Original Article

Dichotomisation of Continuous Data: Review of Methods, Advantages, and Disadvantages

Baneshi MR[1], Talei AR[2]

Abstract

Background: In medical research, dichotomisation of continuous variables is a widespread use approach. However, it has been argued that dichotomisation might be waste of information. The aim of this paper is to review the main methods to dichotomise continuous data, to address practical issues around dichotomisation methods, and to investigate whether dichotomisation is always a bad idea.

Methods: A total of 310 breast cancer patients were recruited. Information on 3 categorical and 1 continuous variable (age at diagnosis) was available. Missing data were imputed applying the Multivariable Imputation via Chained Equations (MICE) method. Then a minimum P-value method was applied to dichotomise the age variable. The Cox regression model was fitted to develop models in which dichotomised versus continuous version of the age variable plus other 3 variables were used. Results were compared in terms of discrimination ability, goodness of fit, and classification improvement.

Results: For the age variable, an optimal split at 47 was found. This split was close to menopause age of women in Shiraz (48) so had biological interpretability. The stability of optimal split was confirmed in bootstrap study. Model in which dichotomised version of age was used showed higher discrimination ability and goodness of fit. Furthermore, dichotomised model assigned 14% of live patients into a more appropriate risk group.

Discussion: Dichotomisation of continuous data is a contentious issue. We have shown that dichotomisation might improve performance of models when it has biological interpretation. More research is needed to understand situations in which dichotomisation might work.

Keywords: Dichotomisation; Breast neoplasm; Minimum P-value; Missing data; Reclassification

Please cite this article as: Baneshi MR, Talei AR. Dichotomisation of Continuous Data: Review of Methods, Advantages, and Disadvantages. Iran J Cancer Prev.2011; Vol4, No1, P.26-32.

1. Health School, Kerman University of Medical Sciences, Department of Biostatistics and Epidemiology, Kerman, Iran
2. Shahid Faghihi Hospital, Shiraz University of Medical Sciences, Shiraz, Iran

Corresponding Author:
Mohammad Reza Baneshi; PhD in Medical Statistics
Tel: (+98) 341 320 50 87
Email: m_baneshi@kmu.ac.ir

Received: 29 Aug. 2010
Accepted: 4 Dec. 2010
Iran J Cancer Prev 2011; 1:26-32

Introduction

Cancer is one of the most major health problems worldwide. Management of patients and treatment selection process is guided using prognostic models. For example, Nottingham Prognostic Index (NPI) was devised to guide the treatment of breast cancer patients [1]. This model has been widely validated and now is one of the central tools in risk prediction [2-4].

In medical applications, to develop prognostic models, researchers often dichotomise continuous covariates prior to modelling analyses. From a statistical point of view, dichotomisation eliminates the need for the linearity assumption, makes data summarisation more efficient, and allows for simple interpretation of results [5]. In the regression setting, for instance, interpretation of the impact of a binary covariate on outcome is easier than that for a change of 1 unit in a continuous covariate. Furthermore, it has been claimed that, from the clinical point of view, binary covariates might be preferred because they offer a simple risk classification into high versus low, assist in making treatment recommendations, and in setting diagnostic criteria [5, 8].

On the other hand, dichotomisation can result in the loss of information and power, if a linear rather than threshold association pertains, and non-linear relationships such as U-shape associations will not be detected [9, 10].

It has been emphasized that dichotomisation is appropriate only when a threshold effect value truly exists. That is, if we can assume some binary split of the continuous covariate creates two relatively distinct but homogeneous groups with respect to a particular outcome [11].

The aim of this paper is to develop prognostic models using dichotomised versus continuous variables, and to address loss or gain in model performance due to dichotomisation. Methods applied analysing a breast cancer data set as an example.

Materials and Methods

Patients and outcome

From 1994 to 2003, the information of 310 breast cancer patients in Shiraz, southern Iran were collected from Hospital-based Cancer Registry of Motahhari Para clinic affiliated to Shiraz University of Medical Sciences. Median follow-up time was 2.5 years. Survival was considered as the time period between diagnosis and death (or last visit) of patient. At the end of the study, there had been 56 deaths.

Variables

Variables offered to the multifactorial models were those showed to have univariate predictive ability [12]: tumour stage with 3 levels (early, locally advanced, and advanced), tumour grade with 3 levels (1, 2, and 3), history of benign breast disease (positive versus negative), and age at diagnosis.

