			0.880	
Brier score		0.009			0.009	

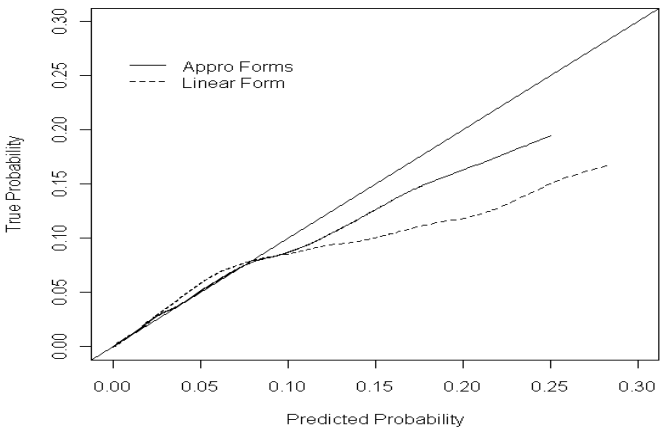

Figure 3: Calibration plot of model with appropriate simple functional forms (solid line) and model with linear terms (dashed line). The diagonal line is for perfect calibration

from Model 0 tended to underestimate the true probabilities, while the predictions from Model 1 were almost perfect within this range. When the predicted probabilities were larger than 0.075, the predictions from both models tended to overestimate the true probabilities. However, the overestimation from Model 0 was much worse than that from Model 1.

Discrimination: Figure 4 displays the ROC curves from Model 1 (solid line) and Model 0 (dashed line). Given the value of 1- specificity, the solid line tended to be higher than the dashed line, indicating that Model 1 separated the patients with CDI from the patients without CDI better than Model 0. In fact, the AUC for Model 1 was 0.879, and for Model 0 was 0.843. The difference was statistically significant. The integrals of sensitivity for Model 1 and Model 0 were 0.0461 and 0.039, respectively. The improvement in the integral of sensitivity for Model 1 over Model 0 was 0.0071 with standard error of 0.0018, yielding a p-value < 0.001. Although the magnitude of the absolute improvement (0.0071) was small, the relative improvement of 28% (0.0071/0.039) was moderate. The integral of specificity for Model 1 and Model 0 was indistinguishable (0.99103 versus 0.99096). The integrated discrimination improvement (IDI) of Model 1 over Model 0 was 0.0072 with standard error of 0.0019, yielding a p-value $<$ 0.001.

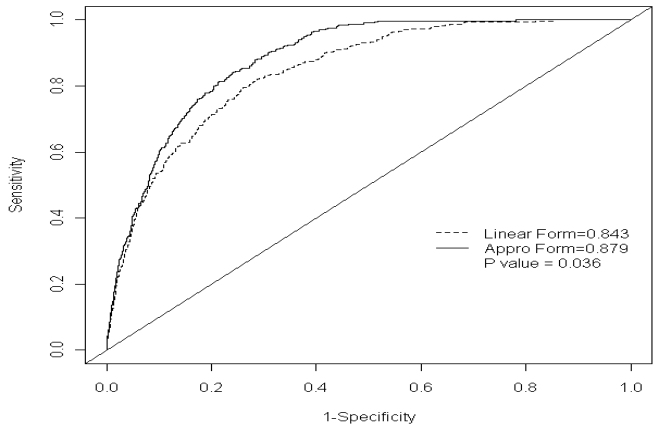

Figure 4: ROC curve and area under the curve (AUC) for the model with appropriate simple parametric forms (solid line) and the model with linear term (dashed line)

5. Discussion and Conclusion

Simple parametric functional forms, if appropriate, are preferred over more complicated functional forms in clinical prediction models [11, 12]. In this paper, we illustrate our practical approach to obtaining the appropriate functional

forms for continuous variables in developing a clinical prediction model for risk of Clostridium difficile infection. First, we used a nonparametric regression smoother to establish the reference curve. Then, we used regression spline function-restricted cubic spline (RCS) and simple parametric forms to approximate the reference curve. Based on the shape of the reference curve, the model fit information (AIC), and the formal statistical test (Vuong test), we selected the simple parametric forms to replace the more elaborated RCS functions. Finally, we refined the simple parametric forms in the multiple variable regression model using the Wald test and the likelihood-ratio test.

Steyerberg *et al.* [13] used an alternative approach to selecting simple parametric forms in developing the clinical prediction rule for testicular cancer. They first fitted RCS function for a continuous variable to obtain a flexible and smooth curve. Based on visual examination and model χ^2 information, they simplified these RCS curves using dichotomization, a linear term, or logarithm and square root transformations. Our work extended their approach by using nonparametric regression smoother as the reference curve and using the more formal test for comparing the superiority between RCS and candidate parametric forms, and by further refining simple parametric forms in the multiple variable regression models.

