*This is the accepted manuscript version of the article: "*Rohit J. Kate and Ramya Nadig. Stage-Specific Predictive Models for Breast Cancer Survivability. International Journal of Medical Informatics, Volume 97, 304-11, January 2017."
The final published version can be downloaded from: http://dx.doi.org/10.1016/j.ijmedinf.2016.11.001

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

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.

Keywords

breast cancer; survivability prediction; machine learning; SEER dataset

[*] Corresponding author at: Department of Health Informatics and Administration, 2025 East Newport Avenue, Milwaukee WI 53211, USA.
Email address: katerj@uwm.edu (Rohit J. Kate).

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[5,6,7], diagnosis[8,9,10,11,12,13,14], recurrence[15,16,17,18,19,20,21] and survivability[22,23,24,25,26,27,28,29,30]. 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[22,23,24,29,30]. 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[25,26,27,28]. 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[22,25,26,27,29]. 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 different 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 ≥ 60 and VSR = "alive" **then** *survived*
else if SM < 60 and COD = "breast cancer" **then** *not survived*
else exclude the patient

Figure 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.

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 Figure 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 survivability for this summary stage. Hence we excluded all the in-situ incidences from the rest of our analysis. SEER dataset has four subcategories of the regional summary stage (direct extension, lymph node extension, extension and nodes, and not specified). We did not distinguish between these subcategories and grouped them all under regional summary stage.

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	3,653	96.57%
Regional	55,340 (31.71%)	46,872	8,468	84.70%
Distant	2,682 (1.54%)	970	1,712	36.17%

2.2 Features

Machine learning methods require data represented in terms of features which are indicative of the class being predicted. For predicting breast cancer survivability, several attributes present in the SEER dataset have been used as features. Table 2 shows the features we used in this study which are largely same as were used in most of the previous work on this task[25,26,27,28]. Note that summary stage is also used as a feature.

Table 2. List of features used to build predictive models for breast cancer survivability.

Feature	Type	Description
Age	Numeric	Age at diagnosis
Behavior code	Nominal	Code based on aggressiveness of tumor
Extension	Nominal	Information on extension of tumor
Grade	Nominal	Category based on the appearance of tumor
Histologic type	Nominal	Form of tumor
Lymph nodes	Nominal	The highest specific lymph node chain that is involved by the tumor
Marital status	Nominal	Marital status at diagnosis
Metastasis at diagnosis	Nominal	Information on distant metastasis
Primary site	Nominal	Site in which the primary tumor originated
Race	Nominal	
Radiation	Nominal	Method of radiation therapy used in the first course of treatment
Regional nodes examined	Numeric	Number of regional lymph nodes removed and examined
Regional nodes positive	Numeric	Number of regional lymph nodes that contained metastases
Site-specific surgery code	Nominal	Code for surgery of primary site as first course of therapy
Summary stage	Nominal	Defined according to the spread of cancer
Tumor size	Numeric	Size in mm

2.3 Machine Learning Methods

We used naïve Bayes[35], logistic regression[36] and decision tree[34] machine learning classification methods to predict breast cancer survivability in this study. We used the free and publicly available Weka software[38] for these methods.

We used Weka's "NaiveBayes" classification method for naïve Bayes, "Logistic" classification method for logistic regression and "ADTree" (alternating decision trees[39]) classification method for decision trees which is based on the machine learning ensemble algorithm of boosting[40]. Among the decision tree classification methods available in Weka, we found it to work best for our dataset through a pilot study.

Our breast cancer survivability dataset is very unbalanced with far more patients surviving than not surviving, except for the distant summary stage for which the reverse is true. For an unbalanced dataset like this, a machine learning method set to maximize accuracy can simply classify all the examples to be of the majority class and get a very high accuracy, but such a classifier will be practically of no use. Hence in order to maximize classification accuracy on both the classes, typically a weight is specified to the minority class which relatively increases the penalty of misclassifying it compared to misclassifying the majority class. Weka implements this mechanism through its cost-sensitive meta-classifier. We ran all our experiments using the cost-sensitive meta-classifier which internally used a base classifier of naïve Bayes, logistic regression or decision trees. The right weight was determined out of 0.25, 0.5, 1, 2, 4, 6, …, 18, 20 through five-fold internal cross-validation within the training data. The weight that maximized the area under ROC curve on the internal cross-validation was then used to train the machine learning method using the entire training data. Weka's default values were used for all other parameters for all the methods.