Imputation of missing data

Not all patients had available data on all 4 variables under study. To avoid attrition in sample size, Multivariable Imputation via Chained Equations (MICE) method was applied to impute missing data [13]. The MICE method replaces each missing value by multiple imputed values, resulting in multiply imputed data sets [14]. Technical details of the MICE method is illustrated elsewhere [15].

Continuous Model

The only continuous variable was age. Keeping this variable in the continuous form, all four variables were offered to the multifactorial model. Multivariable Fractional Polynomial (MFP) modelling was applied to develop the multifactorial regression model and to identify the appropriate (possibly non-linear) form of association between age and outcome [16]. The MFP modelling checks whether power transformation is required in the multifactorial model. The MFP, after fitting of linear factors, ascertains whether model fit would be improved by using a polynomial form for any of the linear variables.

Dichotomised Model

To dichotomise the age variable the minimum P-value procedure applied, as explained below. Dichotomised version of age variable and other 3 categorical variables were then offered to the multifactorial model.

In the minimum P-value approach, after a systematic search across all possible values, the value chosen as the cut point was that with the smallest corresponding P-value in a Log-Rank test, when comparing the survival curve of two groups formed [17]. To avoid groups with very small/ high number of patients, no split at the outer 20% of the covariate distribution was applied [17, 18]. In addition, to take into account the multiple testing undertaken, cut point P-value was adjusted to reach a decision regarding whether or not to adopt the cut point. If ε_{low} and ε_{high} show the proportion of the observations at the bottom and top of the highest cut point value considered (0.10 in our application), derivation below was applied [18]. Cut point selected was adopted only if the adjusted P-value reached significance level at 5% level [18].

$$P_{adj} = -1.63 P_{\min}(1 + 2.35 Ln(P_{\min}))$$

Applying this correction formula, an unadjusted P-value of 0.002 corresponds to 0.05 if one single test applies.

Stability of cut point selected was checked using graphical (minimum P-value graph) ad numerical methods (bootstrap re-sampling) [8]. A minimum P-value graph plots all cut point values of covariate against corresponding P-values to assess whether any other cut off(s) exists with P-value similar to that of minimum P-value [19]. To reduce the instability, the median of optimum splits across bootstrap samples will be used as the split. In the case that competing cut offs are far from each other, the modal statistic will be used (instead of media) [19].

Comparison of performance of models

Models developed were compared in terms of discrimination ability, goodness of fit, and reclassification improvement.

Discrimination refers to the ability to separate patients with different responses [20] and is measured using Harrell's C-index which is a generalisation of the Area Under Curve (AUC). This statistic varies between 0.5 and 1 where values near 1 indicate high discrimination power.

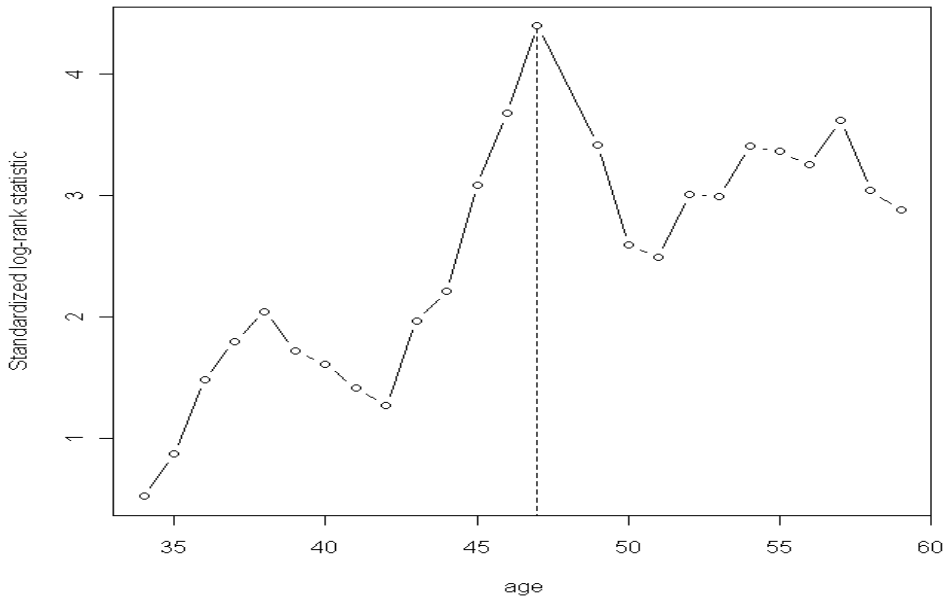

Figure 1. Minimum P-value graph for the Age variable showing optimal split

For all models, the Likelihood Ratio Test (LRT) which indicates how well the model fits the data is reported.

