RESEARCH ARTICLE

Diagnostic Classification Scheme in Iranian Breast Cancer Patients using a Decision Tree

Amal Saki Malehi

Abstract

Background: The objective of this study was to determine a diagnostic classification scheme using a decision tree based model. Materials and Methods: The study was conducted as a retrospective case-control study in Imam Khomeini hospital in Tehran during 2001 to 2009. Data, including demographic and clinical-pathological characteristics, were uniformly collected from 624 females, 312 of them were referred with positive diagnosis of breast cancer (cases) and 312 healthy women (controls). The decision tree was implemented to develop a diagnostic classification scheme using CART 6.0 Software. The AUC (area under curve), was measured as the overall performance of diagnostic classification of the decision tree. Results: Five variables as main risk factors of breast cancer and six subgroups as high risk were identified. The results indicated that increasing age, low age at menarche, single and divorced statues, irregular menarche pattern and family history of breast cancer are the important diagnostic factors in Iranian breast cancer patients. The sensitivity and specificity of the analysis were 66% and 86.9% respectively. The high AUC (0.82) also showed an excellent classification and diagnostic performance of the model. Conclusions: Decision tree based model appears to be suitable for identifying risk factors and high or low risk subgroups. It can also assists clinicians in making a decision, since it can identify underlying prognostic relationships and understanding the model is very explicit.

Keywords: Breast cancer - decision tree - risk factor - Iran

Asian Pac J Cancer Prev, **15 (14)**, 5593-5596

Introduction

Breast cancer is the most common in women in the global burden of cancers (Mamoon et al., 2009; Sadjadi et al., 2009a; Anaya-Ruiz et al., 2014; Keramatinia et al., 2014; Slaoui et al., 2014; Xing et al., 2014). It is stands after lung cancer in developed countries, newly diagnosed in more than 1 million women each year (Sadjadi et al., 2009a; Anaya-Ruiz et al., 2014). In addition mortality rate of the breast cancer has preceded the lung cancer in women (Emami-Razavi et al., 2009; Hosseini and Fakhraee, 2009). The available studies in Iran have showed that the breast cancer attacks Iranian women at least one decade younger than women in developed countries, with the mean age ranging from 47.1 to 48.8 years (Mahouri et al., 2007b; Alireza-Sadjadi et al., 2009b; Sadjadi et al., 2009a). It is estimated that include 21.4% of all malignancies in female and also ranked as the first cancer in Iranian women (Emami-Razavi et al., 2009; Sadjadi et al., 2009a).

So it causes to major focus of attention within public health care systems. However it is a preventable cancer if early was detected (Emami-Razavi et al., 2009). Development of diagnostic classification schemes may provide the improvement in the clinical management of early detection of breast cancer by identifying target women who are at higher risk. It would therefore be useful to assess the risk factors of this disease.

Tree based models that known as decision tree seems to be suited for this role. Tree based models have become one of the most flexible and powerful data analytic tools and applications of these methods are far reaching (Breiman et al., 1894). The best documented and most popular uses of tree based models are in biomedical research, which classification is a central issue (Banerjee and Noone, 2008). Decision tree based models are partitioning procedures; the goal of these methods is to derive a model that predicts the category of a particular individual based on one or more explanatory factors (Spitz et al., 2007). The simplicity of interpretation of results in terms of clinical or other relevant patient characteristics make decision tree an appealing approach in both clinical and epidemiologic investigations (Barnholtz-Sloan et al., 2011; Shen et al., 2012). The aim of the present study was to develop a diagnostic classification scheme by adopting a modeling approach for breast cancerthat is able to identify prognostic relationships underlying data and avoiding restrictive assumptions of conventional modeling approaches.

*Department of Biostatistics and Epidemiology, School of Public Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran *For correspondence: amalsaki@gmail.com*

Materials and Methods

The study was conducted as a retrospective of case-control study of 312 cases and 312 controls that referred to Imam Khomeini hospital in Tehran during 2001 to 2009. Patients who had positive result of pathological diagnostic of breast cancer were included in case group and patients who referred to hospital without any history of breast problems or neoplastic disease were comprised in control group. Women with hysterectomy and artificial menopause were excluded from the study. The two groups matched in term of demographic and socioeconomic status. Demographic and clinical-pathological information were collected from medical record and interview with patients.

Decision tree was implemented to develop diagnostic classification scheme using CART 6.0. This method determines the contrast effects of risk factors and constructs subgroups of patients based on demographic and clinical symptoms by recursive partitioning the subjects.

