Inf Syst Front (2017) 19:1233–1241
DOI 10.1007/s10796-017-9764-0

Revealing determinant factors for early breast cancer recurrence by decision tree

Jimin Guo[1,2,3] · **Benjamin C. M. Fung**[4] · **Farkhund Iqbal**[5] · **Peter J. K. Kuppen**[6] · **Rob A. E. M. Tollenaar**[6] · **Wilma E. Mesker**[6] · **Jean-Jacques Lebrun**[1]

Published online: 1 June 2017

Abstract Early breast cancer recurrence is indicative of poor response to adjuvant therapy and poses threats to patients' lives. Most existing prediction models for breast cancer recurrence are regression-based models and difficult to interpret. We apply a Decision Tree algorithm to the clinical information of a cohort of non-metastatic invasive breast cancer patients, to establish a classifier that categorizes patients based on whether they develop early recurrence and on similarities of their clinical and pathological diagnoses. The classifier predicts for whether a patient developed early disease recurrence; and is estimated to be about 70% accurate. For an independent validation cohort of 65 patients, the classifier predicts correctly for 55 patients. The classifier also groups patients based on intrinsic properties of their diseases; and for each subgroup lists the disease characteristics in a hierarchal order, according to their relevance to early relapse. Overall, it identifies pathological nodal stage, percentage of intra-tumor stroma and components of TGFβ-Smad signaling pathway as highly relevant factors for early breast cancer recurrence. Since most of the disease characteristics used by this classifier are results of standardized tests, routinely collected during breast cancer diagnosis, the classifier can easily be adopted in various research and clinical settings.

Keywords Breast cancer · Recurrence · Decision tree · Classifier · Stroma · TGFβ

✉ Benjamin C. M. Fung
ben.fung@mcgill.ca

✉ Jean-Jacques Lebrun
jj.lebrun@mcgill.ca

Jimin Guo
jimin@hms.harvard.edu

Farkhund Iqbal
farkhund.iqbal@zu.ac.ae

Peter J. K. Kuppen
p.j.k.kuppen@lumc.nl

Rob A. E. M. Tollenaar
r.a.e.m.tollenaar@lumc.nl

Wilma E. Mesker
w.e.mesker@lumc.nl

[1] Division of Medical Oncology, McGill University Health Center, Montreal, QC, Canada

[2] Present address: Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[3] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

[4] School of Information Studies, McGill University, Montreal, QC, Canada

[5] Zayed University, Abu Dhabi, United Arab Emirates

[6] Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands

1 Introduction

Following surgery, breast cancer patients often receive adjuvant therapy for at least 5 years, as a precaution against relapse. However, many patients experience therapeutic failure marked by disease recurrence, in the form of local and regional relapses or distant metastasis (Carlson 2010). These recurrent disease lesions often occur within the period of adjuvant therapy, and indicate that residual tumor cells do not respond to adjuvant therapy or have weak responses (Carlson 2010). If recurrence does not happen during the administration of adjuvant therapy, the incidences of recurrence afterwards tend to be sporadic (Carlson 2010; Brewster et al. 2008). These

sporadic incidences are believed to be due to tumor cells exiting dormancy over time. While late disease recurrence is indicative of some levels of responses to the adjuvant therapy, early breast cancer recurrence poses serious threat to patients' lives. As such, methods that predict whether or not breast cancer patients will develop early recurrence, using disease attributes collected at the time of initial diagnosis, could prove very useful to help determine disease prognosis and the making of clinical decisions.

A widely-explored approach to develop prediction models is to calculate an arbitrary prognostic score established by multivariate regression models, using disease characteristics, immunohistochemistry, gene expression profiles, alone or in combination (Campbell et al. 2010; Galea et al. 1992; Zhang et al. 2013; Barton et al. 2012; Parisi et al. 2010). While these regression models are well curated, they also have a few limitations. They take account of every case in the same manner, even when dealing with highly heterogeneous populations. Moreover, the good performance of a regression model depends on carefully selecting relevant disease characteristics, thus requiring extensive prior knowledge. In addition, it is not always possible to interpret the contribution of individual characteristics from the mathematical formula describing the model. Finally, regression models yield either a score related to an outcome or a probability of an outcome, rather than the outcome itself. To overcome these limitations, we use an alternative approach to build a Decision Tree classifier. The classifier groups patients based on similar disease attributes and outcomes, list the disease attributes in a hierarchical order based on their relevance to the outcomes, and predict for the status of whether or not a patient will develop early breast cancer relapse.

The principle of a Decision Tree algorithm is to continuously partition a group of heterogeneous examples, using the values of several descriptors (feature attributes), to obtain subgroups that are homogeneous of pre-defined classes (class attribute) (Lee and Hsu 1990; Quinlan 1993). As dividing a group based on a feature attribute results in at least two subgroups, which are relatively more homogeneous than the parental group, a decrease in system disorder (or entropy) can be calculated using a probability-based formula and denoted as Information Gain. Partitioning the examples using a feature attribute with higher Information Gain results in a better-organized system with respect to the class attribute. As such, Information Gain serves as a criteria to evaluate the relevance between individual feature attributes and the class attribute (Quinlan 1993; Mitchell 1997). The algorism iteration starts by partitioning examples using the feature attributes that yield the biggest Information Gain; and stops when a subgroup is homogeneous or when the Information Gain of remaining attributes falls below a certain threshold (Lee and Hsu 1990; Quinlan 1993). This results in a tree-like structure with the feature attributes showing as the "branches" and the subgroups showing as the "leaves". By tracing the feature attributes of an incoming example, one can make a prediction for the status of the target attribute of that example.

