

A new methodology to predict the oncotype scores based on clinico-pathological data with similar tumor profiles

Zeina Al Masry¹ · Romain Pic² · Clément Dombry² · Christine Devalland³

Abstract

Purpose The Oncotype DX (ODX) test is a commercially available molecular test for breast cancer assay that provides prognostic and predictive breast cancer recurrence information for hormone positive, HER2-negative patients. The aim of this study is to propose a novel methodology to assist physicians in their decision-making.

Methods A retrospective study between 2012 and 2020 with 333 cases that underwent an ODX assay from three hospitals in the Bourgogne Franche-Comté region (France) was conducted. Clinical and pathological reports were used to collect the data. A methodology based on distributional random forest was developed to predict the ODX score classes ($ODX \leq 25$ and $ODX > 25$) using 9 clinico-pathological characteristics. This methodology can be used particularly to identify the patients of the training cohort that share similarities with the new patient and to predict an estimate of the distribution of the ODX score.

Results The mean age of participants is 56.9 years old. We have correctly classified 92% of patients in low risk and 40.2% of patients in high risk. The overall accuracy is 79.3%. The proportion of low risk correct predicted value (PPV) is 82%. The percentage of high risk correct predicted value (NPV) is approximately 62.3%. The F1-score and the Area Under Curve (AUC) are of 0.87 and 0.759, respectively.

Conclusion The proposed methodology makes it possible to predict the distribution of the ODX score for a patient. This prediction is reinforced by the determination of a family of known patients with follow-up of identical scores. The use of this methodology with the pathologist's expertise on the different histological and immunohistochemical characteristics has a clinical impact to help oncologist in decision-making regarding breast cancer therapy.

Keywords Breast Cancer · Oncotype DX · Clinico-pathological data · Machine Learning · Distributional Random Forest

Zeina Al Masry
E-mail: zeina.almasry@femto-st.fr

¹ SUPMICROTECH, CNRS, institut FEMTO-ST, 24 rue Alain Savary, Besançon, F-25000, France.

² Université de Franche-Comté, CNRS, LmB, F-25000 Besançon, France.

³ Service d'anatomie et cytologie pathologiques, Hôpital Nord Franche-Comté, 100 Route de Moval, 90400 Trévenans, France.

1 Introduction

The Oncotype DX (ODX) test is a commercially available molecular test for breast cancer assay (Genomic Health) that provides prognostic and predictive breast cancer recurrence information for hormone positive, HER2-negative patients. The ODX test is based on 1A-level evidence and it is included in the main international clinical guidelines recommendations such as those of the American Society of Clinical Oncology (ASCO [3]) or the National Comprehensive Cancer Network (NCCN) as well as in the last staging guidelines of AJCC 8th edition [12]. The ODX test is the most widely available molecular test used in the world. This assay analyzes 21 genes by RT-qPCR (16 cancer-related genes and 5 housekeeping genes) and aims to predict the risk of recurrence at 10 years by providing a recurrence score ranging from 0 to 100 and to estimate the benefit of adjuvant chemotherapy. Several retrospective and prospective studies have validated this test and its clinical utility. Paik et al. [23] have shown a correlation between ODX score and disease-free survival in patients with ER-positive/HER2-negative, node negative, tamoxifen-treated breast cancer, based on the NSABP B-14 trial. As for the chemotherapy benefits, Paik et al. [24] and Albain et al. [1] have evaluated the test using the studies related to NSABP-B20 and SWOG 8814. The prospective phase III trial TAILORx study [27] has modified the ODX score's cutoff values (low risk <11, intermediate risk 11-25 and high risk >25) in order to avoid under-treatments of cancer. To be more precise, in the low group, the risk of recurrence at 5 years is very low (< 10%) with hormonal therapy, which confirms the uselessness of adding a chemotherapy [26]. For the intermediate group, chemotherapy has a benefit only for women younger than 50 years old. For the high-risk group, the chemotherapy is highly recommended. Nevertheless, one third of women with hormone-receptor positive breast cancer have a lymph node disease. Thus, the prospective trial RxPONDER trial study analyzes the capacity of the ODX test to predict the benefit of chemotherapy for women with positive lymph node disease [16]. RxPONDER showed that postmenopausal patients with node involvement and an ODX score between 0 and 25 did not benefit from chemotherapy, whereas premenopausal patients with node involvement with 1-3 nodes and ODX scores between 0 and 25 benefited significantly from chemotherapy.

Despite its proven value, the ODX test is not routinely used due to its high cost. For this reason, less than 20% of patients in Europe have access to the ODX test. Health-related economic studies are performed to understand for which patients the assay is the most useful [2]. From this economic point of view, many alternative tools have been developed to predict this score. These tools are based on clinico-pathological data such as Magee equations [18, 28] and the IHC4 score [30]. Indeed, many studies have shown the correlation between the results of the latter tools and the ODX score [11]. Few works used features with machine learning techniques in order to provide an ODX-based methodology to divide the patients into categories corresponding to low or high risk of cancer [17, 22, 5, 25].

The aim of this paper is to propose a novel methodology to assist physicians in their decision-making. It is based on random forests for distributional regression as presented in Meinshausen [21] and Athey et al. [4]. This methodology creates links between a new patient and the cohort used for training based on clinico-pathological characteristics. These links can be used particularly to identify the patients of the training cohort that share similarities with the new patient and to predict an estimate of the distribution of the ODX score ($ODX \leq 25$ and $ODX > 25$). This information is available to clinicians to help them better understand the probable clinical evolution of the tumor in order to optimize the treatment. Moreover, it enables knowledge capitalization by feedback and analysis of patient history. One of the consequences of this study is to weigh the

variability of the anatomo-pathological data, so this new methodology can adapt to the specificities of a cohort.