For both models developed, patients were classified into three risk groups. To do so, a risk score was calculated for each model. This was done multiplying the estimated regression coefficients into the variables. The tertiles of risk scores derived were applied as cut off. Net Reclassification Index (NR Index) was used to compare risk group assignment across different models [21]. This method considers the joint distribution of patients into risk groups, by the two risk grouping schemes being compared. This statistics quantifies the 'correct' movement in the risk group classifications i.e. upwards for patients who did experience the event and downwards for patients who did not [21]. Net gain in cases with event has been defined as the difference in proportion of subjects who moved into a higher or lower risk group. The reverse calculations will be made for event free cases. The NR Index is defined as summation of net gains [21].

Results

The numbers (percentages) of patients with missing value on node status, grade, and history of benign disease were 63 (20.3), 64 (20.6), and 47 (15.2) respectively. In total, out of 310 patients, 203 cases (65%) had data available on all 4 variables of which 54 had died. Applying the MICE method, missing data were imputed 10 times. However, for the purpose of this paper, only first imputed data set was analysed.

We first developed the continuous model. Applying MFP model, we found that a linear risk function was adequate to capture effect of the age variable. Therefore this variable contributed to the multifactorial model using a linear risk function.

We then calculated the optimal split for the age variable. The figure derives was 47 (corrected P-value = 0.0002). To explore the underlying hazard structure of age variable, plotting the tested cut points versus split statistics; minimum P-value graph was plotted (Figure 1). No serious competition split was found. Stability of this split (i.e. 47) was confirmed in the bootstrap study. For subsequent checking of stability, 100 bootstrap samples were drawn. Investigation of the threshold effect of age found that in about 80% of bootstrap samples the optimal split was around 47 (ranged 44 to 52). Additionally, the median and mode of selected optimal thresholds was 47. Therefore, in the dichotomised model, the age variable was dichotomised at 47 and then offered to the multifactorial modelling.

The area under the ROC curve corresponding to dichotomise and continuous models were 79% and 77% respectively, indicating two percentage points improvement in discrimination ability of model after dichotomisation of age variable at 47. Furthermore, offering dichotomised version of age variable to the multifactorial model led to a slight improvement in

Table 1. Risk group assignment by continuous and the dichotomised models

Continuous model	Status	Risk group	Dichotomised model		
			L	I	H
	Alive	L	80	19	0
		I	40	45	2
		H	0	14	53
	Died	L	3	2	0
		I	3	8	3
		H	0	2	35

L: Low risk group I: Intermediate risk group H: High risk group
Numbers show frequency of patients into risk groups based on two models developed

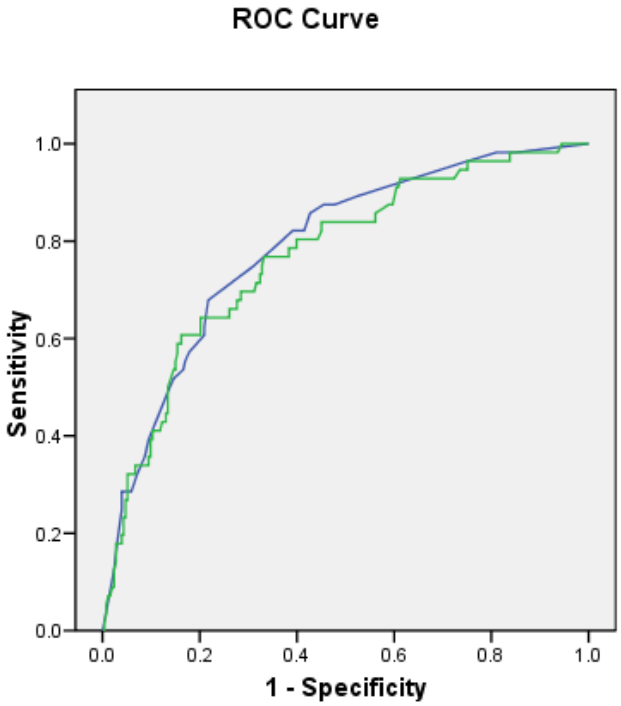

Figure 2. ROC curve analysis comparing discrimination ability of dichotomised (blue line) versus continuous model (green line)

the model's goodness of fit. The LRT statistics corresponding to these two models were 45.7 and 41.8 respectively.

