International Journal of Medical Informatics 97 (2017) 304–311

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

Stage-specific predictive models for breast cancer survivability

Rohit J. Kate [a,*], Ramya Nadig [b]

[a] Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
[b] Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

ARTICLE INFO

Article history:
Received 28 June 2016
Received in revised form
12 September 2016
Accepted 3 November 2016

Keywords:
Breast cancer
Survivability prediction
Machine learning
SEER dataset

ABSTRACT

Background: Survivability rates vary widely among various stages of breast cancer. Although machine learning models built in past to predict breast cancer survivability were given stage as one of the features, they were not trained or evaluated separately for each stage.
Objective: To investigate whether there are differences in performance of machine learning models trained and evaluated across different stages for predicting breast cancer survivability.
Methods: Using three different machine learning methods we built models to predict breast cancer survivability separately for each stage and compared them with the traditional joint models built for all the stages. We also evaluated the models separately for each stage and together for all the stages.
Results and conclusions: Our results show that the most suitable model to predict survivability for a specific stage is the model trained for that particular stage. In our experiments, using additional examples of other stages during training did not help, in fact, it made it worse in some cases. The most important features for predicting survivability were also found to be different for different stages. By evaluating the models separately on different stages we found that the performance widely varied across them. We also demonstrate that evaluating predictive models for survivability on all the stages together, as was done in the past, is misleading because it overestimates performance.

1. Introduction

Breast cancer is the most frequently diagnosed cancer in women [1]. Even though its 5-year survival rate in United States has increased from 75.2% in 1980 to 90.6% in 2013 [2], it is currently the second leading cause of cancer deaths among women after lung cancer [1]. Accurate prediction of breast cancer survivability can enable physicians and healthcare providers to make more informed decisions about a patient's treatment. For example, they may opt for more aggressive new therapies for patients with ominous prognosis.

Several data-driven machine learning methods have been used in recent years for cancer prediction and prognosis [3,4]. These methods learn patterns or statistical regularities from historic data in order to make predictions on new data. Specifically for breast cancer, researchers have used a wide variety of machine learning methods for predicting susceptibility [5–7], diagnosis [8–14], recurrence [15–21] and survivability [22–30]. This paper focusses only on predicting breast cancer survivability. For this task, some of the researchers who developed machine learning models had access to patients' genomic and detailed clinical data from medical centers on which they trained their methods [22–24,29,30]. Although smaller in size (in the order of a few hundred cancer incidences), these datasets were more detailed in patient information. But most other researchers with no access to such detailed patient data used the publicly available SEER cancer dataset [31] for training their methods [25–28]. We have used this dataset for this paper. Although this dataset does not include genomic or detailed clinical information, its large size (in the order of a few hundred thousand cancer incidences) makes it suitable for building accurate models for survivability.

Artificial neural networks (ANN) [32], support vector machines (SVM) [33], decision Trees [34], naïve Bayes [35] and logistic regression [36] are the most common machine learning methods that have been used for predicting breast cancer survivability [22,25–27,29]. In addition, researchers have proposed methods to improve performance on this task through semi-supervised learning [28] as well as through ensemble learning [23,24]. Although a broad range of methods and training mechanisms have been used and evaluated for predicting breast cancer survivability, to the best of our knowledge, no distinction was ever made between differ-

* Corresponding author at: Department of Health Informatics and Administration, 2025 East Newport Avenue, Milwaukee, WI 53211, USA.
E-mail address: katerj@uwm.edu (R.J. Kate).

http://dx.doi.org/10.1016/j.ijmedinf.2016.11.001
1386-5056/

ent cancer stages either for training the predictive models or for evaluating them.

Cancer incidences are assigned stages based on tumor size and the extent of spread, hence survivability varies widely between them. There are more than one cancer staging systems currently in use. One system categorizes cancers to be in Stage 0, Stage I, . . ., Stage IV with further subcategories [37]. TNM (tumor, node, metastasis) is another cancer staging system in which stages are assigned based on the status of tumor, node and metastasis [37]. SEER dataset uses a system in which the stages are: in-situ, localized, regional and distant, based on the spread of cancer [31]. These are called *summary stages*. In in-situ summary stage abnormal cells are confined to the layer of cells in which they developed; in localized summary stage cancer is limited to the organ in which it began; in regional summary stage it has spread to nearby lymph nodes, tissues and organs; and in distant summary stage it has spread beyond to distant lymph nodes, tissues and organs.