2 Materials and methods

2.1 Dataset description

The cohort is a retrospective study between 2012 and 2020 with 333 cases that underwent an ODX assay from three hospitals in the Bourgogne Franche-Comté region (France): Besançon, Belfort and Dijon. All patients have ER-positive and HER2-negative early breast cancer. Clinical and pathological reports were used to collect the data such as the age at diagnosis, the menopausal status, the treatment, the recurrence, the tumor size, the lymph node status, the histological type, the Nottingham grade, hormone receptors for estrogen (ER) expression, hormone receptors for progesterone (PR) expression, the human epidermal growth factor receptor 2 (HER2) status and the protein p53 and Ki67 proliferation index. Immunohistochemical staining was performed (Ventana Benchmark XT system[®], Roche[™]) on the tumor block of ODX testing with UltraView Universal DAB detection with ER antibody (clone SP1; Roche/Ventana Medical Systems, Tucson, USA), PR antibody (clone 1E2; Roche/Ventana Medical Systems, Tucson, USA), HER2 antibody (clone 4B5; Roche/Ventana Medical Systems, Tucson, USA), Ki67 antibody (clone Mib-1, Dako, Glostrup, Denmark) and p53 antibody (clone DO-7, Dako, Glostrup, Denmark). The HER2 immunostaining was interpreted using the 2018 ASCO/College of American Pathologists guidelines [29]. The Ki67 proliferation index was evaluated by manual counting with counter on at least 200 tumor cells. The protein p53 was assessed by immunohistochemistry. The positive threshold is greater than 10% of the tumor cells' nuclei. The ODX test was realized by Genomic Health (Redwood City, CA, USA) and analyzed 21 genes by RT-qPCR from paraffin-embedded blocks of tumor tissue. The ODX score was obtained from the clinical reports. The three ODX categories were the same as the ones defined in the ODX assays using TAILORx and RxPONDER: low risk (< 16), intermediate risk (16-25) and high risk (> 25). The institution review board approved this study.

The cohort contains more than 50 features, from which we selected the most critical ones using feature importance in random forest and physicians' assessments. Table 1 describes the tumor characteristics using the features selected for our study.

2.2 Distributional Random Forest

Random Forest [7] is a powerful machine learning algorithm that can be used for prediction in various settings and has been successfully applied in the field of medicine [9, 10, 31]. Our goal here is to predict the result of the expensive ODX test based on clinico-pathological features. We propose the use of Distributional Random Forest that provides a predictive distribution for the ODX score based on the clinico-pathological features. We shall expose the methodology for Random Forest (RF) and Distributional Random Forest (DRF).

Standard regression links the mean of the response variable Y to a set of features X based on observations from a training sample of feature–response pairs, say (X_i, Y_i) for $i = 1, \dots, n$. RF is an ensemble method that consists in the bootstrap aggregation [6] of randomized classification and regression trees (CART, [8]). The predictive mean can be written as the average

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^B T_b(X), \quad (1)$$

Table 1 ODX score distribution by patient and tumor characteristics.

		Percentage of patient by category			
		< 16	16 – 25	> 25	Total
Population		113	138	82	333
Age	≤ 50 yr	14.41	10.51	5.71	30.63
	> 50 yr	19.52	30.92	18.92	69.37
Tumor size	< 1 cm	3.90	4.51	3.00	11.41
	1-2 cm	15.62	22.52	15.62	53.76
	> 2 cm	14.41	14.41	6.01	34.83
p53	≤ 10%	18.62	23.12	12.01	53.75
	> 10%	15.32	18.32	12.61	46.25
SBR grade	1	5.41	3.30	0.00	8.71
	2	21.02	24.03	10.81	55.86
	3	7.51	14.11	13.81	35.43
Mitotic grade	1	12.61	14.42	4.50	31.53
	2	17.12	20.12	12.01	49.25
	3	4.20	6.91	8.11	19.22
ER status	Negative	0.00	0.00	0.00	0.00
	Positive (≥ 10%)	33.93	41.44	24.63	100
PR status	Negative	2.10	7.51	8.11	17.72
	Positive (≥ 10%)	31.83	33.93	16.52	82.28
Ki67-positive cells	< 10%	0.00	0.30	0.00	0.30
	10 – 20%	16.22	15.92	4.80	36.94
	> 20%	17.72	25.22	19.82	62.76
Lymph node status	0	15.02	15.52	13.81	45.35
	1	10.81	14.11	4.20	29.13
	2	3.61	3.30	0.90	7.81
	3	1.80	2.10	2.10	6.00
	NA	2.70	5.41	3.60	11.71

where $T_1(X), \dots, T_B(X)$ corresponds to the prediction of the different trees built on different bootstrap samples. Each single tree prediction takes the form of an average across a neighborhood of X in the tree, i.e.

$$T_b(X) = \frac{1}{|R_b(X)|} \sum_{X_i \in R_b(X)} Y_i,$$

where $R_b(X)$ denotes the neighborhood of X in the tree T_b and $|R_b(X)|$ the numbers of observations that fall into this region. Consequently, the RF prediction (1) has the equivalent form

$$\hat{Y} = \sum_{i=1}^n w_i(X) Y_i, \quad (2)$$

with the RF weights defined by

$$w_i(X) = \frac{1}{B} \sum_{b=1}^B \frac{\mathbb{1}_{\{X_i \in R_b(X)\}}}{|R_b(X)|}, \quad 1 \leq i \leq n. \quad (3)$$

These weights are non-negative with sum 1 (probability weights).

The main idea of DRF relies on Equation (2): the prediction \hat{Y} is the sample mean of the weighted sample Y_i with weights $w_i(X)$ which can be seen as an approximation of the conditional distribution of Y given X . The cumulative distribution function $F(y|X) = \mathbb{P}(Y \leq y|X)$ is thus approximated by

$$\hat{F}(y|X) = \sum_{i=1}^n w_i(X) \mathbb{1}_{\{Y_i \leq y\}}. \quad (4)$$

This idea was first suggested by Meinshausen [21] who proposed the construction of quantile regression forest by approximating the conditional quantile of Y given X by the quantiles of the weighted empirical distribution (4).

Figure 1 presents a synthetic representation of the DRF procedure with the different steps: subsampling of the original sample, tree construction on each subsample, computation of the neighborhood/weights at the point to predict, averaging of weights given by the different trees that finally provide the predictive distribution.

The RF weights (3) are interesting in themselves and provide relevant information in terms of similar/influential observations. Given a new feature X , the weight $w_i(X)$ is interpreted as the proportion in which the observation Y_i contributes to the prediction of Y given X . Observations with the largest weights are interpreted as the nearest neighbors of X in terms of an implicit metric on the predictor space that is tailored for predicting the response, see [19]. The RF weights make it possible to identify the most similar/influential individuals in the training data. Comparing a new individual with clinico-pathological features X to observations with similar clinico-pathological features help us to predict the ODX score and understand the relationship between clinico-pathological features and the ODX score.