We were particularly interested in using a Decision Tree classifier to study whether stroma percentage and TGFβ signaling biomarkers are relevant to early breast cancer recurrence. Piling studies using regression models show that these factors have different or even contrasting associations with breast cancer recurrence in subgroups of patients. As such, their implications in breast cancer pathology are context-dependent. However, these contexts remain to be defined in a systematic manner. By grouping patients based on their similar outcomes and disease characteristics, a Decision Tree classifier is capable to achieve this goal.

Stroma percentage in tumor core is an emerging prognostic indicator for several types of cancer (Gujam et al. 2014; Downey et al. 2014; Huijbers et al. 2013; Moorman et al. 2012; de Kruijf et al. 2011). In breast cancer, the prognostic value of stroma percentage is context dependent. While high stroma percentage is associated with shorter times of relapse-free survival and overall survival in triple negative breast cancer patients (Moorman et al. 2012), it is associated with longer times of relapse-free survival and overall survival among patients with ER+ breast tumors (Downey et al. 2014). In a mixed population of various subtypes, intra-tumor stroma loses its prognostic value, as determined by a multivariate analysis (Ahn et al. 2012), likely because this method fails to highlight differences within a highly heterogeneous population.

The canonical TGFβ/Smad pathway is also implicated in breast cancer pathology in a context-dependent manner (Massague 2008; Lebrun 2012). In normal mammary gland and early stage, low-grade breast carcinomas, TGFβ functions to maintain homeostasis and this effect is largely due to its growth-inhibitory and pro-apoptotic functions (Mazars et al. 1995). However, in advanced-stage breast tumors, TGFβ promotes aggressive behaviors such as cell migration, cell invasion and homing at distant metastatic sites (Muraoka et al. 2002; Padua et al. 2008). Binding of the TGFβ ligand to its two serine/threonine kinase receptors, results in the recruitment and subsequent activation of specific downstream signaling molecules, called Smads (Smad2, 3 and 4), which then translocate to the nucleus to regulate gene transcription (Shi and Massague 2003).

The Decision Tree classifier that we generated identifies the status of lymph node involvement, intra-tumor stroma percentage, and percentages of tumor cells expressing components of TGFβ-Smad signalling to be highly relevant to the status of early breast cancer relapse. It is estimated to be about 70% accurate, and correctly predicted for 55 out of 65 patients in an independent validation dataset.

2 Materials and methods

2.1 Dataset

The dataset contained the following types of information of 574 patients of non-metastatic invasive breast cancer who received surgeries in the Leiden University Medical Center: age, pathological grade, TNM (tumor, node, metastasis) stage, local and systemic therapy, recurrence status (local, regional and distant), time of recurrence following initial treatment and overall survival. Tumor cores were subjected to Haematoxylin and Eosin (H&E) staining for scoring percentages of intra-tumor stroma by two investigators. In addition, percentages of cells expressing the following factors were determined by standard immunohistochemistry procedure: estrogen receptor (ER), progesterone receptor (PgR), epidermal growth factor receptor 2 (HER2) and Ki-67. A tissue microarray (TMA) was constructed from tumor cores of these patients, subjected to immunohistochemistry (IHC) and scored for percentages of tumor cells expressing the following factors: TGFβ type I and type II receptors (TGFBR1 and TGFBRII, respectively), nuclear Smad4 and nuclear phospho-Smad2.

Details on the patient cohort, methods of stroma percentage scoring and materials and methods of IHC are reported in previous studies (de Kruijf et al. 2011, 2013; Dekker et al. 2013). These procedures are in accordance with those listed in REporting recommendations for tumour MARKer prognostic studies (REMARK) (LM et al. 2005). Names and brief descriptions of the attributes used are included in Table 1.

2.2 Decision tree

We defined the class attribute as the status of breast cancer recurrence in the first 3 years after diagnosis (disease free or tumor recurred). We arbitrarily chose this endpoint as these patients had minimal benefit from adjuvant therapy. Therefore, their disease outcomes help to predict for patients who likely do not respond to adjuvant therapy.

We used 55 breast cancer disease characteristics as feature attributes (Table 1). Most of them are well-established disease characteristics, used by physicians worldwide to describe breast tumors and form treatment plans such as pathological grades, clinical stages and expression of molecular markers. In addition, we also included several characteristics whose roles in breast cancer recurrence are controversial, as determined by linear regression methods. These characteristics include stroma percentage in tumor core and percentage of cells expressing TGFβ signaling components.

We used Rapidminer 6.0 to implement the Decision Tree. Rapidminer's Decision Tree operator is derived from Quinlan's C4.5 Decision Tree (Quinlan 1993). We chose to rank attributes based on Information Gain-Ratio. This is a modified Information Gain method, which normalizes Information Gains of all attributes to minimize bias towards attributes that contain large numbers of unique values (distinctive yet non-relevant information) (Mitchell 1997). We set the minimum size of split as 4, minimum leaf size as 2, the minimum gain ratio to split with an attribute as 0.1. We grew the tree for up to 10 steps and do post-pruning.

2.3 Estimation of accuracy

To estimate the accuracy of the Decision Tree classifier during the building step, we coupled the model building process with 2 different resampling validation methods: 10-fold bootstrapping validation with a 0.9 sampling ratio and 10-fold cross validation with stratified sampling. As such, these two methods are comparable, that each round of repeating validation uses 90% of the available data to build a model and then uses the remaining to test the accuracy of the model. Results show an estimated accuracy with standard deviation obtained from the 10 slightly different models.

