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

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 (Bartholz-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 cancer that is able to identify prognostic relationships underlying data and avoiding restrictive assumptions of conventional modeling approaches.
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., 1984) 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 menstrual 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
Diagnostic Classification Scheme in Iranian Breast Cancer Patients using a Decision Tree

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 another 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


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.

**Table 1. Variable Included in CART Analysis**

<table>
<thead>
<tr>
<th>Age</th>
<th>Ovarian Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of benign breast disease</td>
<td>Age of menarche</td>
</tr>
<tr>
<td>Family history of breast cancer</td>
<td>Menstrual pattern</td>
</tr>
<tr>
<td>History use of alcohol</td>
<td>Marriage status</td>
</tr>
<tr>
<td>History of smoking</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Case (%)</th>
<th>Control (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marriage status</td>
<td>Single 243 (77.9) 280 (89.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Divorced 7 (2.2) 3 (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Married 62 (19.9) 29 (9.3)</td>
<td></td>
</tr>
<tr>
<td>Menstrual pattern</td>
<td>Regular 60 (79.3) 285 (91.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Irregular 230 (20.7) 27 (8.7)</td>
<td></td>
</tr>
<tr>
<td>Family history of ovarian cancer</td>
<td>No 306 (98) 312 (100)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes 6 (2) 0</td>
<td></td>
</tr>
<tr>
<td>Family history of Breast Cancer</td>
<td>No 274 (88) 306 (98)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes 38 (12) 6 (2)</td>
<td></td>
</tr>
<tr>
<td>History use of alcohol</td>
<td>No 310 (99.4) 312 (100)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes 2 (6.6) 0</td>
<td></td>
</tr>
<tr>
<td>History of smoking</td>
<td>No 311 (99.7) 312 (100)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes 1 (3.1) 0</td>
<td></td>
</tr>
<tr>
<td>History of benign breast disease</td>
<td>No 300 (96.2) 311 (99.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes 12 (3.8) 1 (0.3)</td>
<td></td>
</tr>
</tbody>
</table>