Finally, let us mention that the weights (2)-(3) depend on the specific structure of the trees that are used for prediction. Trees are grown by recursive binary splitting, maximising a homogeneity criterion; the goal is to partition the feature space into different regions that are as homogeneous as possible. In the standard CART algorithm, the variance is used as the homogeneity criterion, resulting in a partition adapted to the prediction of the mean. Several different splitting rules have been considered in the statistical literature that put the emphasis on the prediction of quantiles [4, Generalized Random Forest] or on the overall distribution [32, Distributional Random Forest].

For implementation to our data set, we use the software R with package `grf` (Generalized Random Forest; [4]) to compute the Random Forest and the associated weights.

2.3 Applications of Distributional Random Forest

DRF is a fully non-parametric and model-free method that performs probabilistic forecast and distributional regression. For a set of features X , it provides the full predictive distribution of the response variable Y , that is to say exhaustive information for its possible fluctuations knowing the features. The method is very informative and powerful (see Figure 2) as it provides:

- (distributional regression) a predictive distribution for each new case that can be represented by a histogram;
- (mean or median prediction) a predictive mean or median when a point estimate is needed - the mean is commonly used while the median is more robust to outliers;
- (uncertainty assessment) a graphical assessment of the uncertainty with the shape of the histogram (either peaked or flat) or numerical statistics such as standard error or confidence interval for the prediction;
- (classification) the probability of classes of particular interest can be instantly computed - for the ODX score, the classes $ODX \leq 25$ and $ODX > 25$ are considered;
- (similar/influential patients) the patients in the cohort (training set) that are the most similar to a new case can be easily identified through the RF that are interpreted as a measure of proximity - this proximity is meant in the sense of an implicit distance that is learnt by the model and that gives more importance to the relevant features; this information can allow the practitioner to make meaningful and informative comparisons between the new case and the patients from the cohort.

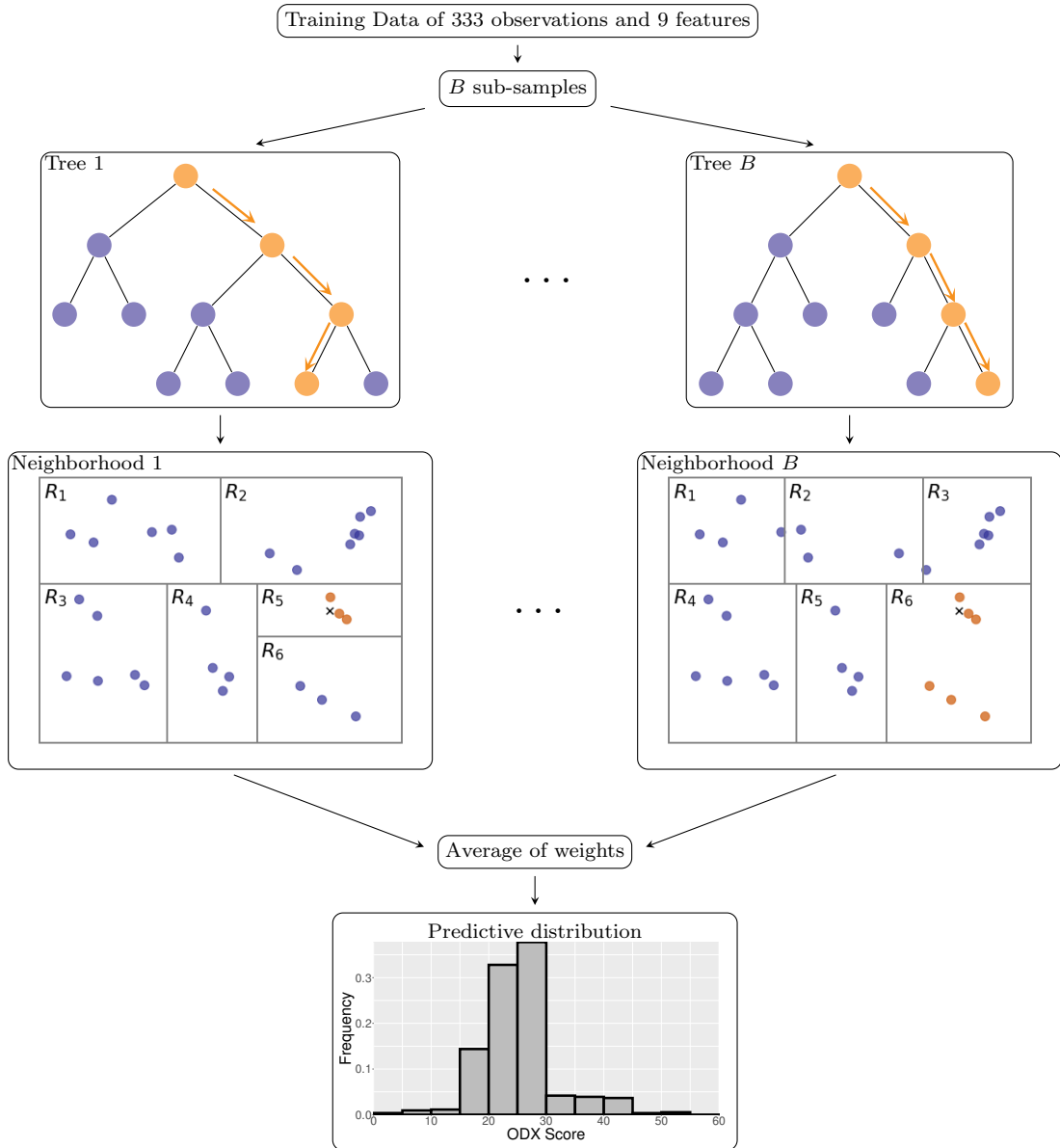


Fig. 1 Flowchart of Distributional Random Forest. Starting from the training data, a large number of subsamples are randomly chosen and binary trees are constructed on each subsamples; the neighborhood/weights at the point to predict are computed in each tree and then averaged so as to give the forest weights; the predictive distribution corresponds to the weighted sample of the original training data with these forest weights.