2.4 Validation after model building

In the model building process, we excluded a dataset of 69 patients with missing Smad4 values (Smad4 null). This dataset served as an independent validation dataset, because it was excluded from model building and estimation processes. We eliminated 4 patients in this cohort, as they died within 3 years of diagnosis but did not develop disease recurrence. The prevalence rates of early breast cancer relapse in the original cohort, in the cohort that we used to build the classifier and in the Smad4 null cohort were comparable as 22.47%, 23.4% and 23.08%, respectively. Using the standard truth table, we calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) with predictions for patients of this dataset.

3 Results

3.1 Data pre-cleaning

The original dataset contained missing values for every TGFβ signaling components, due to tissue falling off from slides during the immunohistochemistry process. To maximize the utilization of real data in the algorism training process, we did not fill missing values. Instead, we excluded 69 cases that had missing values for nuclear Smad4 (Smad4 null), since this attribute contained the most missing values. We then combined local recurrence, regional recurrence and distant relapse into one status, and defined the time of recurrence as the earliest time when any of the recurrence event occurred.

We further eliminated 22 patients who did not develop disease recurrence but died of non-breast cancer causes within

Table 1 A list of 55 attributes used as inputs of the Decision Tree

Attribute	Descriptions	Type	Used by classifier
age	age	numerical	*
BR	total score of grading (sum of BR_MAI, BR_duct and BR_kern)	integer	*
BR_duct	percentage of duct	integer	
BR_kern	nuclear atypia grade	integer	
BR_MAI	amount of mitogen	integer	*
CARatio	carcinoma ratio	integer	
cMstag	clinical M stage	integer	
cNStag	clinical N stage	integer	*
cNstag2	clinical nodal stage in 5 scales	integer	
CT	chemotherapy	binominal	
cTStag	clinical T stage	integer	*
cTstag2	clinical tumor stage	integer	
ER	Estrogen Receptor (ER) status	binominal	
ER_percpos	mean percentage of ER stained	numerical	*
ER10	histological status of ER, border is 10	binominal	
ER80	histological status of ER, border is 80	binominal	*
ExtC	clinical tumor type (early or local advanced)	binominal	*
ExtP	pathological tumor type (early or local advanced)	binominal	*
GR	tumor grade (I, II and III)	integer	
GR_2	tumor grade (I/II and III)	integer	
Gra	tumor grade (I/II and III)	binominal	
Her2	Her2 status(normal and overexpressed)	binominal	
Her2_M	mean Her2 histological score of 3 cores	numerical	*
HT	Hormonal therapy	binominal	
inv	status of lymphangioinvasion	binominal	
IT	Immunotherapy	binominal	
Ki67_0	histological score of Ki67, border is 0	binominal	
Ki67_10	histological score of Ki67, border is 10	binominal	
Ki67_5	histological score of Ki67, border is 5	binominal	
Ki67_Mean	mean of Ki67	numerical	*
loct2	surgery type (mastectomy or breast-conserving surgery)	binominal	
loct	surgery type with radiotherapy (MAST + RT, MAST-RT or BCS)	polynominal	
MAI_Gr	mitogen grade	integer	
OK	status of receiving surgery	binominal	
PgR	status of Progestrone Receptor (PgR), 10% and above = positive	binominal	
PgR_percpos	mean percentage of PgR stained	numerical	*
PgR10	histological status of PgR, border is 10	binominal	
PgR80	histological status of PgR, border is 80	binominal	
pMstag	pathological M stage	integer	
pN2	pathological nodal stage on 2 scales	binominal	
pNStag	pathological N stage	integer	
pNstag2	pathological nodal stage on 5 scales	integer	
pSmad2_nt_perc	percentage of cells expressing nuclear phospho-Smad2	integer	*
pT3	pathological T stage into 3 groups (T1, T2 and T3/4)	integer	
pTstag	pathological T stage	integer	*
pTstag2	pathological tumor stage on 3 scales (pT1, pT2 and other)	integer	
RT	radiotherapy	binominal	
smad4_perc	percentage of cells expressing nuclear Smad4	numerical	*
stroma_perc	stroma percentage in the tumor core	numerical	*

Table 1 (continued)

Attribute	Descriptions	Type	Used by classifier
Surg	type of surgery (mastectomy or conserving surgery)	binominal	*
T_type	tumor type (ductal, lobular and other)	polynominal	
T_Type2	tumor type in 2 status(ductal and other)	binominal	
T2	tumor stage in 2 status (T1/2 and T3/4)	integer	
TGFRI_perc	percentage of cells expressing TGFbRI	numerical	*
TGFRII_perc	percentage of cells expressing TGFbRII	numerical	*

The Decision Tree classifier identified 19 of them to be relevant to early breast cancer recurrence (marked with asterisk)

3 years. Data pre-cleaning resulted in a dataset of 487 examples with less than 10% missing values for nuclear Smad2, TGFβ type I and type II receptors. The missing values of each attribute were then filled with the average of known values of that attribute.

We assigned one of the following attribute types to each of the 55 feature attributes. Numerical attributes contain values of real numbers. Nominal attributes contain values of a category. Integer attributes, such as clinical and pathological stages, are orderly nominal attributes and therefore also have a numerical nature. Of the target attribute (3-year relapse), we assigned a binominal value for each patient. Patients who were disease-free received 0, and patients who developed relapse received 1.