2.4 Experimental Methodology

We built two types of predictive models for breast cancer survivability using each of the machine learning methods. The first type, which we call *joint* model, was trained using training data that included incidences of all the summary stages. This is how the predictive models for breast cancer survivability were built in the past by other researchers. The second type, which we call *summary stage-specific* models, were built separately for each summary stage using training data of incidences from only that respective summary stage.

After training, a joint model was first evaluated on test incidences of all the summary stages, as was done in the past work. But in this study, it was then also separately evaluated on each summary stage. This did not require applying the model again, but only required doing separate evaluations for each summary stage on the output already

generated for the earlier evaluation. After training, each summary stage-specific model was first evaluated on test incidences of its respective summary stage. Next, outputs of all the three summary stage-specific models were also combined to evaluate their performance on all summary stages together. This is equivalent of using a meta-model that first looks at the summary stage of a test incidence and then applies the corresponding summary stage-specific model.

Various evaluation measures have been used to report performance of predictive models for survivability in the past, including accuracy, precision, recall, sensitivity, and specificity. However, most machine learning methods used to build predictive models also assign confidences to their outputs and by varying threshold on these confidences one can obtain a range of different values for these evaluation measures. For example, precision can be traded off with recall while sensitivity can be traded off with specificity. Thus reporting single values for these evaluation measures does not present the full spectrum of performance and a better way to gauge performance of a predictive model is by plotting an entire precision-recall curve or an ROC curve[41]. The latter is a curve between a model's true positive rate (fraction of positive examples the model correctly classified as positive, also known as sensitivity) and false negative rate (fraction of negative examples the model incorrectly classified as positive, equivalent to 1-specificity). In our task the positive class is "survived" and the negative class is "not survived".

A useful property of ROC curve is that, unlike precision-recall curve, it is independent of the class distribution, i.e. it will not change if the distribution of positive and negative examples is changed in the test dataset[42]. Area under ROC curve (AUC) is a representative summary statistic of an ROC curve and is widely used to report performance as a single value. For all the reasons outlined above, we used AUC to report performance of our predictive models in this paper. A higher AUC means a better performance. A random classifier will have a diagonal ROC curve with an AUC of 0.5 while a perfect classifier will have the maximum AUC of 1.

When we combined the outputs of all the three summary stage-specific models to evaluate their performance on all summary stages together (as described before), we needed to generate one ROC curve using three different thresholds. Although there exists a principled way to combine multiple ROC curves generated using multiple methods on the same test set by taking their convex hull[43], our case is different in which we have to instead combine

results obtained on three different test sets (one for each summary stage). In order to do this, we considered all combinations of the three confidence thresholds (by varying them in small step size of 0.0025), and for each combination we determined the combined true positive rate and false negative rate for the data which gave us one point on the ROC graph. In this way, different threshold combinations gave us different points on the ROC graph. Using all the points thus obtained on the ROC graph, we plotted the ROC curve by choosing the highest true positive rate point for every false negative rate. We are justified in doing so because when deploying the model one would always choose the threshold combination that gives the best point in ROC graph over other threshold combinations that give comparably worse points.

All evaluations were done using the standard 10-fold cross-validation procedure[41]. Given the unbalanced nature of our data in terms of survived and not survived classes, as well as given the uneven distribution among different summary stages, all folds were stratified for uniformity so that each fold had the same distribution of classes and summary stages. In each fold, the training data for separate summary stage-specific models were the disjoint subsets of the training data for the joint model in the corresponding fold. This makes the results of the model more comparable. The exact same folds were used across different experiments. In our results, we reported average AUC scores of the ten folds of cross-validation and used the individual scores of the folds for statistical significance testing.

3 Results and Discussion

3.1 Joint vs. Summary Stage-Specific Predictive Models

Table 3 shows the AUC results obtained by joint and summary stage-specific models trained using naïve Bayes, logistic regression and decision trees. The numbers shown in bold were found statistically significant ($p < 0.05$) to the corresponding numbers in the same row by the same machine learning method using two-tailed paired t-test.

Table 3. Area under ROC curve (AUC) obtained by joint and summary stage-specific predictive models for breast cancer survivability using three different machine learning methods. The numbers shown in bold were found to be statistically significant ($p < 0.05$; two-tailed paired t-test) to the corresponding numbers in the same row for the same machine learning method. It can be seen from the rest of the rows that the numbers in the first row for all stages combined are misleadingly high.