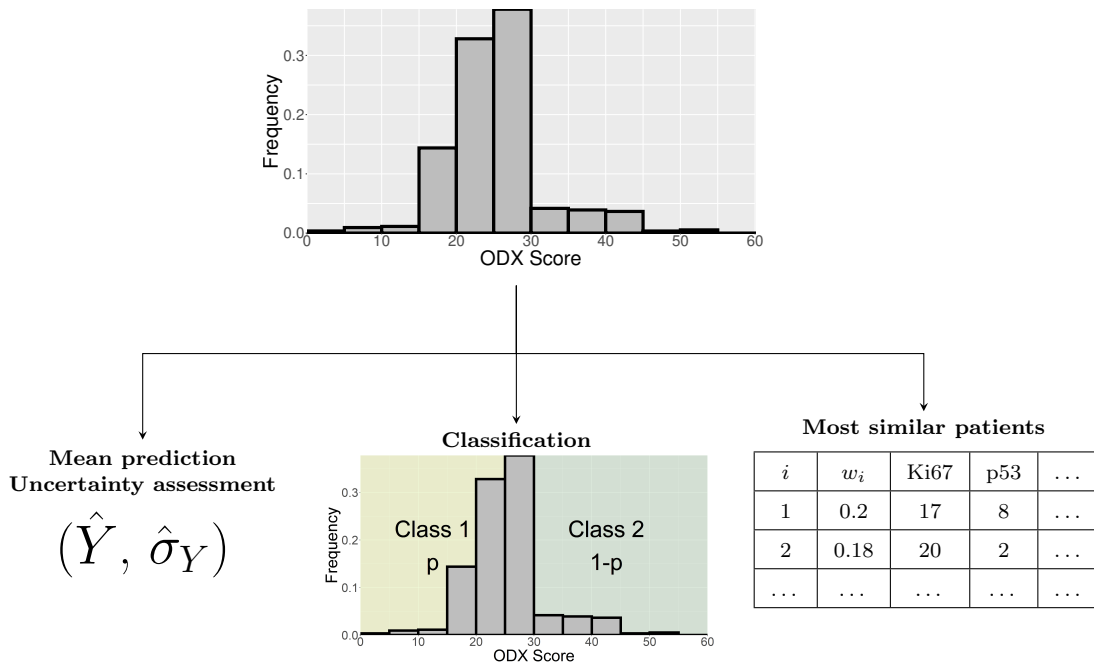


Fig. 2 Applications of Distributional Random Forest. Once the DRF is trained, prediction of classes (classification) or of conditional mean or median (regression) together with an uncertainty estimate is straightforward. Furthermore, the weights at the point to predict make it possible to identify the most similar neighbors in the training data, with an adaptive notion of similarity tailored for the purpose of prediction.

2.4 Evaluation of predictive performance

In order to evaluate the DRF algorithm and compare it with concurrent methods, the theory of a proper scoring rule [14] is used. In probabilistic forecasting, a scoring rule compares a predictive distribution F and the outcomes y . It plays the role of a measure of error similar to the mean squared error in regression or the misclassification rate in classification. A scoring rule is strictly proper if the expected score is minimal when the predictive distribution F matches the outcome distribution. A strictly proper scoring rule can be used for the evaluation of probabilistic forecast and distributional regression [13].

The most popular scoring rule is the Continuous Ranked Probability Score (CRPS) [20] and is defined by

$$\text{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbf{1}_{\{y \leq z\}})^2 dz.$$

In a case where the predictive distribution F corresponds to a weighted sample $(y_i)_{1 \leq i \leq n}$ with weights $(w_i)_{1 \leq i \leq n}$, the CRPS is easily computed by

$$\text{CRPS}(F, y) = \sum_{i=1}^n w_i |y_i - y| - \sum_{1 \leq i < j \leq n} w_i w_j |y_i - y_j|.$$

The first term compares the predictive distribution F and observation y (calibration) while the second term assesses the precision of the prediction (sharpness). This expression also shows that $\text{CRPS}(F, y)$ is reported in the same unit as the observation y and

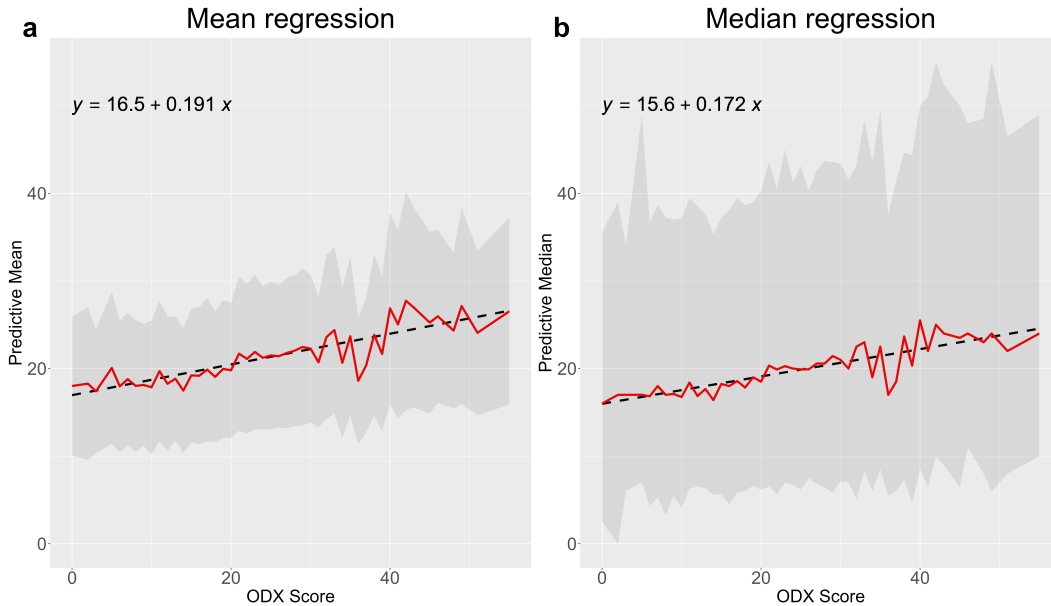


Fig. 3 Evaluation with regression. Mean-regression (left figure - a) plots the ODX observation versus their predictive mean; the grey ribbon represents the standard errors. Median-regression (right figure - b) plots the ODX score versus their predictive median; the grey ribbon represents the 90%-confidence interval.

that it generalizes the absolute error to which it is reduced if F is a deterministic forecast, that is to say a point measure.