Performing the decision tree using CART contains 4 steps: (Breiman et al., 1894) 1) Growing tree (over fitting data); 2) Pruning the over fit tree; 3) Select best subtree of pruning tree that exhibit best possible structure of data; 4) Statistical summaries for terminal nodes of selected tree.

Terminal nodes constitute interested subgroups in term of concern outcome. In the CART each tree's structure depended on the initial split of the patients. A default tree was generated by allowing the CART program to determine the important variable with optimal first split. Also selecting other variables for partitioning procedure is based on importance score of each variable. The measure of importance is the sum over all nodes of the decrease in impurity produced by the best split on x_m at each node.

To evaluate and verify the validation of the diagnostic classification, the CART analysis was performed using cross validation with 10 fold (Barnholtz-Sloan et al., 2011).

The AUC (area under curve), can also be interpreted as a measure of the overall performance of diagnostic classification of the decision tree (Spitz et al., 2007; Barnholtz-Sloan et al., 2011). Computing AUC, the area under curve for receiver operating characteristic (ROC) is based on probability of correct classification of patients.

Results

A total 624 patients included in this study, only 312 patients of them were admitted with diagnostic of breast cancer and 75.64% of them were 17-40 years old. The Mean±SD age of case and control was 36.9±4.98 and 32.9±4.3 respectively.

Decision tree was performed with baseline characteristics variables as depicted in Table 1. The baseline characteristics of patients were described in Table 2.

The diagram of tree structure was presented in Figure 1. The decision tree had an initial split on age, then age of menarche was selected with importance score 39.54 and family history of breast cancer, marriage status, menstrual pattern were selected with importance score 29.72, 10.73, 5.89 respectively. These variables were identified as risk factors and had been determined the tree structure. Nine terminal nodes as classification subgroups were formed the structure of decision tree which six of them introduced as high-risk subgroups. This classification identified the following variables as risk factors of breast cancer: 1) Age more than 39.5 years old; 2) Low age of menarche (≤12 years old); 3) History of family breast cancer; 4) Single or divorced status; 5) Irregular menarche pattern.

The incidence of cancer detection in groups with low age of menarche (≤12) and have history familial breast cancer is 71.3%, 86.4% respectively. This result indicated that these are key risk factors in the Iranian breast cancer patients.

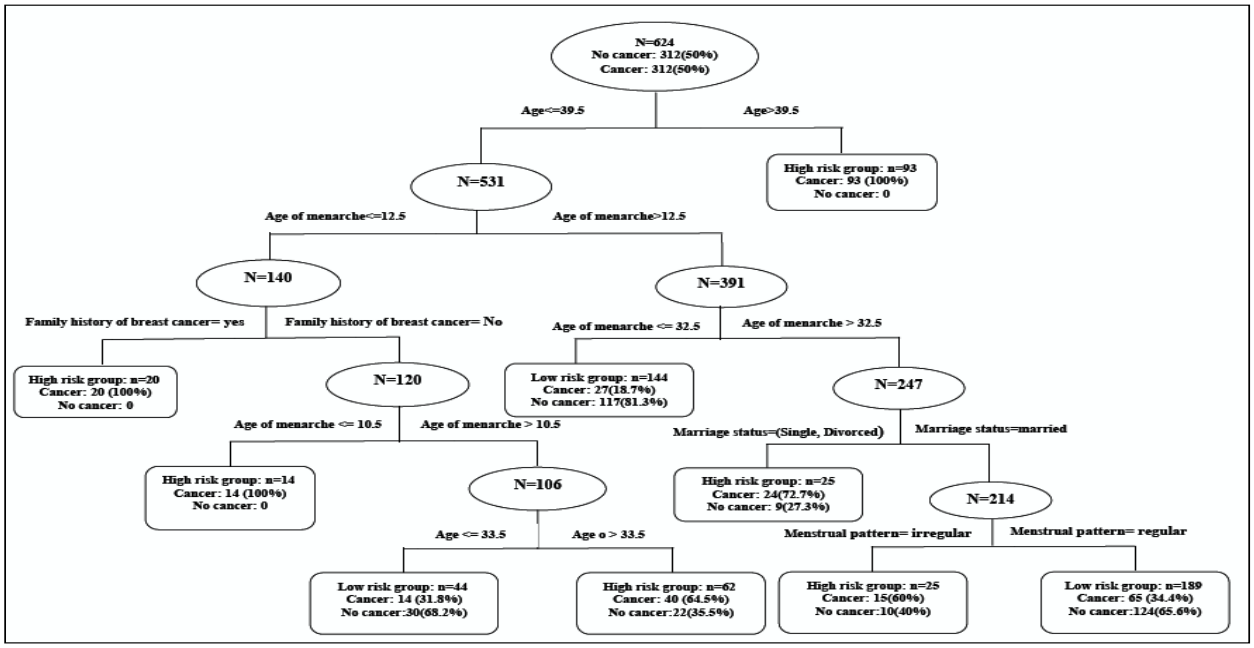

Figure 1. Decision Tree of Women Referred in Imam Khomeini Hospital in Tehran

Table 1. Variable Included in CART Analysis

Age	Family history of ovarian cancer
History of benign breast diseases	Age of menarche
Family history of breast cancer	Menstrual pattern
History use of alcohol	Marriage status
History of smoking	