We finally checked distribution of patients into low, intermediate, and high risk group, for dichotomised model relative to continuous model (Table 1). The proportion of alive patients that moved into more appropriate and less appropriate risk groups were 22% ([40+0+14]/253) and 8%, respectively ([19+0+2]/253). The corresponding figures for died patients were 9% ([2+0+3]/56) and 9% ([3+0+2]/56).

The net gain in the proportion of patients that were reclassified was 14% (22% - 8%) for those that alive and 0% (9% - 9%) for those that died, giving an NRI at 14% (14% + 0%).

Discussion

Dichotomisation of continuous data is a debatable issue. It has been argued that it is statistically non-intuitive, although it makes the presentation and interpretation of results easier. However, dichotomisation might work if it creates groups with similar biological characteristics.

It should be emphasized that the main goal of this study was to demonstrate the application of statistical methods to dichotomise a continuous variable, to highlight advantages and disadvantages of dichotomisation, and to address impact of dichotomisation on model performance. To achieve our goals, we simply used a breast cancer data set as an example. Therefore discussion of risk factors of breast cancer and their contribution to the disease course is beyond the scope of this paper. This issue has been addressed here [22, 23].

Before any modelling practice, we imputed missing data to avoid biased estimates [24]. In our data set, the optimal split derived was 47. Furthermore, it has been shown that menopause age of women in Shiraz is about 48. In other words, the dichotomised age variable was a surrogate for approximate menopausal status. The split applied created two homogenous subgroup of patients as menopause status is liked to activity Estrogen and Progesterone receptors. These receptors promote the growth of cancer cells, are present in nearly two thirds of breast cancer specimens, and can affect patient's outcome [25].

We have found that, offering dichotomised version of age variable to the model –instead of continuous version- led to two percentage point improvement in discrimination ability of model (i.e. C-index).

It should be noted that although the C-index is the most frequently used criteria in the literature [26,27], this statistics is not sensitive to assess the benefit of addition of a new risk factor to a set of standard risk factors, or to compare impact of change in the form of risk function (i.e. continuous versus dichotomised) [28-30]. As an example, the C-index has been used to assess the incremental coronary risk prediction using C- reactive protein and other novel risk factors. C-index for the basic model was 80%. This model was included age, race, sec, cholesterol level, high-density lipoprotein cholesterol level, systolic blood pressure, antihypertensive medications, smoking status, and diabetes. Although most of novel markers showed age-adjusted significant association predict CHD, out of 19 variables studied only 4 of them added the most to the AUC ranged from 0.006 to 0.011 [31]. In another study, the extent to which inclusion of seven single nucleotide polymorphisms (SNP) improves assessment of breast cancer risk was assessed [32]. AUC of new models were compared with that of the National Cancer Institute's Breast Cancer Risk Assessment Tool (BCRAT), which is based on ages at menarche and at first live birth, family history of breast cancer, and history of breast biopsy examinations. Only two percentage point increase was seen in AUC (0.61 versus 0.63). It has been shown that addition of a new risk factor with an odds ratio as large as 3, to a set of prognostic factors, may have little impact on C-index [28]. Therefore, we believe two percentage points improvement, as a consequence of using a different form of risk function (i.e. dichotomised versus continuous), might be remarkable.

When biological knowledge cannot guide selection of an appropriate split, alternative approaches (i.e. media or optimal split methods) can be used. Although dichotomisation based on biological evidence is attractive, for the majority of variables the biological knowledge needed is not available.

Another method commonly used is to categorise the covariate at a pre-determined split such as the median [33]. In this way an equal proportion of patients (50%) are assigned to each group. It should be added that dichotomisation at median leads to different threshold values from one study to another, and creates difficulties in comparing findings across different studies [10]. As an example, in a meta analysis of eleven studies on the role of cathespin D on Disease Free Survival (DFS) of breast cancer patients, the cut points used to define high/low cathespin D concentration ranged from 20 to 78 [34]. Therefore this approach might not be useful in practice. In our data set, the median of age variable was 46 which were very close to optimal split. Therefore, no separate model was developed.