In order to evaluate the generalization capacity of the model, i.e. its predictive performance on a new sample, different validation methods can be used to assess the prediction error. Simple validation uses a training set to fit the model and an independent test set to compute error (CRPS). K -fold cross validation is more involved and splits the data into K groups that successively play the role of the test set. More precisely, K different models are fitted on training sets consisting of all folds but one which is left-out during training and used as a test set to compute the CRPS; this results in K different test errors which are averaged so as to obtain the K -fold cross validation error. In the specific case of bagging including DRF, the out-of-bag (OOB) method can be used instead. It usually provides similar results as K -fold cross validation but is much more numerically efficient since only one fit of the model is required.

3 Results

The DRF was applied to 333 patients to predict the ODX score using the 9 features presented in Table 1. In order to compare with the literature, we emphasize the classification into two classes ($ODX \leq 25$ and $ODX > 25$).

Before presenting the results of the DRF, we shall first present the evaluation of our model. Simple graph can be performed by considering the regression model deduced from DRF. The results of the regression are presented in Figure 3, where the predictive mean (Figure 3a) and predictive median (Figure 3b) versus the real ODX score are plotted. We can observe a rather good fit, and that an important proportion of the observations are within in their confidence intervals. In Figure 3a, the grey ribbon has a semi-amplitude equal to the standard error and accounting for uncertainty. The grey ribbon in Figure 3b represents the credibility interval with a level of 90%.

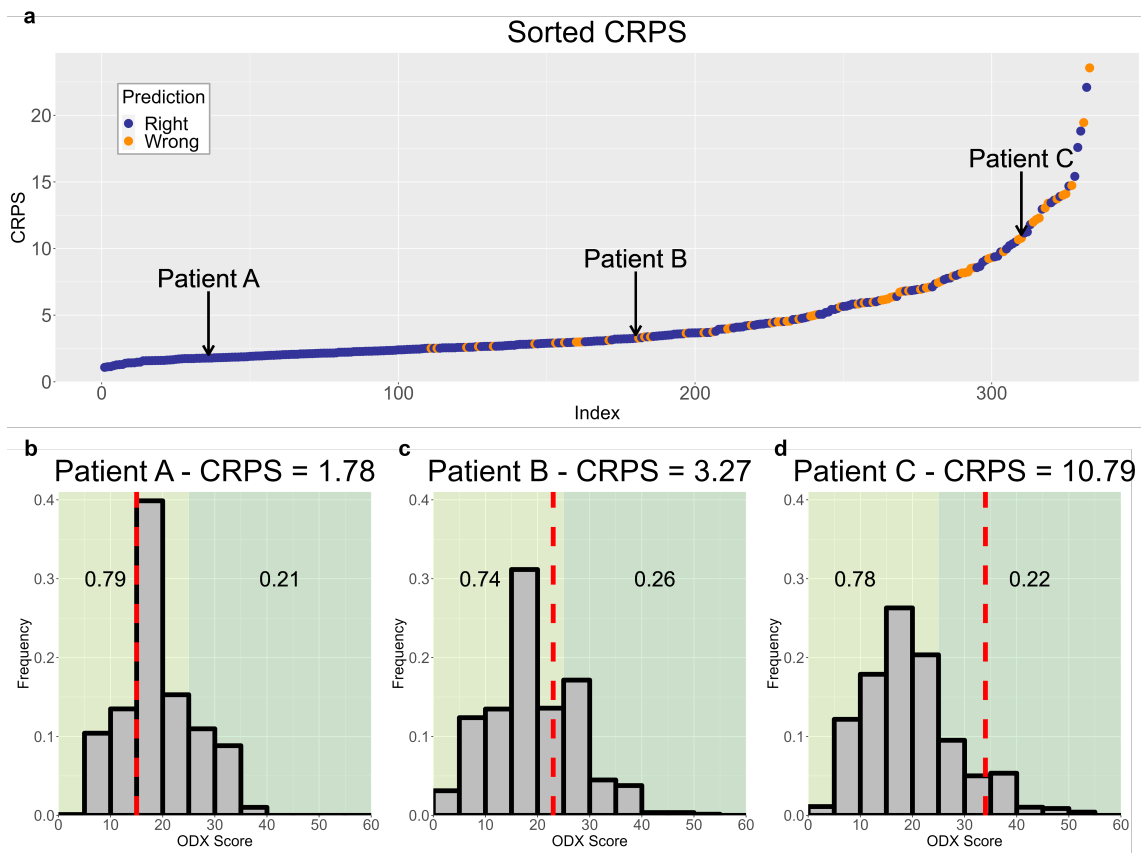


Fig. 4 Out-of-Bag evaluation of the prediction with the CRPS - a low CRPS corresponds to a precise forecast. The subfigures b, c and d correspond to three different patients chosen in different range of the CRPS presented in the subfigure a. In the three lower subfigures, the gray histogram corresponds to the predicted distribution of the ODX score obtained by the DRF. The red dashed line represents the true ODX score of the patient. The two-classes ($ODX \leq 25$ and $ODX > 25$) are represented as areas of different colors and the predicted probabilities of each class is given for each patient.

Additionally, in order to assess the prediction quality, we compared the OOB predictive distribution and the actual observation for the ODX, for each observation. The prediction error is measured in terms of the CRPS introduced in Section 2.4. The different scores are represented in Figure 4a. The smaller the CRPS, the more accurate the forecast. We can observe that most of the predictions have a small or medium CRPS, which indicates the overall good quality of prediction. A smaller number of observations have a large CRPS, indicating individuals for whom the ODX score notably differs from what we might expect in comparison with the overall population. Together with the CRPS, the figure provides the results for the binary classification task ($ODX \leq 25$ or $ODX > 25$): classification errors are indicated with the color orange while the color blue corresponds to correctly classified observations. We can observe a good match between classification errors and a large CRPS, which confirms the ability of the CRPS to assess forecast quality. Then, for each patient, the DRF provides a predictive distribution represented by a histogram that can be compared with the actual ODX score. We also indicate the two class probabilities corresponding to the light-green/left or dark-green/right classes.