3.2 Performance of the decision tree classifier

We generated a Decision Tree classifier to predict for breast cancer recurrence within 3 years of the initial diagnosis, using a patient dataset containing information on clinical diagnosis, pathological diagnosis, stroma percentage and expression of TGFβ signaling components (Table 1). The Decision Tree operator nested with bootstrapping validation or cross validation generated similar tree structures and similar estimated accuracy, even if the sampling methods differed. Table 2 shows the estimated accuracy, estimated sensitivity (class recall) and estimated specificity (class precision) of the classifiers. Furthermore, Bayesian Boosting, which generated 9 additional tree structures every round during model building to vote for consensus, did not remarkably improve model accuracy (data not shown). We also found that growing the tree to the depth of 10 was ideal for this dataset. Neither growing the tree deeper nor not pruning the tree changed the major structure of the tree (data not shown). Altogether, these results suggest that the classifiers that we obtained captured major properties of the dataset.

Figure 1 shows the decision tree validated by cross validation. The classifier presents patients in 66 leaves. Each leaf represents a subset of breast cancer patients with similar disease characteristics. Even though 2 different leaves could have the same patient outcome, each leaf is independent and can be summarized with a distinct subset of attributes. As such, the classifier grouped breast cancer patients into different subsets based on their intrinsic properties.

Out of 66 leaves in total, 60 leaves contained patients only with or only without recurrence (no mix), indicating that in most cases, the combined attributes that describe a group of patients were sufficient to predict for a finite outcome. Six leaves contained mixed populations of patients, indicating that for these subgroups, additional attributes are required to further distinguish the disease-free and disease-recurred status.

Independent validation of the classifier's performance was achieved using a set of 65 patients for whom the values for the Smad4 attribute were missing. In the event that a prediction process reaches a branch with a missing Smad4 value (or any other missing value), the classifier assigned the consensus results of all lower branches as the final prediction. Interestingly, the classifier predicted correctly for 55 out of the 65 patients, an accuracy of 85%. Table 3 summarizes the predictions, sensitivity, specificity likelihood ratios and predictive values. For the disease-relapsed status, the classifier achieved 40% sensitivity (95% CI: 16.43% -

Table 2 Estimated accuracy of the decision tree classifier

	true disease-free	true recurred	Precision
predicted disease-free	298	77	79.47%
predicted recurred	75	37	33.04%
Recall	79.89%	32.46%	
Cross Validation: 68.8−/+6.3% mikro 68.79%			
predicted disease-free	1183	297	79.93%
predicted recurred	333	178	34.83%
Recall	78.03%	37.47%	
Bootstrapping Validation: 68.28−/+4.63% mikro 68.36%			

The accuracy of the classifier was estimated with cross-validation (top) and with bootstrapping validation (bottom). For each class, the performance was also evaluated with Precision (percentage of the predictions that are correct) and Recall (percentage of an outcome that is correctly predicted)