	Naïve Bayes		Logistic Regression		Decision Trees	
	Joint	Stage-specific	Joint	Stage-specific	Joint	Stage-specific
All Stages	0.821	**0.842**	0.848	0.850	**0.846**	0.840
Localized	0.717	**0.768**	0.771	**0.774**	0.773	**0.779**
Regional	0.759	**0.778**	0.794	0.794	0.781	**0.789**
Distant	0.644	**0.712**	0.712	0.721	0.665	**0.711**

From Table 1, there were 106,390 incidences in localized summary stage, 55,340 in regional and 2,682 in distant. Thus the total number of incidences in all summary stages were 164,412 (note that we excluded in-situ incidences). Since we used 10-fold cross-validation, the size of the training data was $9/10^{th}$ of the above numbers in each fold. Thus given that a joint model was trained with significantly more number of incidences than any of the summary stage-specific models, one would have expected them to also perform significantly better than them for each machine learning method. However, as can be observed from Table 3, joint models do either significantly worse or nearly equal to the summary stage-specific models (except for decision tree in the "All Stages" row, however, we later show that the results of this row are overestimations and hence misleading). This clearly points out that there is no advantage of building a joint model over building separate summary stage-specific models, in fact, it can lead to worse performance in some cases. Training separate models that use smaller training datasets also has an added advantage of faster training times for the machine learning methods that have higher than linear training time complexity.

It can be also observed from Table 3 that for naïve Bayes the performance of joint model is always significantly worse than summary stage-specific models while for logistic regression and decision trees the differences are smaller. Given that in each fold, the training data for joint model always included all the incidences of the training data of each summary stage model, a good machine learning method will be expected to do at least at par with each summary stage-specific model when tested on the corresponding summary stage. This seems to be the case majority

of times with logistic regression and decision trees. However, naïve Bayes joint model has been consistently negatively affected by adding examples of other summary stages during training. We will look into this more closely in the next subsection using learning curves.

Next important observation from Table 3 is that the performance on distant summary stage is much worse than other two summary stages. This is true for both joint as well as summary stage-specific models and for all the machine learning methods. This shows that it is harder to predict breast cancer survivability for the distant summary stage and there is a larger room for improvement. Note that this observation was not made in any previous work because prediction models had been always evaluated together on all summary stages. The worse performance on distant summary stage by distant summary stage-specific model could be because of its smaller data size available for training (and again, including incidences from other summary stages for training as done in joint models did not help). We will further look into this in the next subsection.

An interesting observation form Table 3, looking at each column separately, is that the performance on all summary stages together was always found to be better than performance on any of the individual summary stages, whether for the joint model or for the summary stage-specific models. At first this may appear counter-intuitive because one is used to seeing performance on a combination as an average of the performance on individual components. However, ROC curves are less intuitive, and in fact, as we demonstrate next with an example, evaluation on all summary stages together leads to an overestimation of performance.

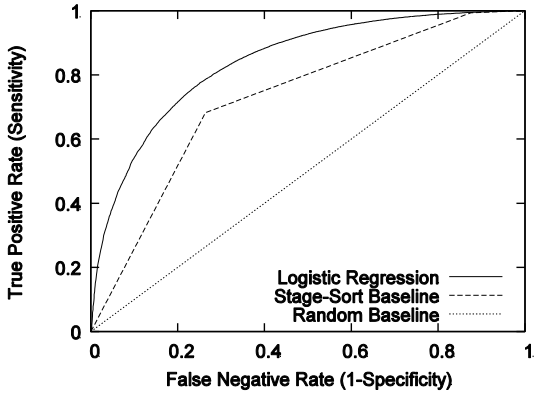

Figure 2. ROC curves obtained for survivability prediction when evaluated on all summary stages together. The ROC curve for the stage-sort baseline model illustrates that it is easy to obtain a seemingly reasonable performance

in this evaluation setup. The ROC curves for logistic regression joint model and random baseline model are shown for comparison.