We have selected three patients respectively with a low (Figure 4b, Patient A), medium (Figure 4c, Patient B) and large CRPS (Figure 4d, Patient C). The predictions

associated to these patients can be considered "good", "average" and "bad", respectively. We now provide a description for the patients A and C. Patient A is 50 years old and peri-menopausal. The tumor size and SBR grade are respectively 13 mm and 2. She has macro metastases and is ER and PR positive. The Ki67, p53 and ODX score are respectively 15, 8 and 15. Patient A did not undergo chemotherapy. The patients listed as neighbors of Patient A ranged in age from 46 to 55, with an average age of 50. Six patients were non-menopausal, three menopausal and one peri-menopausal. Tumor size ranged from 11 to 20 mm. Seven patients had grade 2 SBR, two patients grade 1 and one patient grade 3. Five patients have macro metastases, one patient micro metastases and four patients with no metastases. Tumors were non-specific carcinomas and ER positive, PR positive in 7 HER2-negative patients. Ki67 ranged from 8 to 20%. p53 was unmutated. Oncotype ranged from 7 to 30, with no ODX over 30 and a recurrence risk of 8 to 23% at 10 years. In terms of treatment, all had surgery, radiotherapy and hormone therapy. One patient underwent chemotherapy. None of the patients had a recurrence. As for Patient C, she is 62 years old and menopausal. The tumor size and SBR grade are respectively 15 mm and 3. She has no metastases and is ER and PR positive. The Ki67, p53 and ODX score are respectively 24, 0 and 34. Patient C had chemotherapy. The patients listed as neighbors of Patient C are ranged in age from 60 to 78, with an average age of 68. All of them are menopausal. Tumor size ranged from 9 to 44 mm. Seven patients had grade 3 SBR and three patient grade 2. All patients have no metastases. Tumors were non-specific carcinomas and ER positive, PR positive in 8 HER2-negative patients. Ki67 ranged from 20 to 28%. p53 was unmutated for four patients. Oncotype ranged from 6 to 36 and a recurrence risk of 5 to 24% at 10 years. In terms of treatment, all had surgery, radiotherapy and hormone therapy. Two patients underwent chemotherapy. None of the patients had a recurrence. In Figure 4b we can observe a sharp predictive distribution (peaked histogram) and an ODX score close to the peak. In Figure 4c, the histogram is flatter, indicating more uncertainty, and the true ODX score is contained in a high probability region. In Figure 4d, the predictive distribution has also a large dispersion and the ODX score is contained in a low probability region, which means that the match between the two is poor. We insist on the fact that a large CRPS does not necessary mean a misclassification of a patient as it can be seen for some of the higher CRPS values in Figure 4a. The CRPS considers the distributional regression and is not explicitly related to the binary classification presented here.

Due to the impact of the classification of ODX in the two classes $ODX \leq 25$ and $ODX > 25$, we shall present the detailed evaluation of the classification model deduced from DRF (see Table 2). This evaluation is based on the standard classification metrics such as confusion matrix and standard metrics. The standard metrics are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \quad (5)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad (6)$$

$$\text{Specificity} = \frac{TN}{FP + TN}, \quad (7)$$

$$\text{Positive Predictive Value} = \frac{TP}{TP + FP}, \quad (8)$$

$$\text{Negative Predictive Value} = \frac{TN}{FN + TN}, \quad (9)$$

$$\text{F1-score} = \frac{2 * \text{Positive Predictive Value} * \text{Sensitivity}}{\text{Positive Predictive Value} + \text{Sensitivity}} \quad (10)$$

Table 2 Evaluation with classification. Confusion matrix (left) together with standard metrics (right).

		Predicted	
		ODX \leq 25	ODX $>$ 25
True	ODX \leq 25	231	20
	ODX $>$ 25	49	33

Accuracy	79.3%
Sensitivity	92.0%
Specificity	40.3%
Positive Predictive Value	82.5%
Negative Predictive Value	62.3%
F1-score	0.870
Area Under Curve	0.759

where TP is the number of patients correctly classified as ODX \leq 25, FP is the number of patients incorrectly classified as ODX \leq 25, TN is the number of patients correctly classified as ODX $>$ 25 and FN is the number of patients incorrectly classified as ODX $>$ 25.

We have correctly classified 231 out of 251 patients (92%) in low risk and 33 out of 75 patients (40.2%) in high risk. The overall accuracy is 79.3% and the p-value is less than 0.05. The proportion of low risk correct predicted value (PPV) is 82%. The percentage of high risk correct predicted value (NPV) is approximately 62.3%. The F1-score and the Area Under Curve (AUC) are of 0.87 and 0.759, respectively. The DRF will provide additional information such as the nearest neighbor patients, the distribution of the ODX score and the uncertainty prediction (see Figure 2). We now consider the 69 misclassified patients with low and high risks. First of all, we notice that the majority of these patients have predictions that are close to the decision border (i.e. close to ODX = 25). These patients are misclassified because of the binary decision and additional information available with the DRF method shows either that the patient’s ODX score is close to the decision border or that the neighborhood of the patient is not realistic because of limitations of the training cohort. This first part of the misclassified patients might have a small CRPS as the CRPS accounts for both the dispersion of the prediction and its bias. The second part of the misclassified patients correspond to extremal values of the ODX score within our cohort. The nearest patients provided by the DRF for these misclassified patients are thus less informative as they are taken within the cohort that is not representative of these outlier patients. In order to give more quantitative results, we compared the mean absolute difference for the ODX score, Ki67 and p53 between the 69 misclassified patients and the weighted average value of their neighborhoods. The misclassified patients have a mean absolute difference of ODX score compared to their neighborhood of 9.84 where the correctly classified patients have an average absolute difference of 6.29. In terms of Ki67 and p53, the average absolute difference is 24.56% and 5.77% respectively when the average absolute difference for the correctly classified patient is 16.77% for Ki67 and 5.84% for the p53 respectively.

These classification results are then compared with state-of-the-art techniques [18, 15, 17, 22, 5, 25]. A detailed comparison is given in Table 3.