Using nonparametric regression smoother as the reference curve, we found the RCS functions did not always represent the data better than the simple parametric forms. For AGE, which was well distributed, RCS function tracked the reference curve better than the simple functional form. For HRABX and ADMIT60, however, RCS functions did not track the reference curve as well as the simple functional forms. More formal statistical evidence (AIC and Voung test) also indicated that the simple functional forms fit the data better than the RCS for CP and HRABX. Steyerberg *et al.* also found one of the continuous variables (prechemotherapy HCG) that had a higher model χ^2 with dichotomized functional form than with the RSC form (14 versus 10) [13]. Recognizing that the RCS may not represent the underlying shape of a continuous variable with the outcome in certain situations, we should avoid blindly using the RCS without checking the appropriate reference curves.

When using RCS, we need to specify the number and location of knots. Stone has found that the performance of RCS depends more on the number of knots than on the location of knots in most situations [14]. Harrell has suggested that it is a good approach to place the knots at fixed quantiles of a predictor's marginal distribution [15]. Stone has also found that more than 5 knots are seldom required in RSC [14]. In our RCS modeling, we used 5 knots with equally spaced quantiles to ensure enough data points in each interval as recommended by Harrell [15]. In this article, we use RCS for comparison since it is widely used in medical research,

and can be easily implemented by the R function [10].

To our knowledge, there are few studies presenting the information about the difference in predictive accuracy between the model with the appropriate forms for continuous variables and the model with simple linear terms. In our study, we found that compared to the model with simple linear terms for all continuous variables, the model with appropriate functional forms not only fit the data much better, but it also had higher predictive accuracy. The improvement of predictive accuracy mainly came from the discrimination aspect, especially the sensitivity component. This observation further highlights the importance of appropriate functional forms for continuous variables when developing clinical prediction models.

Acknowledgements

This work was supported by Barnes-Jewish Hospital Foundation grant, 1R21NR011362-01,1K23AI065806-01A2.

References

Altman D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine* **19**, 453-473.

Dubberke, E. R., Reske, K. A., Yan, Y., Olsen, M. A., McDonald, L. C. and Fraser, V. J. (2007). Clostridium difficile−associated disease in a setting of endemicity: identification of novel risk factors. *Clinical Infectious Diseases* **45**, 1543-1549.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1992). Local regression models. In: Chambers, J. M. and Hastie, T. J. (Eds.), *Statistical Models in S*, pp. 309-376. Chapman & Hall, New York.

Harrell, F. E., Jr., Lee, K. L. and Pollock, B. G. (1988). Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute* **80**, 1198-1202.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307-333.

D'Agostino, R. B., Griffith, J. L., Schmidt, C. H. and Terrin, N. (1997). Measures for evaluating model performance. *Proceedings of the Biometrics Section*, 253-258. Alexandria, VA., USA.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845.

Pencina, M. J., D'Agostino, Sr., R. B., D'Agostino, Jr., R. B. and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157-172.

R Development Core Team. (2010). *R*: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Harrell, F. E., Jr. (2009). *Design*: *Design Package.* R package version 2.3-0. http://CRAN.R-project.org/package=Design

Harrell, F. E., Jr., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387.

Steyerberg, E. W. (2009). *Clinical Prediction Models*: *A Practical Approach to Development, Validation, and Updating.* Springer, New York.

Steyerberg, E. W., Vergouwe, Y., Jan Keizer, H. and Habbema, J. D. F. (2001). Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. *Statistics in Medicine* **20**, 3847-3859.

Stone, C. J. (1986). Comment: generalized additive models. *Statistical Science* **1**, 312-314.

Harrell, F. E., Jr. (2001). *Regression Modeling Strategies*: *With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer, New York.

Received April 27, 2011; accepted August 10, 2011.

Yan Yan
Department of Surgery and Division of Biostatistics
Washington University School of Medicine
660 South Euclid Avenue
St. Louis, MO 63110, USA
yany@wudosis.wustl.edu

Kimberly A. Reske
Department of Internal Medicine
Washington University School of Medicine
660 South Euclid Avenue
St. Louis, MO 63110, USA
Kreske@DOM.wustl.edu

Victoria J. Fraser
Department of Internal Medicine
Washington University School of Medicine
660 South Euclid Avenue
St. Louis, MO 63110, USA
VFRASER@DOM.wustl.edu

Graham A. Colditz
Department of Surgery
Washington University School of Medicine
660 South Euclid Avenue
St. Louis, MO 63110, USA
Colditzg@wudosis.wustl.edu

Erik R. Dubberke
Department of Internal Medicine
Washington University School of Medicine
660 South Euclid Avenue
St. Louis, MO 63110, USA
Edubberke@DOM.wustl.edu