Consider a baseline joint model that simply looks at the summary stage of an incidence (which is a feature) and assigns survivability confidence to that incidence equal to 1 if it is localized, 0.5 if it is regional and 0 if it is distant (the actual numbers do not matter as long as they are in a decreasing order by the summary stages). That is, this model simply predicts survivability in the order of the known survivability rates of the summary stages. We will call this the *stage-sort baseline* model. Figure 2 shows the ROC curve for this model. For comparison, ROC curves for the logistic regression joint model and a random classification baseline model are also shown. The ROC curve for the stage-sort baseline has essentially two points in the interior area of the graph, one point for which the model predicts everyone in localized summary stage to survive and everyone else to not survive, and another point for which the model predicts everyone in localized and regional summary stages to survive and the rest to not survive. The second point can be seen close to the top of the graph. The point (0,0) is obtained by predicting everyone to not survive and the point (1,1) is obtained by predicting everyone to survive. The in-between points can be obtained by predicting random subsets to survive within each summary stage. The AUC of this curve was found to be 0.726. Hence one can obtain what looks like a reasonable performance on an evaluation on all summary stages together by simply exploiting the known survivability rates of the individual summary stages. If we now evaluate performance of this stage-sort baseline model separately on each summary stage, we will get the diagonal baseline ROC curve that connects point (0,0), i.e. predicting none to survive and (1,1), i.e. predicting all to survive, with in-between points obtained if random subsets are predicted to survive. The AUC of these ROC curves obtained by the stage-sort baseline model when evaluated on each summary stage separately will be thus 0.5. Hence this example demonstrates that one can obtain a higher AUC by evaluating a model on all summary stages together than by evaluating it separately on each of the summary stages. This is possible by exploiting the differences in the known survivability rates of different summary stages. Note that while this argument was specific to ROC and AUC, the stage-sort baseline will also have a misleadingly good precision and recall (and hence F-measure) on an evaluation on all the summary stages together. For example, predicting everyone in localized and regional summary stages to survive and everyone in distant summary stage to not survive gives precision of 92.5% (fraction of predicted to be survived actually survived) and recall of 99.36% (fraction of actually survived that were predicted correctly).

The above illustrative example was for a baseline model, but in general, a machine learning predictive model can easily learn to rank localized summary stage incidences generally above other summary stage incidences, and regional summary stage incidences generally above distant summary stage incidences. This will give it an artificial boost in performance when measured on all summary stages together. We believe this is what happened with all our models in the "All Stages" row of Table 3 as well as with all the models reported in the previous work on this task. Note that this also happened when the outputs of summary stage-specific models were combined, because in this case as well the models will generally have higher confidences for incidences in summary stages with higher survivability rates. However, exploiting the differences in survivability of different summary stages in assigning confidences to survivability is not practically useful. For example, in order to predict survivability for a patient who is known to be in the localized summary stage, giving her an extra confidence of survivability because of other patients who are in worse summary stages is although not incorrect, but is not helpful, because it is essentially stating the obvious. To avoid overestimation of performance when measured on all summary stages together, we recommend evaluating prediction models for survivability only separately on each summary stage. However, in case performance is measured on all summary stages together, it should be at least compared with the stage-sort baseline.

3.2 Learning Curves

We plotted learning curves to see how the performance of joint and summary stage-specific models vary with increasing amounts of training data when evaluated on each summary stage. To obtain a point on the learning curve, we used the same training and evaluation procedure as before, but used only a portion of the entire training data available in each fold. The training data was cumulatively increased to obtain higher points on the learning curve. For uniformity, through stratification each portion of the training data used for training had the same distribution of summary stages (in case of joint models) and survivability rate. The test data remained exactly the same.

Joint model is trained using all the summary stages and hence it has more incidences available for training, but in order to fairly compare it to a particular summary stage-specific model across learning curve, the joint model was given maximum same number of training examples as were available to that summary stage-specific model. For example, for the localized summary stage, both the models were given maximum 95,751 ($9/10^{th}$ of 106,390) training examples in each fold; for the summary stage-specific model they were all localized summary stage incidences, but

for the joint model they were a mix of the three summary stage incidences randomly selected from the original training data of the fold while maintaining their relative distribution.

Figure 3 shows all the learning curves obtained by naïve Bayes and logistic regression methods by evaluating joint and summary stage-specific models on each of the summary stages. A common trend that can be observed in the graphs is that summary stage-specific models did better than joint models with less training data. This trend continued for naïve Bayes method even with full training data (Figures 3(a), 3(c) and 3(e)). From Table 3 we know that even after including all the incidences of other summary stages available to train the joint model, its performance lagged behind that of summary stage-specific models. However, for logistic regression method, joint model caught up with the summary stage-specific models with more training data, but its performance never significantly exceeded even after including all the incidences of other summary stages (Table 3). This indicates that it is only the presence of incidences of the particular summary stage in the training data that contributes towards its performance on that summary stage. Although the presence of incidences of other summary stages in the training data did not help logistic regression method, it was not adversely affected by it unlike naïve Bayes method. It is clear why naïve Bayes method should be adversely affected by the presence of incidences of other summary stages given that it computes all its probabilities using frequency counts from the entire training data. Logistic regression method, on the other hand, is less susceptible to this and can learn to be more discriminative with various summary stages. An intelligent method can, in fact, internally learn effectively a different model for each summary stage (for example, a decision tree with summary stage as the root node). Hence whether a joint model will do worse than a summary stage-specific model or not depends on the individual machine learning method, however, in our study we did not find a joint model ever doing statistically significantly better than a summary stage-specific model.