4 Discussion

The ODX test is the most commonly available breast genomic test used in early stage ER positive/ HER2-negative breast cancer. It makes it possible to define patients who are unlikely to benefit from chemotherapy. The ODX score is based on 6 gene groups. These groups correspond to the markers analysis in pathological reports. Some have compared the ODX score to this immuno-histological data and proved the predictive relationship with the ODX score. Several studies were published using this clinico-pathological data to predict the ODX score with different methods (see Table 3). The present study was realised to predict the ODX score from a specific regional cohort of 333 patients with clinical and immuno-histological data using Distributional Random Forest. This

prediction is associated with a predictive error on the one hand, and the ability to determine the similar patients on the other hand. The proposed DRF model detected 82% of lower risk patients ($ODX \leq 25$) and 62.3% of high risk patients ($ODX > 25$).

A few studies have proposed some prediction tools for the ODX score [18, 15, 17, 22, 5, 25]. Each study is based on the specific categorization of patients according to the original ODX categories and TAILORx (see ODX Prediction Threshold in Table 3). The prediction results of the different studies are similar and based on clinico-pathological data. The tumor size, tumor grade and PR are used in all the six selected published studies as well as for our current study. The Ki67 is not used in [22] and [25]. In our study, we integrated the p53. The threshold used for the ODX score is different from one study to another. Our DRF model performs as well as the other prediction tools. The novelty is in providing additional information to the prediction (see Figure 2) such as the probability of classes (low and high risk), the similar profiles and the uncertainty prediction.

The correct predicted values are 82.5% and 62.3% for low and high risk, respectively. We used the CRPS score to distinguish the best and worst prediction. A negative correlation between ODX and PR for the best and worst results can be observed. The best results were obtained for ODX profiles below 16. The average Ki67, for the first best ten results, is under 14%, which corresponds to the low-risk profile of our previous study [5]. The average percentages of ER and PR are 93% and 77%, respectively, which fits into the same low risk profiles. When looking at the RF weights and the profile of closest patients, we observe that similar profiles vary between 0 and 25 for ODX. The similarities fall in the low risk profile. The Ki67 scores of similar profiles for the first ten results are below 25%. We can point out from the description of the results of Patient A in Section 3 that the prediction of our model is coherent with the listed neighbors.

Most cases that were misclassified by our model lie in the high risk class. The averages of the ODX score, Ki67 and PR for the last worst ten results are respectively 40, 28% and 26%. Patient C, described in Section 3, was misclassified by our model. This patient has a high ODX score and probability of lying in the high class of 22%. The real ODX score is 34 and the predicted ODX score is near to the cut-off. The similar profiles are very dispersed, which is difficult to analyze. Most of the nearest neighbors have an SBR grade of 3. The prediction is bad, but nevertheless, the similar profiles have a low ODX score and a high SBR grade. This is due to the small number of cases in the high risk category.

The size of the cohort and the training and testing phase could impact the prediction results. In addition, we have an unbalanced cohort in our study, since we have less patients in the high risk class. In that case, the factors of the similar profiles that influences ODX score such as PR, Ki67 and p53 should be considered. The size of the cohort and the training and testing phase could impact the prediction results. In addition, we have an unbalanced cohort in our study, since we have less patients in the high risk class. In that case, the factors of the similar profiles that influences ODX score such as PR, Ki67 and p53 should be considered. The strengths of our proposed model include the distribution of identical profiles which can allow to the clinician to retrieve similar historical cases in terms of evolution. Moreover, the proposed model can be applied even when there is missing data. It makes it possible to predict the low risk class with high certitude, which means no chemotherapy to plan. Our study is related to the dataset and it is therefore difficult to generalize to a different cohort because of known inter-cohort variability, especially on some biomarkers such as Ki67.

5 Conclusion

This paper proposes a new methodology for ODX score prediction. This methodology is based on 9 clinico-pathological features and uses Distributional Random Forest to perform prediction. It makes it possible to predict the distribution of the ODX score for a patient and provides an explanation of the predicted score by computing the probability of belonging to the low or high risk category and identifying the nearest similar profiles. The proposed Distributional Random Forest model detects 82% of lower risk patients (≤ 25) and 62.3% of patients with high risk (> 25). However, DRF presents certain limitations. The use of DRF with the pathologist's expertise on the different histological and immunohistochemical characteristics has a clinical impact to help oncologist in decision-making regarding breast cancer therapy. Thanks to the family of patients with identical profiles determined by the proposed methodology, with knowledge of their evolution and prognosis, the oncologists can support their therapeutic decision. The medico-economic interest of this strategy is obvious. Additional studies are needed to further validate the DRF method and improve knowledge extraction from pathological data.

Table 3 Comparison of our study with six selected published studies [18, 15, 17, 22, 5, 25] to predict the ODX score. For three classes only the sensitivity and specificity of the lower class are given.

		Klein et al. (2013)	Hou et al. (2017)	Kim et al. (2019)	Orucevic et al. (2019)	Baltres et al. (2020)	Pawloski et al. (2022)	Current study (DRF)
Patients	(n_{train}, n_{test})	(817, 255)	(-, 163)	(208,76)	(65,754, 18,585)	(152, 168)	(2,587, 1,293)	(333, OOB)
Age	Mean	-	58.6	-	-	-	-	56.9
	Median	-	-	44.0	58	57.5	62	58.0
	Range	-	34-82	-	19-90	30-84	56-69	30-84
Clinico-pathologic features used for modelling	Tumor size	✓	✓	✓	✓	✓	✓	✓
	Tumor grade	✓	✓	✓	✓	✓	✓	✓
	Lymphovascular invasion			✓			✓	
	Lymph node status			✓		✓		✓
	ER	✓	✓	✓	✓	✓		✓
	PR	✓	✓	✓	✓	✓	✓	✓
	Ki67	✓	✓	✓		✓		✓
	p53							✓
ODX Prediction	Type	Continuous	Continuous	Classification	Classification	Classification	Classification	Distributional
	Threshold	< 18 18 – 30 > 30	< 18 18 – 30 > 30	< 11 > 25	≤ 25 > 25	< 18 18 – 30 > 30	≤ 25 > 25	≤ 25 > 25
Method		Multiple Linear Regression	Multiple Linear Regression	Neural Network Decision Jungle	Binomial Logistic Regression	Deep Multi-Layer Perceptron	Random Forest	Distributional Random Forest
Precision	Low risk	62.5-69.4%	72.6%	100%	87.5%	58.3%	92.9%	82.5%
	High risk	68.8-77.8%	-	25.0%	79.6%	63.0%	65.1%	62.3%
Sensitivity		58.6-59.1%	85.7%	11.0%	99.2%	55%	96.3%	92.0%
Specificity		70.5-77.4%	41.4%	100%	18.3%	78%	48.3%	40.2%
AUC		-	-	0.744	0.81	0.63	-	0.759

Author Contributions All authors contributed to the study conception, design, data analysis and manuscript preparation. All authors reviewed and approved the final version of the manuscript.