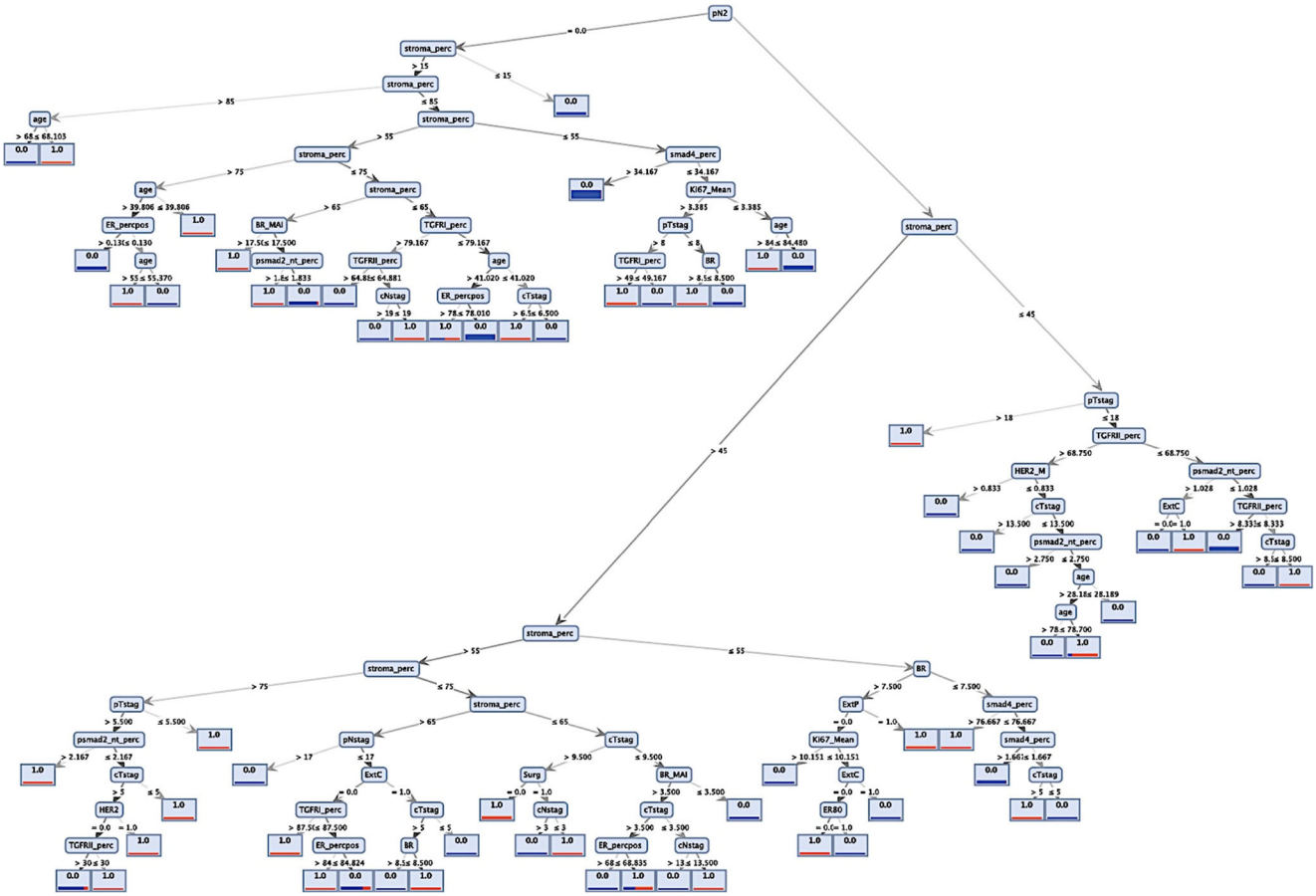

Fig. 1 Structure of the Decision Tree classifier. A cohort of 483 patients was continuously divided into 67 subgroups, based on intrinsic similarities of their diseases. The branches of the tree showed the disease characteristics used to divide the patients. And each subgroup was labeled with the outcomes of the patients 3 years after diagnosis: 1 as recurrence and 0 as disease-free

67.67%). For the disease-free status, the classifier achieved 92% specificity (95% CI: 80.75% - 97.73%). Respectively, these values are notably higher than the penetrance (23.08%) and percentage of disease free patients (76.92%). These results suggest that the classifier was capable of distinguishing disease outcomes for most patients in the independent validation set.

3.3 Pathological nodal stage, stroma percentage and TGFβ signaling are predictive attributes of early breast cancer recurrence

Among the 55 attributes that we used, the Decision Tree classifier selectively presented 20 disease attributes on 9 levels. These attributes are marked with an asterisk (*) in Table 1.

Table 3 Predictions for the Smad4 null dataset (top) and a truth table showing the performance of this prediction (bottom)

Predictions:		
	true disease-free	true recurred
Predicted disease-free	46	6
Predicted recurred	4	9
Results:		
Sensitivity	40.00%	95% CI: 16.43% - 67.67%
Specificity	92%	95% CI: 80.75% - 97.73%
Positive Likelihood Ratio	5	95% CI: 1.62–15.42
Negative Likelihood Ratio	0.65	95% CI: 0.43–0.99
Positive Predictive Value	60%	95% CI: 26.37% - 87.6%
Negative Predictive Value	83.64%	95% CI: 71.19% - 92.22%
Prevalence: 23.08%		

The structure captured several well-documented traits of breast cancer recurrence. The first attribute used to divide patients was the status of lymph node involvement (pN2), highlighting lymph node positivity as the most relevant attribute to early breast cancer recurrence. The classifier splits patients into 2 groups; defined as pN2 = 0 (not spreading to lymph node) and pN2 = 1 (containing all patients with lymph node involvement, regardless of the level of involvement). This is highly consistent with clinicians' emphasis on lymph node involvement when making prognosis for breast cancer recurrence.

In addition, we also observed that stroma percentage (Fig. 1, stroma_perc) was the only secondary attribute appearing on both branches, following the division based on pathological node status. This indicates that, alongside lymph node status, stroma percentage was an utmost relevant attribute for all cases. For both branches, the classifier divided patients into multiple groups based on stroma percentage, suggesting that tumor-stroma interaction levels define different subgroups of breast tumors, with respect to early breast cancer relapse. Notably, the classifier identified a subgroup of 11 disease-free patients who had no lymph node involvement (pN2 = 0) and low stroma percentage (stroma_perc = 0% or 10%). This is consistent with the notion that patients with low grade, well-encapsulated tumors tend not to develop early relapse (Esposito et al. 2009).

Aside from lymph node status and stroma percentage, the classifier also highlighted several molecular characteristics, commonly used in the clinic for defining breast cancer subtypes and prognosis, as being determinant for status of early breast cancer relapse. These include the Estrogen Receptor α (ER_percpos), Progesterone Receptor (PgR_percpos), HER2, Ki67 (Ki67_Mean) and clinical tumor stage (CTstag) (Table 1 and Fig. 1). In multiple branches of the tree structure, we also found TGFβ receptors (type I and type II) as well as nuclear Smad4 and phospho-Smad2. In particular, TGFβ receptor II and Smad4 were the third level attributes of their respective branches. These results highlight the subgroup-specific prognostic values of TGFβ signaling components. Equally importantly, these results also indicate that TGFβ signaling components are better attributes than many of the commonly used clinical criteria (those not shown in the tree, Table 1) when predicting for early breast cancer relapse.

4 Discussions

In this study, we took a data mining approach to generate a Decision Tree classifier that can predict for breast cancer relapse status within the first 3 years following diagnosis. The tree classified patients into disease-free or disease-relapsed categories. The tree subdivided patients, using disease characteristics that display a defined and relevant threshold for disease recurrence (Information Gain Ratio = 0.1), and listed these characteristics in hierarchy order. As such, the model building process was also a "feature selection" process that helped identify important disease characteristics.