Another important observation from Figure 3 is that for localized and regional summary stages, learning has plateaued for all the models. But it has not plateaued for the distant summary stage for either type of model using logistic regression (Figure 3 (f)). In that case, summary stage-specific model is doing better than joint model with same number of training examples, but from Table 3 we know that eventually with equal number of distant summary stage incidences in the training data, the joint model does as well as the summary stage model. However, from Figure 3 (f) it is clear that the summary stage-specific model can do better with more training incidences.

Hence while the current models are not expected to do any better on localized and regional summary stages with more training data, there is a scope of improvement on the distant summary stage. This indicates that collecting more data for distant summary stage is likely to improve performance on its survivability prediction. We would not have made this observation without separately evaluating the models on different summary stages.

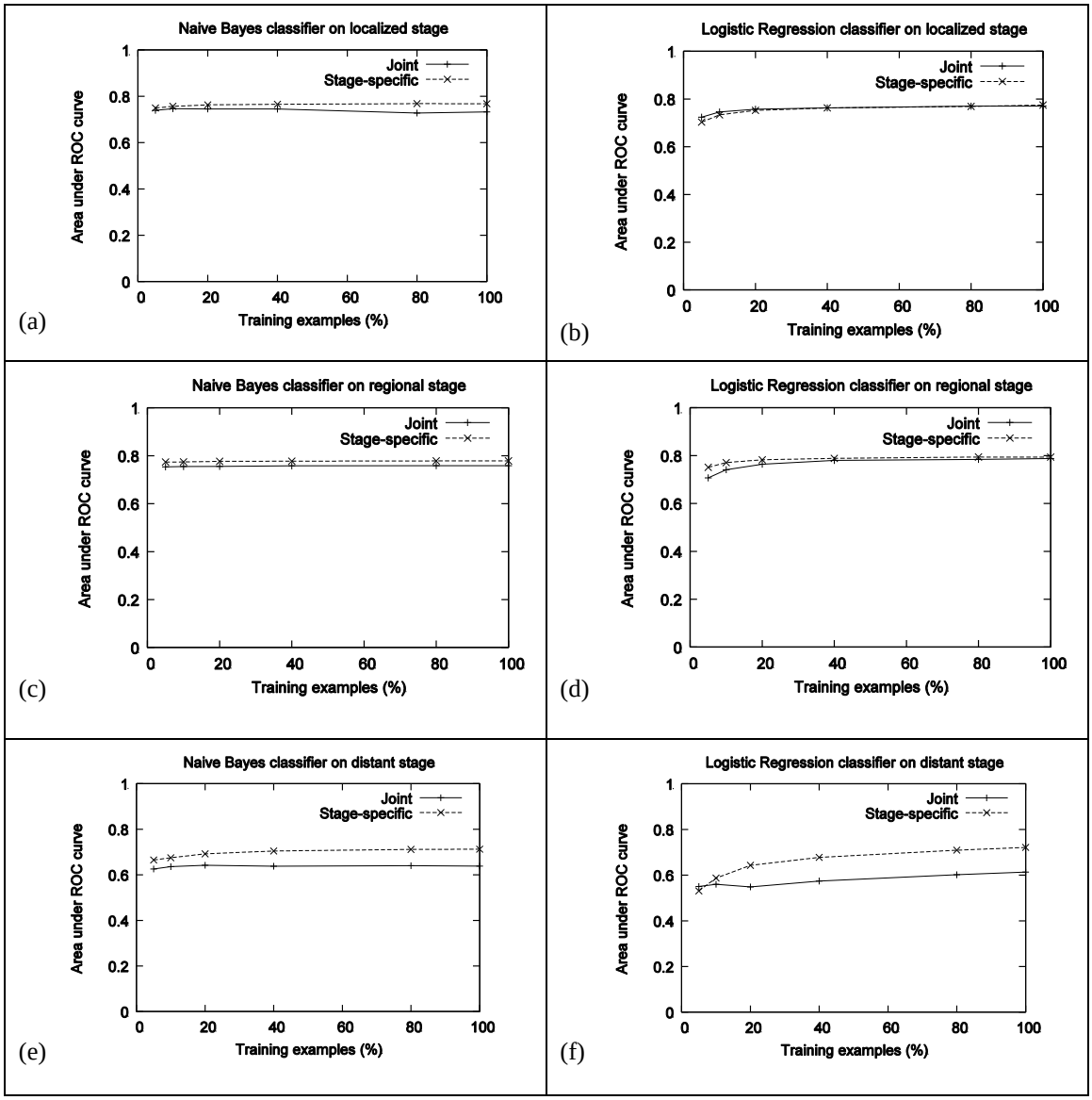

