Exploratory Data Analysis on Breast Cancer Prognosis

Mohammad Mehdi Owrang O.
American University, USA

Yasmine M. Kanaan
Howard University, USA

Robert L. Copeland Jr.
Howard University,USA

Melvin Gaskins
Howard University Hospital, USA

Robert L. DeWitty Jr.
Providence Hospital, USA

INTRODUCTION

Breast cancer is the most common female cancer in the US, the second most common cause of cancer death in women ("American Cancer Society", 2016), and the main cause of death in women ages 40 to 59 (Siegel et al., 2012).

In 2016, it is estimated that 249,260 new cases of breast cancer will be diagnosed and estimated 40,890 breast cancer deaths; and an invasive breast cancer will be diagnosed in about 246,660 women and 2,600 men. An additional 61,000 new cases of in situ breast cancer will be diagnosed in women ("American Cancer Society", 2016). The lifetime probability of developing breast cancer is one in six overall (one in eight for invasive disease) ("American Cancer Society", 2013; Siegel et al., 2012).

Worldwide, breast cancer is the most frequently diagnosed cancer among women in 140 of 184 countries, according to the World Cancer Research Fund International ("BCRF", n.d.). Since 2008, breast cancer incidence has increased by more than 20 percent and mortality has increased by 14 percent ("BCRF",n.d.). Nearly 1.7 million new breast cancer cases were diagnosed in the last report of 2012.

Medical prognosis is an evaluative component of medicine that encompasses the science of estimating the complication and recurrence of disease and predictive survival of patients (Ohno-Machado, 2001). Medical prognosis plays an increasing role in health care outcome. Reliable prognostic models that are based on survival analysis statistics and techniques have been applied to a variety of domains with varying degrees of success (i.e., APACHE IV (Zimmerman, 2006)).

Breast cancer prognosis is a vital element of providing effective treatment for patients. It has become increasingly important that clinicians are provided with accurate prognostic information on which to base therapeutic decision as the range of options for the treatment of patients with breast cancer widens. A large number of factors, including tumor grade, tumor size, and lymph node status including other aspects may influence or correlate with prognosis for breast cancer patients.

Breast cancer prediction survivability has mainly been studied through clinical approaches, based on pathological factors such as tumor grade, tumor size and number of the positive lymph nodes, estrogen (ER), progesterone (PR), and human epidermal growth factor receptor 2 (Her2) receptors,

DOI: 10.4018/978-1-5225-2255-3.ch156

etc. Most studies are carried out in an effort to find factors that clarify the large unexplained variation in prognosis of the breast cancer patients. There is still uncertainty about the importance of most prognostic factors. There are other non-clinical factors such as age, ethnicity, obesity, and marital status that may have prognostic impact but are not used routinely in clinical practice.

This chapter is a survey of the significance of the non-clinical prognostic factors (i.e., age, ethnicity) in finding the prognosis for breast cancer patients. The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Public-Use Data (years 1977-2013; 421,056 records) ("SEER," n.d.) and the Howard University Cancer Center Tumor Registery (1995-2013; 1599 records) data were analysed. NPI (Nottingham Prognostic Index (Galea, 1992)) is a prognostic tool that enables grouping of patients based on calculated prognosis (i.e., excellent, good, moderate, poor) and the impact of each non-clinical prognostic factor on these subgroups. In addition, survival analysis tools such as Cox proportional hazards and Kaplan-Meier survival curve ("Cox proportional-hazards regression", 2013) were used in analyzing the data.

Our data analysis suggested that age, ethnicity, and marital status have some influence on the survivability rate of breast cancer patients. Whether such influence is significant enough in relationship with the clinical factors such as tumor size and grade remains to be further studied.

BACKGROUND

Several studies have been carried out on the survivability prediction of breast cancer using Naïve Bayes and Classification Trees, Artificial Neural Networks and statistical techniques of regression (Delen et al., 2005; Gupta et al., 2011). Delen and colleagues (Delen et al., 2005) have used data mining algorithms Artificial Neural Networks, decision trees, and logistic regression to develop the breast cancer prediction models. The results indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the sample, artificial neural networks came second with 91,2% accuracy and the logistic regression models came to be the worst of the three with 89.2% accuracy.

Investigators (Owrang & Hosseinkhah, 2013; Owrang, 2014) have used the association rules data mining technique and data mining tool XLMiner ("XLMiner", n.d.) to investigate the significance of the ethnic factor on the patient's prognosis. The results, the association rules, suggested that the ethnicity had some significance in the survivability rate of the patient.

Existing reports suggest that young breast cancer patients have poorer outcomes compared to their older counterparts, which are in part attributed to later stage disease, more aggressive tumors, and less favorable receptor status (Anders, 2009; Bharat, 2009; Saghie et al., 2006). It appears that ethnicity may further influence breast cancer prognosis (Biffl et al., 2001; Liu et al., 2013). The large SEER Program and National Cancer Data Base analyses indicated that African-American women have more advanced disease at presentation and a higher death rate compared with non-Hispanic Caucasians. Investigators (Biffl et al., 2001) studied the influence of ethnicity on the prognosis of young breast cancer patients ($\leq$ 35 years). Their study shows that Latino patients suffer more aggressive disease and poorer prognosis than other young women.

Most existing studies reflect single institution experiences that may not be representative of the whole population. Additionally, existing reports did not attempt to compare patients based on age and ethnicity relative to patients' prognostic group (i.e., excellent or good). This motivated us to examine the prognosis for breast cancer patients with respect to age, ethnicity, and marital status with a new approach based on patients' prognostic groups (i.e., excellent, moderate, good, poor).