The classifier identified pathological nodal status as the most relevant feature to disease recurrence. While we supplied 3 different ways to categorize lymph node statues to the algorism, including pN2 (binary attribute of lymph node involvement), pNstag2 (integer attribute denoting pN0, pN1, pN2, pN3 and pNx), pNstag (integer attribute further subdividing each pNstag2 stage), the Decision Tree classifier identified pN2 as the only attribute among the three that was relevant to early breast cancer relapse. This indicates that, lymph node involvement is relevant to predicting early breast cancer relapse, independently, of the number of nodes involved. This is also highly consistent with the longstanding notion that pathological lymph node status is the most significant predictor of breast cancer recurrence (Aubele et al. 1995). As such, this fact validates the capacity of the Decision Tree classifier to identify and hierarchically present important features in our dataset.

Stroma percentage showed as the only second level attribute of all branches while TGFβ signaling components showed in various branches on lower levels. Current literature suggests that stroma percentage and TGFβ signaling components are relevant to breast cancer recurrence, but their predictive values differ, or even contrast, depending on the context. The Decision Tree classifier not only identified these attributes to be highly relevant, but also provided detailed description of the individual contexts.

With respect to the model performance, the classifier achieved over 80% precision for predicting a disease-free status, but only 34.15% recall for predicting early recurrence. This suggests that additional attributes are needed to better describe patients with early recurrence. Potentially, including immunohistochemistry scores of additional oncogenic or tumor suppressive signaling pathways, such as those of PI3K-AKT-mTOR, EGFR, p53 or Rb, could improve the classifier.

Nevertheless, the performance is comparable and potentially better than existing methods. For the Smad4 null independent validation set, the Decision Tree classifier predicted correctly for 85% of the patients in the Smad4 null validation set. In particular, 40% of the patients predicted to have early relapse within 3 years indeed had relapse. By comparison, another study using the Breast Cancer Index (BCI), a well-curated method to predict outcomes for ER+, lymph node negative (LN-) patients, classified patients in 3 groups of increasing risks of distant recurrence; using a combination of HOXB13:IL17BR gene expression ratio and molecular grade index (Jerevall et al. 2011; Ma et al. 2008). In 2 different

patient cohorts, the estimated percentage of patients classified into high-risk group by BCI, and developed distant relapse within 5 years are 2.6%–21% and 14.6–33.3%, respectively (Zhang et al. 2013). BCI and the Decision Tree classifier have different advantages. BCI is capable of predicting for distant relapse and overall survival for various endpoints, but only for ER+, LN- patients. The Decision Tree classifier can be applied to all types of patients but predicts for 3-year relapse as its current design stands. However, predicting for other endpoints can be easily done, as it only requires creating a new target attribute for that endpoint. As such, the Decision Tree classifier could potentially be a powerful prognostic tool. Especially, the classifier can be easily adopted in different academic and clinical settings, as the attributes that we used are empirical and easy to assess. All nominal attributes, such as stage and grade, are assessed based on established quantitative methods in clinical practice at the time of diagnosis. All numeric attributes are established from quantitative immunohistochemistry staining.

In summary, we generated a Decision Tree classifier that hierarchically organizes breast cancer disease characteristics based on their relevance to early breast cancer relapse. One can easily trace down the tree structure to obtain the description of the intrinsic similarity of each subgroup of patients. The classifier also highlights the prognostic values of pathological nodal status, stroma percentage and TGFβ signaling components. To our knowledge, this is the first Decision Tree model that utilizes standardized disease characteristics that can be easily obtained by different clinics.

Acknowledgements We would like to thank Drs. C. C. Engels, J. W. T. Dekker and E. M. de Kruijf for conducting immunohistochemistry staining, evaluating stroma percentage and recording original data; and Drs. A. Dibrov and Catalin Mihalcioiu for valuable discussions. J. Guo is supported by a Traineeship from the Breast Cancer Research Program of Congressionally Directed Medical Research Program (CDMRP). B. C. M. Fung is a Canada Research Chair in Data Mining for Cybersecurity. J.-J. Lebrun is a Sir William Dawson Research Chair of McGill University. This work was supported in part by grants from the Canadian Institutes for Health Research (CIHR) (fund codes 230670 and 233716 to J.-J. Lebrun), the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (fund code 356065-2013 to B. C. M. Fung), Canada Research Chairs Program (fund code 950-230623 to B. C. M. Fung), and Zayed University Research Incentive Fund and Research Cluster Award (fund codes R15048 and R16083 to F. Iqbal and B. C. M. Fung).

Contribution J. Guo, B. C. M. Fung and J.-J. Lebrun designed the study, analyzed and interpreted the results. F. Iqbal participated in interpreting the results. P. J. K. Kuppen, R. A. E. M. Tollenaar and W. E. Mesker collected patient samples and designed the tumor tissue microarrays.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ahn, S., Cho, J., Sung, J., Lee, J. E., Nam, S. J., Kim, K. M., & Cho, E. Y. (2012). The prognostic significance of tumor-associated stroma in invasive breast carcinoma. *Tumour biology : the Journal of the International Society for Oncodevelopmental Biology and Medicine, 33*(5), 1573–1580.

Aubele, M., Auer, G., Voss, A., Falkmer, U., Rutquist, L., & Hofler, H. (1995). Disease-free survival of node-positive breast-cancer patients - improved prognostication by cytometric parameters. *Pathology, Research and Practice, 191*(10), 982–990.