Figure 3. Learning curves comparing joint and summary stage-specific models trained using naïve Bayes (left) and logistic regression (right) and evaluated on localized (top), regional (middle) and distant (bottom) summary stages. In each graph, the maximum number training incidences were same for each model.

3.3 *Differences between Summary Stage-Specific Predictive Models*

To further see the differences between breast cancer survivability predictive models trained using incidences of different summary stages, we plotted information gain statistic for all the features separately for each summary stage. Information gain statistic evaluates the importance of a feature with respect to the class[44] for classification and can be obtained using Weka. Figures 4 (a)-(d) show the plots for each of the summary stages and for all of them combined. The features are ranked in each graph in the order of their importance. The y-axis scale has been kept same in all the graphs for direct comparison.

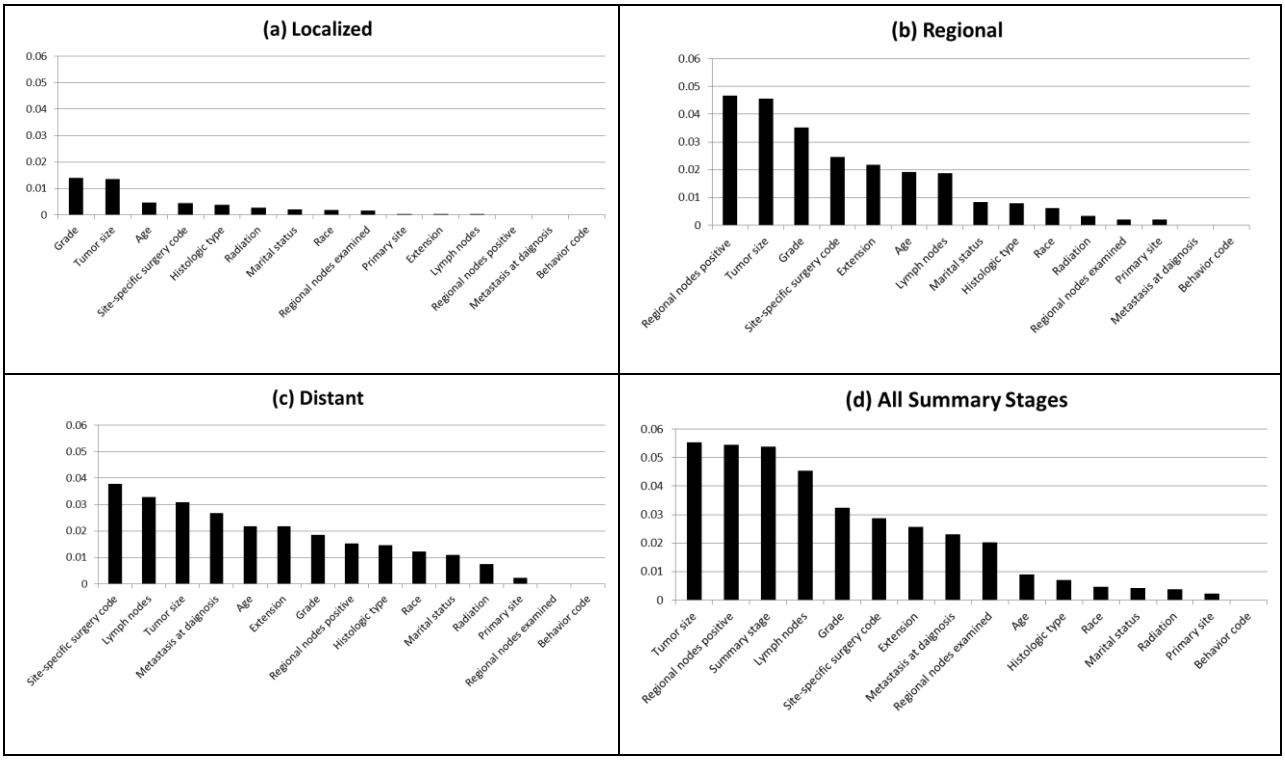

Figure 4. Information gain statistic for all the features for predicting survivability on the different summary stages separately and on all summary stages together.