In majority of situations, there is no biologic evidence or priori information regarding the underlying relationship between the covariate and the outcome. In such situations, it is possible to seek the cut point which gives us the largest difference between individual outcomes in the resulting two groups [35]. It should be emphasized that multiple testing is a regrettable consequence of minimum P-value method. To reflect the multiple testing undertake, we adjusted the P-value estimated. Furthermore, we checked the stability of optimal split selected through graphical (minimum P-value graph) and numerical methods (bootstrap study). The stability of optimal split derived was confirmed.

Use of minimum P-value method might result in a type one error as high as 40% [36]. This rate might be inflated to 50% if examining 50 cut points [37]. To show the danger of application of minimum P-value method without P-value adjustment, a series of 686 node-positive breast cancer patients was divided into two equally sized samples (training and test samples). A Minimum P-value method was applied to the training sample. An optimal split for

age at 43 years old was proposed in the training set (unadjusted P-value= 0.02). On the other hand, if the adjusted P-value had been calculated for the training sample, it would have been 0.31, far from significant and preventing a misleading impression of association with age younger or older than 43. Applying this cut point to the test sample gave a P-value of 0.23. Furthermore, application of this split to another independent sample (n=139) resulted in P-value of 0.38 [38].

Two-fold cross-validation and sample-split techniques are two alternative approaches to deal with multiple testing [39]. In a two-fold cross-validation approach the data is divided into two equally sized subsets. Minimum P-value method is applied separately in each subset to find the optimal cut points (say C1 for first subset, C2 for second subset). Cut points derived are applied to the other subset. The subgroups of patients with low values for the covariate is a combination of the below cut point patients in each subset. High risk patients are defined in a similar way. The P-value of the covariate is estimated using a Log-Rank or Cox model [40]. Simulation studies show that the type one error for this method is approximately correct [41].

In the sample-split method, the data will be divided into training and test samples. The optimal cut point derived in the training set will be applied in the test set to find the correct P-value. It might be that neither two-fold cross-validation nor two-sample statistics are feasible when sample size and number of events is low.

In a similar study, Royston et al. assessed the issue of loss or gain in model performance due to dichotomisation of continuous data. This has been illustrated in an analysis of 207 patients with primary biliary cirrhosis [42]. The association between 2 continuous (age and logarithm of bilirubin) and 2 binary variables (central cholestasis and cirrhosis), and treatment was evaluated. The two variables were dichotomised with both optimal cut point and at median. Different multifactorial models were developed in which continuous variables were modelled in continuous and binary form. The model in which continuous data were treated as being continuous had highest discrimination ability and model goodness of fit [42]. However, it has not been argued that whether splits applied had biological interpretation. In our data set optimal split applied was fairly close to menopause age. This might explain differences seen between our results and the Royston's study [42].

We believe that dichotomisation might improve performance of prognostic models when it creates groups with similar biological features. There is room in science for trying several approaches with a given data set and reviewing the results critically' [43]. Comparison of results enriches the body of the literature and enhances the understanding of the situations in which dichotomization of continuous data improves performance of prognostic models.

Acknowledgment

We should thank staff of Motahhari Para clinic and Shahid Faghihi hospital who facilitated our access to patients' folder and information.

Conflict of Interest

There was no conflict of interest.

Authors' Contribution

The data set analyzed in this project was collected under the direction of Prof.TAR at Shiraz University of Medical Sciences. All analyses and writing of manuscript has been done by BMR.

References

1. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. Br J Cancer 1982 Mar; 45(3):361-6.

2. Todd JH, Dowle C, Williams MR, Elston CW, Ellis IO, Hinton CP, et al. Confirmation of a prognostic index in primary breast cancer. Br J Cancer 1987 Oct; 56(4):489-92.

3. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat 1992; 22(3):207-19.

4. Balslev I, Axelsson CK, Zedeler K, Rasmussen BB, Carstensen B, Mouridsen HT. The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). Breast Cancer Res Treat 1994; 32(3):281-90.

5. Williams BA, Mandrekar JN, Mandrekar SJ, Cha SS, Furth AF. Finding optimal cutpoints for continuous covariates with binary and time-to-event outcomes. 2006.

6. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer-Verlag .; 2000.

7. Harrell FE. Regression modelling strategies with application to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.

8. Mazumdar M, Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. Stat Med 2000 Jan 15; 19(1):113-32.

9. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006 May 6; 332(7549):1080.

10. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods 2002 Mar; 7(1):19-40.

11. Abdolell M, LeBlanc M, Stephens D, Harrison RV. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Stat Med 2002 Nov 30; 21(22):3395-409.

12. Rajaeefard AR, Baneshi MR, Talei AR, Mehrabani D. Survival Models in Breast Cancer. Iranian Red Crescent Medical Journal 2009; 11(3):295-300.

13. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values were preferred. J Clin Epidemiol 2006 Oct; 59(10):1092-101.

14. Schafer JL. Analysis of Incomplete Multivariate Data. Florida: Chapman and Hall; 1997.

15. Baneshi MR. Statistical Models in Prognostic Modelling of Many Skewed Variables and Missing Data: A Case Study in Breast Cancer (PhD thesis submitted at Edinburgh University) 2009.

16. Royston P, Sauerbrei W. Multivariable Model Building a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: John Wiley; 2008.

17. Lausen B, Schumacher M. Maximally selected rank statistics. Biometrics 1992; 48:73-85.

18. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 1994 Jun 1; 86(11):829-35.

19. Dannegger F. Tree stability diagnostics and some remedies for instability. Stat Med 2000 Feb 29; 19(4):475-91.

20. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999 Mar 16; 130(6):515-24.

21. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008 Jan 30;27(2):157-72.

22. Baneshi MR, Warner P, Anderson N, Bartlett JSM. Tamoxifen resistance in early breast cancer: statistical modelling of tissue markers to improve risk prediction. British Journal of Cancer 2010; 102:1503-10.

23. Baneshi MR, Warner P, Anderson N, Tovey S, Edwards J, Bartlett JM. Can biomarkers improve ability of NPI in risk prediction? A decision tree model analysis. Iranian Journal of Cancer Prevention 2010; 2:62-74.

24. Baneshi MR, Talei AR. Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. Iranian Journal of Cancer Prevention 2010; 3(3):127-31.

25. Martin M. Molecular biology of breast cancer. Clin Transl Oncol 2006 Jan; 8(1):7-14.

26. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr; 143(1):29-36.

27. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004 Jul 15; 23(13):2109-23.

28. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004 May 1; 159(9):882-90.

29. Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. Arch Intern Med 2005 Nov 28; 165(21):2454-6.

30. Ware JH. The limitations of risk factors as prognostic tools. N Engl J Med 2006 Dec 21; 355(25):2615-7.

31. Folsom AR, Chambless LE, Ballantyne CM, Coresh J, Heiss G, Wu KK, et al. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study. Arch Intern Med 2006 Jul 10; 166(13):1368-73.

32. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst 2008 Jul 16; 100(14):1037-41.

33. Linderholm B, Grankvist K, Wilking N, Johansson M, Tavelin B, Henriksson R. Correlation of vascular endothelial growth factor content with recurrences, survival, and first relapse site in primary node-positive breast carcinoma after adjuvant treatment. J Clin Oncol 2000 Apr; 18(7):1423-31.

34. Ferrandina G, Scambia G, Bardelli F, Benedetti PP, Mancuso S, Messori A. Relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients: a meta-analysis. Br J Cancer 1997; 76(5):661-6.

35. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer-Verlag; 2003.

36. Altman DG. Suboptimal analysis using 'optimal' cutpoints. Br J Cancer 1998 Aug; 78(4):556-7.

37. Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out? Breast Cancer Res Treat 1992; 22(3):197-206.

38. Hollander N, Schumacher M. On the problem of using 'optimal' cutpoints in the assessment of quantitative prognostic factors. Onkologie 2001 Apr; 24(2):194-9.

39. Hilsenbeck SG, Clark GM. Practical p-value adjustment for optimally selected cutpoints. Stat Med 1996 Jan 15; 15(1):103-12.

40. Mazumdar M, Smith A, Bacik J. Methods for categorizing a prognostic variable in a multivariable setting. Stat Med 2003 Feb 28; 22(4):559-71.

41. Faraggi D, Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. Stat Med 1996 Oct 30; 15(20):2203-13.

42. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regressions: a bad idea. Stat Med 2006 Jan 15; 25(1):127-41.

43. Royston P, Sauerbrei W, Altman DG. Modeling the effects of continuous risk factors. J Clin Epidemiol 2000 Feb; 53(2):219-21.