Barton, S., Zabaglo, L., A'Hern, R., Turner, N., Ferguson, T., O'Neill, S., Hills, M., Smith, I., & Dowsett, M. (2012). Assessment of the contribution of the IHC4+C score to decision making in clinical practice in early breast cancer. *British Journal of Cancer, 106*(11), 1760–1765.

Brewster, A. M., Hortobagyi, G. N., Broglio, K. R., Kau, S. W., Santa-Maria, C. A., Arun, B., Buzdar, A. U., Booser, D. J., Valero, V., Bondy, M., & Esteva, F. J. (2008). Residual risk of breast cancer recurrence 5 years after adjuvant therapy. *Journal of the National Cancer Institute, 100*(16), 1179–1183.

Campbell, H. E., Gray, A. M., Harris, A. L., Briggs, A. H., & Taylor, M. A. (2010). Estimation and external validation of a new prognostic model for predicting recurrence-free survival for early breast cancer patients in the UK. *British Journal of Cancer, 103*(6), 776–786.

Carlson, R. (2010). Surveillance of patients following primary therapy. In *Diseases of the breast*, 4 edn. Lippincott Williams and Wilkins.

de Kruijf, E. M., van Nes, J. G., van de Velde, C. J., Putter, H., Smit, V. T., Liefers, G. J., Kuppen, P. J., Tollenaar, R. A., & Mesker, W. E. (2011). Tumor-stroma ratio in the primary tumor is a prognostic factor in early breast cancer patients, especially in triple-negative carcinoma patients. *Breast Cancer Research and Treatment, 125*(3), 687–696.

de Kruijf, E. M., Dekker, T. J., Hawinkels, L. J., Putter, H., Smit, V. T., Kroep, J. R., Kuppen, P. J., van de Velde, C. J., ten Dijke, P., Tollenaar, R. A., & Mesker, W. E. (2013). The prognostic role of TGF-beta signaling pathway in breast cancer patients. *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO, 24*(2), 384–390.

Dekker, T. J., van de Velde, C. J., van Pelt, G. W., Kroep, J. R., Julien, J. P., Smit, V. T., Tollenaar, R. A., & Mesker, W. E. (2013). Prognostic significance of the tumor-stroma ratio: Validation study in node-negative premenopausal breast cancer patients from the EORTC perioperative chemotherapy (POP) trial (10854). *Breast Cancer Research and Treatment, 139*(2), 371–379.

Downey, C. L., Simpkins, S. A., White, J., Holliday, D. L., Jones, J. L., Jordan, L. B., Kulka, J., Pollock, S., Rajan, S. S., Thygesen, H. H., Hanby, A. M., & Speirs, V. (2014). The prognostic significance of tumour-stroma ratio in oestrogen receptor-positive breast cancer. *British Journal of Cancer, 110*(7), 1744–1747.

Esposito, N. N., Dabbs, D. J., & Bhargava, R. (2009). Are encapsulated papillary carcinomas of the breast in situ or invasive? A basement membrane study of 27 cases. *American Journal of Clinical Pathology, 131*(2), 228–242.

Galea, M. H., Blamey, R. W., Elston, C. E., & Ellis, I. O. (1992). The Nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment, 22*(3), 207–219.

Gujam, F. J., Edwards, J., Mohammed, Z. M., Going, J. J., & McMillan, D. C. (2014). The relationship between the tumour stroma percentage, clinicopathological characteristics and outcome in patients with operable ductal breast cancer. *British Journal of Cancer, 111*(1), 157–165.

Huijbers, A., Tollenaar, R. A., van Pelt, G. W., Zeestraten, E. C., Dutton, S., McConkey, C. C., Domingo, E., Smit, V. T., Midgley, R., Warren, B. F., Johnstone, E. C., Kerr, D. J., & Mesker, W. E.

(2013). The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: Validation in the VICTOR trial. *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO, 24*(1), 179–185.

Jerevall, P. L., Ma, X. J., Li, H., Salunga, R., Kesty, N. C., Erlander, M. G., Sgroi, D. C., Holmlund, B., Skoog, L., Fornander, T., Nordenskjold, B., & Stal, O. (2011). Prognostic utility of HOXB13:IL17BR and molecular grade index in early-stage breast cancer patients from the Stockholm trial. *British Journal of Cancer, 104*(11), 1762–1769.

Lebrun, J. J. (2012). The dual role of TGF in human cancer: From tumor suppression to cancer metastasis. *ISRN Molecular Biology, 2012*, 1–28.

Lee, H. M., & Hsu, C. C. (1990). *A new model for concept classification based on linear threshold unit and decision tree*. Proceedings of the International Joint Conference on Neural Networks (IJCNN-90-Wash D.C. IEEE/INNS), Washington, D.C., USA, vol. 2, pp. 631–634.

LM, M. S., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., Clark, G. M., & Statistics Subcommittee of the NCIEWGoCD. (2005). REporting recommendations for tumour MARKer prognostic studies (REMARK). *European Journal of Cancer, 41*(12), 1690–1696.

Ma, X. J., Salunga, R., Dahiya, S., Wang, W., Carney, E., Durbecq, V., Harris, A., Goss, P., Sotiriou, C., Erlander, M., & Sgroi, D. (2008). A five-gene molecular grade index and HOXB13:IL17BR are complementary prognostic factors in early stage breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 14*(9), 2601–2608.