Table 2. Characteristic of 624 Subjects in Case and Control Group

Parameter		Case (%)	Control (%)
Marriage status	Single	243 (77.9)	280 (89.7)
	Divorced	7 (2.2)	3 (1)
	Married	62 (19.9)	29 (9.3)
Menstrual pattern	Irregular	60 (79.3)	285 (91.3)
	Regular	230 (20.7)	27 (8.7)
Family history of ovarian	No	306 (98)	312 (100)
	Yes	6 (2)	0
Family history of Breast Cancer	No	274 (88)	306 (98)
	Yes	38 (12)	6 (2)
History use of alcohol	No	310 (99.4)	312 (100)
	Yes	2 (.6)	0
History of smoking	No	311 (99.7)	312 (100)
	Yes	1 (.3)	0
History of benign breast disease	No	300 (96.2)	311 (99.7)
	Yes	12 (3.8)	1 (0.3)

Our model showed an acceptable accuracy 76.4% (477 of 624), it means that the model was 76.4% accurate in discriminating patients with Breast cancer from healthy controls.

The AUC analysis also showed an excellent classification and diagnostic performance for decision tree that was 0.82.

Discussion

As regards the breast cancer is a major cancer in women and it is the leading cause of cancer mortality (Emami-Razavi et al., 2009; Hosseini and Fakhraee, 2009; Mamoon et al., 2009; Sadjadi et al., 2009a), it requires a focus of attention for public health authorities and policy-makers. However early diagnosis of Breast cancer reduces mortality rate and improves long-term survival. Therefore, it will be needed to improve early detection of women who suspected to have breast cancer and those at high risk of the disease. We suggested decision tree model to derive a diagnostic classification scheme and classify low and high-risk subgroups. It can also investigate the risk factors by exploring associations of patient characteristics in the data set and identify high-risk subgroups. It seems that performing decision tree based models have been highlighted for predicting outcomes in cancer patients (Cong and Tsokos, 2010; Barnholtz-Sloan et al., 2011; Shen et al., 2012). Our model that have derived from a case–control study, construct six high-risk subgroup and provided evidence that age, family history of the disease, marital status, menstrual pattern and age of menarche influence Breast cancer risk. Moreover, It should be noted that based on associations between these factors, low and high-risk subgroups were constructed. The acceptable accuracy and high AUC of our decision tree–based model showed a good performance of the model in classifying the low and high-risk subgroups.

Increasing age is one of the common effective factors that its role proved in many of cancers (Washbrook, 2006; Amin et al., 2009). It was introduced as the primary risk factor in this study.

A significant proportion of Breast cancer patients had a family history of cancer and early age at menarche in this study. A number of other studies (Washbrook, 2006; Mahouri et al., 2007a; Amin et al., 2009; Kruk, 2009) have shown that women with a family history of breast cancer and early age at menarche are at increased risk of the disease. It was found that reduction in risk were associated with late age at menarche. But there is not exact in cut-off for age at menarche and various of cut-offs used in these studies.

Furthermore, marital status as other risk factors recognized in decision tree model. It's according to many studies; they have demonstrated that single and divorced women have significantly higher risk of breast cancer (Mahouri et al., 2007a; Datta and Biswas, 2009; Parsa and Parsa 2009). But other studies had reported that marital status had no effect on risk of breast cancer (Al-Shaibani et al., 2006; Amin et al., 2009).

Moreover, menstrual cycle was associated with risk of Breast cancer in our study, it was considered in a few studies and they have established that it influenced risk of breast cancer (Parsa and Parsa, 2009). In this study we support using the decision tree based model as a diagnostic classifier for recognizing high-risk subgroups based on important risk factors. However our result regarding to risk factors in agreement with previous studies, but none of them could to create a diagnostic classification scheme. The present study based on limited data set, but if one can implement decision tree on population based data, it will be referral diagnostic scheme.

In conclusion, the decision tree based model appears to be suitable for identifying the risk factors and high or low risk subgroups. It can also assists clinician in making a decision and prognostic inference for future clinical trials, since understanding this model is very explicit and need not a statistical experience.