In the part of the SEER dataset that we used in the current study, we found that the survivability rate for in-situ summary stage breast cancer was 99.42% while for distant summary stage breast cancer it was 36.17%. The survivability rates for other summary stages were in between. Clearly, it is far easier to predict survivability for in-situ summary stage than for other summary stages. Hence an evaluation of any breast cancer survivability prediction model should distinguish between these summary stages. In addition, given their wide range of survivability rates and the differences between them in terms of the spread of cancer, it is conceivable that a machine learning method trained specifically on a summary stage would be more suitable to predict survivability for that summary stage. However, this was not tested in previous work which had used summary stage only as one of the several features for training machine learning methods.

In this paper, we compare breast cancer survivability prediction models trained on all summary stages and trained separately on each summary stage. In addition, we compare how the performance changes with increasing amounts of training data in each case. We also present which features are most indicative of survivability for different summary stages. We present our evaluation results separately for each summary stage to show the differences between them in terms of the prediction performance. We also show that presenting evaluation results together for all summary stages, as had been done previously, leads to an overestimation of the performance because of the inherent high to low variation in survivability rates between different summary stages.

2. Materials and methods

2.1. Dataset

We used the publicly available SEER cancer dataset [31]. This data is collected on an ongoing basis from various registries in the US representing around 28% of the US population. It is part of the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program. The data is publicly available and can be obtained after signing a data use agreement. Its latest version, which we used in this study, covers de-identified cancer incidences from years 1973 through 2013 to a total of 9.18 million cancer incidences which includes 1.47 million breast cancer incidences. The dataset associates unique identifiers with patients using which one can track multiple incidences of cancer for every patient. In case a patient had multiple breast cancer occurrences, we only considered the last occurrence for predicting survivability (we found that using the number of occurrences as a feature did not improve prediction models).

if SM $\geq$ 60 and VSR = "alive" **then** *survived*
else if SM < 60 and COD = "breast cancer" **then** *not survived*
else exclude the patient

Fig. 1. Logic used to determine survivability of breast cancer patients from the SEER dataset using the attributes– survival months (SM), vital status recode (VSR) and cause of death (COD). Survivability is defined as surviving for five years (60 months) after diagnosis.

Table 1
Summary stage-wise survivability statistics of the breast cancer data used in this study which was subset of the SEER dataset.

	Total incidences	Survived	Not survived	Percent survived
All stages	174,518 (100%)	160,626	13,892	92.04%
In-situ	10,106 (5.79%)	10,047	59	99.42%
Localized	106,390 (60.96%)	102,737	3653	96.57%
Regional	55,340 (31.71%)	46,872	8468	84.70%
Distant	2682 (1.54%)	970	1712	36.17%

Each cancer incidence in the SEER dataset is associated with several cancer relevant attributes in addition to patients' demographic information. Three of these attributes can be used to determine survivability of a patient: survival months (SM) which tells the number of months a patient survived, vital status recode (VSR) which takes value "alive" or "dead", and the cause of death (COD). Cancer survivability is most commonly defined as surviving for five years (60 months) after diagnosis. Using this definition of survivability, we used the logic shown in Fig. 1 to determine whether a breast cancer patient in SEER dataset survived or not. The same logic was used in prior work [26] (attribute survival months (SM) was formerly called survival time recode (STR)). Given that we are building model for breast cancer survivability, the logic excludes all the incidences in which the patient died due to some cause other than breast cancer. We also excluded the patients if any one of these three attributes was not known for them. This logic is used to determine the survivability gold-standard for the purpose of training and evaluating the predictive models. The predictive models do not use these three attributes as features.

Although less common, breast cancer also occurs in men accounting for 1% percent of all the incidences [1]. However, for the purpose of this study we only focused on incidences in women. The next subsection describes the attributes of the SEER dataset that we used in our predictive models as features. Codes for some of these attributes were redefined in the year 2004 and a few new attributes were also introduced in the same year. Hence for consistency and given the abundancy of incidences each year, we decided to exclude incidences of breast cancer diagnosis before 2004. Note that most of the previous work on breast cancer survivability from SEER dataset had instead excluded incidences diagnosed after 2004. Given that survivability rates have changed over the years, it is better to use the more recent data as we have done in this study. We also excluded incidences if any of their feature values were unknown. Given that survivability is defined as surviving for five years after diagnosis, we had to also exclude incidences which were diagnosed less than five years ago from the latest year of submission for the current data. In the Results section, we show through learning curves that even after all the exclusions we were left with more than sufficient data for training and evaluating the prediction models.

Table 1 shows the statistics of the data we used in this study categorized by summary stage which is one of the attributes in the SEER dataset. There were a total of 174,518 incidences of breast cancer with 92.04% survival rate which is consistent with the current survival rate [1]. Given that in-situ summary stage had an almost sure survival rate of 99.42% and had only 5.79% incidences, we did not see much value in building models for predicting sur-