Tumor size always ranked among the top three features in all the graphs implying that it is highly indicative of survivability of a patient in any summary stage. But there are differences in the importance of rest of the features in

various summary stages. While grade was the top feature for localized summary stage, it progressively gets lower rank from regional to distant summary stages. Regional nodes positive was the most important feature for regional summary stage, however it ranks far lower in the other two summary stages. Site-specific surgery code was the most important feature for distant summary stage but ranked fourth in the other summary stages. Metastasis at diagnosis was ranked fourth in distant summary stage but ranked very far behind in other summary stages. This makes clinical sense because metastasis is a distinguishing characteristic of the distant summary stage. These examples show that the predictive power for most of the features in predicting survivability significantly varies across the summary stages.

The actual information gain values of all the features for localized summary stage were lower than the other two summary stages. Figure 4 (d) shows that the information gain values for all the summary stages combined which appears to be a mix of what was found separately for each summary stage. Note that it also includes summary stage as a feature which is ranked third. Overall these plots show that each summary stage has a unique set of features that are indicative of survivability for that summary stage. This further supports separately building predictive models for breast cancer survivability for each summary stage.

4 Conclusions

In this paper we built predictive models for breast cancer survivability using SEER dataset and machine learning methods. Unlike previous work, besides building one joint predictive model for all summary stages, we also built separate predictive models for different summary stages. Our experiments showed that a joint model offers no advantage over separate summary stage-specific models, and in fact, can lead to worse performance. We also showed that each summary stage is different from others in terms of the features that indicate survivability which further supports building separate models for them. Based on our findings, to predict survivability of a patient in a particular summary stage we recommend using a model specifically trained with incidences of only that summary stage. By evaluating the models separately on each summary stage, our experiments also revealed differences in the performance on different summary stages. Predictive models were found to perform worst on the distant summary stage with still room for improvement. We also illustrated that an evaluation on all the summary stages together, as was done in the past, is misleading because it has an inherent inclination towards overestimating the performance.

Hence we recommend reporting performances separately for each summary stage which is more appropriate as well as more informative.

Authors' Contributions

RJK conceived the research idea, did the experiments presented in the paper and wrote the manuscript. RN did early ground work of processing the data and building prediction models.

Conflicts of Interests

None

Acknowledgements

We thank National Cancer Institute's Surveillance, Epidemiology, and End Results Program for making the dataset used in this work available.

Summary Points

What was already known on this topic:

- Machine learning methods have been used to predict survivability of breast cancer.
- These methods were always trained and evaluated jointly on all the stages of breast cancer.
- Breast cancer stage was used only as a feature in training machine learning methods.

What this study added to our knowledge:

- There are enough differences between various stages of breast cancer that it is best to train machine learning methods separately for each stage.
- Survivability prediction performance varies widely across different stages of breast cancer with worst performance on the distant summary stage.
- Evaluating predictive models for survivability on all the stages together as was done in the past has an inherent shortcoming which leads to an overestimation of performance, hence such models should be evaluated separately for each stage.

References

1. American Cancer Society, Cancer facts & figures 2016, Atlanta: American Cancer Society 2016.

2. National Cancer Institute: Surveillance, epidemiology, and end results program, SEER stat fact sheets: female breast cancer. URL: http://seer.cancer.gov/statfacts/html/breast.html (Accessed September 9, 2016).

3. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informatics 2006;2:59-77.

4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 2015;13:8-17.

5. Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, Mackey J, Wishart D, Greiner R, Zanke B. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clinical Cancer Research. 2004;10(8):2725-2737.

6. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited. Cancer. 2010;116(14):3310-3321.

7. Ayer T, Chhatwal J, Alagoz O, Kahn Jr CE, Woods RW, Burnside ES. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics. 2010;30(1):13-22.

8. Wang XH, Zheng B, Good WF, King JL, Chang YH. Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. International Journal of Medical Informatics. 1999;54(2):115-126.

9. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007;17(4):694-701.

10. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications. 2009;36(2):3240-3247.

11. Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial Intelligence in Medicine. 2002;25(3):265-281.

12. Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications. 2011;38(7):9014-9022.

13. Liu HX, Zhang RS, Luan F, Yao XJ, Liu MC, Hu ZD, Fan BT. Diagnosing breast cancer based on support vector machines. Journal of Chemical Information and Computer Sciences. 2003;43(3):900-907.

14. Kiyan T, Yildirim T. Breast cancer diagnosis using statistical neural networks. Journal of Electrical and Electronics Engineering. 2004;4(2):1149-1153.

15. Dai H, van't Veer L, Lamb J, He YD, Mao M, Fine BM, Bernards R, van de Vijver M, Deutsch P, Sachs A. Stoughton R. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Research. 2005;65(10):4059-4066.

16. Jerez-Aragonés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine. 2003;27(1):45-63.

17. De Laurentiis M, De Placido S, Bianco AR, Clark GM, Ravdin PM. A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. Clinical Cancer Research. 1999;5(12):4133-4139.

18. Marchevsky AM, Shah S, Patel S. Reasoning with uncertainty in pathology: artificial neural networks and logistic regression as tools for prediction of lymph node status in breast cancer patients. Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc. 1999;12(5):505-513.

19. Naguib RN, Adams AE, Horne CH, Angus B, Smith AF, Sherbet GV, Lennard TW. Prediction of nodal metastasis and prognosis in breast cancer: a neural model. Anticancer Research. 1996;17(4A):2735-2741.

20. Mariani L, Coradini D, Biganzoli E, Boracchi P, Marubini E, Pilotti S, Salvadori B, Silvestrini R, Veronesi U, Zucali R, Rilke F. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. Breast Cancer Research and Treatment. 1997;44(2):167-178.

21. Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, Park RW. Development of novel breast cancer recurrence prediction model using support vector machine. Journal of Breast Cancer. 2012;15(2):230-238.

22. Jonsdottir T, Hvannberg ET, Sigurdsson H,Sigurdsson S. The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. Expert Systems with Applications 2008;34(1):108-118.

23. Thongkam J, Xu G, Zhang Y, Huang F. Breast cancer survivability via AdaBoost algorithms. Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management 2008;80:55-64.

24. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, Sauerwine BA, Shimoni Y, Vollan HK, Mecham BH, Rueda OM. Improving breast cancer survival analysis through competition-based multidimensional modeling. PLOS Computational Biology 2013;9(5)1-15.

25. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine 2005;34:113-127.

26. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. Proceedings of the Scientific Data Mining Workshop at the SIAM Conference in Data Mining 2006.

27. Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. Engineering Applications of Artificial Intelligence 2013;26(9):2194-2205.

28. Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. Journal of the American Medical Informatics Association 2013;20(4):613-618.

29. Xu X, Zhang Y, Zou L, Wang M, Li A. A gene signature for breast cancer prognosis using support vector machine. Proceedings of 5th International Conference on Biomedical Engineering and Informatics (BMEI) 2012;928-931.

30. Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics. 2006;22(14):184-190.

31. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2013), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission.

32. Haykin S. Neural network: A comprehensive foundation, 2nd ed. Prentice Hall; 1998.

33. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press 2000.

34. Quinlan R. C4.5: Programs for machine learning. Morgan Kaufmann Publishers;1993.

35. Lewis DD. Naive (Bayes) at forty: the independence assumption in information retrieval. Proceedings of the European Conference on Machine Learning (ECML) 1998;4-15.

36. Hosmer Jr DW, Lemeshow S. Applied logistic regression. John Wiley & Sons; 2004.

37. Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti AI. AJCC cancer staging manual, Vol. 649. New York: Springer 2010.

38. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P,Witten IH. The WEKA data mining software: an update. SIGKDD Explorations 2009;11(1):10-18.

39. Freund Y, Llew M. The Alternating Decision Tree Algorithm. Proceedings of the 16th International Conference on Machine Learning (ICML) 1999;124-133.

40. Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. European Conference on Computational Learning Theory 1995;23:37.

41. Japkowicz N, Shah M. Evaluating learning algorithms: a classification perspective. Cambridge University Press 2011.

42. Provost F, Fawcett T. Data science for business: what you need to know about data mining and data-analytic thinking, Chapter 8: Visualizing model performance. O'Reilly Media, Inc. 2013.

43. F Provost, Fawcett T. Robust classification for imprecise environments. Machine Learning 2001;42(3):203-231.

44. Kent JT. Information gain and a general measure of correlation. Biometrika 1983;70(1):163-173.