Acknowledgements

The valuable contribution of Cancer Institute of the Imam Khomeini Hospital in this study is greatly appreciated.

References

Al-Shaibani H, Bu-Alayyan S, Habiba S, et al (2006). Risk factors of breast cancerin Kuwait: case-control study. *Iranian J Med Sci*, **31**, 61-4.

Amin TT, Al Mulhim AR, Al Meqihwi A (2009). Breast cancer knowledge, risk factors and screening among adult Saudi women in a primary health care setting. *Asian Pac J Cancer Prev*, **10**, 133-8.

Anaya-Ruiz M, Vallejo-Ruiz V, Flores-Mendoza L, Perez-Santos M (2014). Female breast cancer incidence and mortality in Mexico, 2000-2010. *Asian Pac J Cancer Prev*, **15**, 1477-9.

Banerjee M, Noone AM (2008). Tree-based methods for survival data, in A Biswas (ed.), Advances in the BiomedicalSciences (New Jersey: John Wiley& Sons, Inc), 265-85.

Barnholtz-Sloan JS, Guan X, Zeigler-Johnson C, Meropol NJ, Rebbeck TR (2011). Decision tree-based modeling of androgen pathway genes and prostate cancer risk. *Cancer Epidem Biomar Prev*, **20**, 1146-55.

Breiman L, Friedman JH, OlShen RA, Stone CJ (1894). Classification and regression trees. A Division of Wadsworth Inc, California:, USA.

Cong C, Tsokos CP (2010). Parametric and nonparametric analysis of breast cancer treatments. *Int J Biol Life Sci*, **6**, 134-7.

Datta K, Biswas J (2009). Influence of dietary habits, physical activity and affluence factors on breast cancer in East India: a case-control study. *Asian Pac J Cancer Prev*, **10**, 219-22.

Emami-Razavi SH, Aaghajani H, Haghazali M, et al (2009). The most common cancers in Iranian women. *Iranian J Public Health*, **38**, 109-12.

Hosseini N, Fakhraee R (2009). The association between abortion and risk of breast cancer in Iranian women.*Navid No*, **40**, 72-7.

Keramatinia A, Mousavi-Jarrahi SH, Hiteh M, Mosavi-Jarrahi A (2014). Trends in incidence of breast cancer among women under 40 in Asia. *Asian Pac J Cancer Prev*, **15**, 1387-90.

Kruk J (2009). Lifetime occupational physical activity and the risk of breast cancer: a case-control study. *Asian Pac J Cancer Prev*, **10**, 443-48.

Mahouri K, Dehghani Zahedani M, Zare S (2007). Breast cancer risk factors in south of Islamic Republic of Iran: a case-control study. *East Med Health J*, **13**, 1265-73.

Mahouri K, Zahedani M, Dehghani, Zare S (2007). Breast cancer risk factors in South of Islamic Republic of Iran: a case-control study. *East Med Health J*, **13**, 1265-73.

Mamoon N, Hassan U, Mushtaq S (2009). Breast carcinoma in young women aged 30 or less in northern Pakistan - the armed forces institute of pathology experience. *Asian Pac J Cancer Prev*, **10**, 1079-82.

Parsa P, Parsa B (2009). Effects of reproductive factors on risk of breast cancer: a literature review. *Asian Pac J Cancer Prev*, **10**, 545-50.

Sadjadi A, Nouraie M, Ghorbani A, et al (2009). Epidemiology of breast cancer in the Islamic Republic of Iran: first results from a population-based cancer registry. *East Med Health J*, **15**, 1426-31.

Sadjadi A1, Hislop TG, Bajdik C, et al (2009). Comparison of breast cancer survival in two populations: Ardabil, Iran and British Columbia, Canada. *BMC Cancer*, **9**, 1-6.

Shen C, Yang H, Chang Y, et al (2012). A decision tree-based approach for cervical smears. *IJICIC*, **8**, 3251-63.

Slaoui M, Razine R, Ibrahimi A, et al (2014). Breast cancer in Morocco: a literature review. *Asian Pac J Cancer Prev*, **15**, 1067-74.

Spitz MR, Hong WK, Amos CI, et al (2007). A risk model for prediction of lung cancer. *J Natl Cancer Inst*, **99**, 715-26.

Washbrook E (2006). Risk factors and epidemiology of breast cancer. *Women's Health Med*, **3**, 8-14.

Xing MY, Xu SZ, Shen P (2014). Effect of low-fat diet on breast cancer survival: a meta-analysis. *Asian Pac J Cancer Prev*, **15**, 1141-4.