Massague, J. (2008). TGFbeta in cancer. *Cell, 134*(2), 215–230.

Mazars, P., Barboule, N., Baldin, V., Vidal, S., Ducommun, B., & Valette, A. (1995). Effects of TGF-beta 1 (transforming growth factor-beta 1) on the cell cycle regulation of human breast adenocarcinoma (MCF-7) cells. *FEBS Letters, 362*(3), 295–300.

Mitchell, T. M. (1997). *Machine learning*. The McGraw-Hill Companies, Inc., New York.

Moorman, A. M., Vink, R., Heijmans, H. J., van der Palen, J., & Kouwenhoven, E. A. (2012). The prognostic value of tumour-stroma ratio in triple-negative breast cancer. *European Journal of Surgical Oncology : the Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 38*(4), 307–313.

Muraoka, R. S., Dumont, N., Ritter, C. A., Dugger, T. C., Brantley, D. M., Chen, J., Easterly, E., Roebuck, L. R., Ryan, S., Gotwals, P. J., Koteliansky, V., & Arteaga, C. L. (2002). Blockade of TGF-beta inhibits mammary tumor cell viability, migration, and metastases. *The Journal of Clinical Investigation, 109*(12), 1551–1559.

Padua, D., Zhang, X. H., Wang, Q., Nadal, C., Gerald, W. L., Gomis, R. R., & Massague, J. (2008). TGFbeta primes breast tumors for lung metastasis seeding through angiopoietin-like 4. *Cell, 133*(1), 66–77.

Parisi, F., Gonzalez, A. M., Nadler, Y., Camp, R. L., Rimm, D. L., Kluger, H. M., & Kluger, Y. (2010). Benefits of biomarker selection and clinico-pathological covariate inclusion in breast cancer prognostic models. *Breast Cancer Research, 12*(5), R66.

Quinlan, J. R. (1993). *C4.5 : programs for machine learning*. San Mateo: Morgan Kaufmann Publishers.

Shi, Y., & Massague, J. (2003). Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell, 113*(6), 685–700.

Zhang, Y., Schnabel, C. A., Schroeder, B. E., Jerevall, P. L., Jankowitz, R. C., Fornander, T., Stal, O., Brufsky, A. M., Sgroi, D., & Erlander, M. G. (2013). Breast cancer index identifies early-stage estrogen receptor-positive breast cancer patients at risk for early- and late-distant recurrence. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 19*(15), 4196–4205.

Jimin Guo is a Research Fellow in Biomedical Informatics at Harvard Medical School and Harvard School of Engineering and Applied Sciences. He received graduate training in Cancer Biology at McGill University. His research focuses on understanding cellular and molecular mechanisms of cancer metastasis and relapse. And he also develops high-throughput and high-resolution transcriptome profiling methods. He is an associate member of the American Association for Cancer Research.

Benjamin C. M. Fung is the Canada Research Chair in Data Mining for Cybersecurity and an associate professor in the School of Information Studies (SIS), McGill University. He has over 100 refereed publications that span the research forums of data mining, privacy protection, cyber forensics, services computing, and building engineering. His data mining works in crime investigation and authorship analysis have been reported by media worldwide. He is a licensed professional engineer in software engineering. He is a senior member of the IEEE and ACM.

Farkhund Iqbal holds the position of Associate Professor and Graduate Program Director in the College of Technological Innovation. He holds a Master (2005) and a Ph.D. degree (2011) from Concordia University, Canada. He has served as a chair and TPC member of several IEEE/ACM conferences and is the reviewer of high rank journals. He is the member of several professional organization including ACM and IEEE Digital society. He has published more than 50 papers in high impact factor journals and conferences.

Peter J. K. Kuppen is a biomedical researcher, with a focus on tumor-immune interactions, especially in breast cancer and colorectal cancer. His work involves large panels of human tumor tissues documented with clinical follow-up data, enabling to make a link between parameters that determine tumor-immune interaction and clinical outcome.

Rob A. E. M. Tollenaar is the head of the Department of Surgery, LUMC, Leiden. In this position he is responsible for patient care, education and research within the surgical oncology field. His expertise is in diagnostic and prognostic factors in colon-, pancreatic- and breast cancer within the field of proteomics and tumor-stromal interactions. He is one of the founders and president of the scientific committee of the DSCA (Dutch Surgical Clinical Auditing).

Wilma E. Mesker is an associate professor at the Department of Surgery, LUMC, Leiden. Her research field relates to the development and application of novel technology in translational areas, as in cancer screening. She has expertise in diagnostic and prognostic factors in colon-, pancreatic- and breast cancer within the field of proteomics and tumor-stromal interactions, minimal residual disease in bone marrow and lymph nodes, DNA analysis.

Jean-Jacques Lebrun holds the ranks of Full Professor of Medicine and Associate Dean of graduate and postdoctoral studies at McGill University, Canada. Dr Lebrun's research program aims at understanding processes underlying tumor suppression, cancer stem cells, drug resistance, metastasis and tumor relapse, all considered major challenges in the management of cancer patients. Using several central lines of research utilizing various cellular model systems (in vitro), pre-clinical studies using Xenograft mouse models (in vivo), human cancer specimens as well as bioinformatics analyses of large human tumors datasets, Dr Lebrun's research has expanded our knowledge of the molecular mechanisms that drive breast cancer invasion, metastasis, drug resistance and cancer stemness. Professor Lebrun's leadership is further reflected by his high-level publication record, successful funding and frequent invitations to International meetings.