Conflict of Interest Statement All authors of this work declare that there are no conflicts of interest in the authorship or publication of this contribution.

Funding Statement The authors declare that no funds, grants or other support were received during the preparation of this manuscript.

References

1. Kathy S Albain, William E Barlow, Steven Shak, Gabriel N Hortobagyi, Robert B Livingston, I-Tien Yeh, Peter Ravdin, Roberto Bugarini, Frederick L Baehner, Nancy E Davidson, et al. Prognostic and predictive value of the 21-gene recurrence score assay in a randomized trial of chemotherapy for postmenopausal, node-positive, estrogen receptor-positive breast cancer. *The lancet oncology*, 11(1):55, 2010.
2. Joan Albanell, Christer Svedman, Joseph Gligorov, Simon DH Holt, Gianfilippo Bertelli, Jens-Uwe Blohmer, Roman Rouzier, Ana Lluch, and Wolfgang Eiermann. Pooled analysis of prospective european studies assessing the impact of using the 21-gene recurrence score assay on clinical decision making in women with oestrogen receptor-positive, human epidermal growth factor receptor 2-negative early-stage breast cancer. *European Journal of Cancer*, 66:104–113, 2016.
3. Fabrice Andre, Nofisat Ismaila, N Lynn Henry, Mark R Somerfield, Robert C Bast, William Barlow, Deborah E Collyar, M Elizabeth Hammond, Nicole M Kuderer, Minetta C Liu, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: Asco clinical practice guideline update—integration of results from tailorx. *Journal of Clinical Oncology*, 37(22):1956–1964, 2019.
4. Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Ann. Statist.*, 47(2):1148–1178, 2019. ISSN 0090-5364. doi: 10.1214/18-AOS1709. URL <https://doi.org/10.1214/18-AOS1709>.
5. Aline Baltres, Zeina Al Masry, Ryad Zemouri, Severine Valmary-Degano, Laurent Arnould, Nouredine Zerhouni, and Christine Devalland. Prediction of oncotype dx recurrence score using deep multi-layer perceptrons in estrogen receptor-positive, her2-negative breast cancer. *Breast Cancer*, 27(5):1007–1016, 2020.
6. Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996. ISSN 0885-6125.
7. Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324.
8. Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Classification and regression trees. The Wadsworth Statistics/Probability Series. Belmont, California: Wadsworth International Group, a Division of Wadsworth, Inc. X, 358 p. \$ 29.25; \$ 18.95 (1984)., 1984.
9. Gaoxiang Chen, Qun Li, Fuqian Shi, Islem Rekik, and Zhifang Pan. Rfdcr: Automated brain lesion segmentation using cascaded random forests with dense conditional random fields. *NeuroImage*, 211:116620, 2020.
10. Carlos Fernandez-Lozano, Pablo Hervella, Virginia Mato-Abad, Manuel Rodríguez-Yáñez, Sonia Suárez-Garaboa, Iria López-Dequidt, Ana Estany-Gestal, Tomás Sobrino, Francisco Campos, José Castillo, et al. Random forest-based prediction of stroke outcome. *Scientific reports*, 11(1):1–12, 2021.

11. Melina B Flanagan, David J Dabbs, Adam M Brufsky, Sushil Beriwal, and Rohit Bhargava. Histopathologic variables predict oncotype dx™ recurrence score. *Modern Pathology*, 21(10):1255–1261, 2008.
12. Armando E Giuliano, James L Connolly, Stephen B Edge, Elizabeth A Mittendorf, Hope S Rugo, Lawrence J Solin, Donald L Weaver, David J Winchester, and Gabriel N Hortobagyi. Breast cancer—major changes in the American joint committee on cancer eighth edition cancer staging manual. *CA: a cancer journal for clinicians*, 67(4):290–303, 2017.
13. Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, jan 2014. doi: 10.1146/annurev-statistics-062713-085831.
14. Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.*, 102(477):359–378, 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL <https://doi.org/10.1198/016214506000001437>.
15. Yanjun Hou, Gary Tozbikian, Debra L Zynger, and Zaibo Li. Using the modified Magee equation to identify patients unlikely to benefit from the 21-gene recurrence score assay (oncotype dx assay). *American Journal of Clinical Pathology*, 147(6): 541–548, 2017.
16. Kevin Kalinsky, William E Barlow, Julie R Gralow, Funda Meric-Bernstam, Kathy S Albain, Daniel F Hayes, Nancy U Lin, Edith A Perez, Lori J Goldstein, Stephen KL Chia, et al. 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. *New England Journal of Medicine*, 385(25):2336–2347, 2021.
17. Isaac Kim, Hee Jun Choi, Jai Min Ryu, Se Kyung Lee, Jong Han Yu, Seok Won Kim, Seok Jin Nam, and Jeong Eon Lee. A predictive model for high/low risk group according to oncotype dx recurrence score using machine learning. *European Journal of Surgical Oncology*, 45(2):134–140, 2019.
18. Molly E Klein, David J Dabbs, Yongli Shuai, Adam M Brufsky, Rachel Jankowitz, Shannon L Puhalla, and Rohit Bhargava. Prediction of the oncotype dx recurrence score: use of pathology-generated equations derived by linear regression analysis. *Modern Pathology*, 26(5):658–664, 2013.
19. Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.*, 101(474):578–590, 2006. ISSN 0162-1459. doi: 10.1198/016214505000001230. URL <https://doi.org/10.1198/016214505000001230>.
20. James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22, jun 1976. doi: 10.2307/2629907.
21. Nicolai Meinshausen. Quantile regression forests. *J. Mach. Learn. Res*, pages 983–999, 2006.
22. Amila Orucevic, John L Bell, Megan King, Alison P McNabb, and Robert E Heidel. Nomogram update based on tailorx clinical trial results—oncotype dx breast cancer recurrence score can be predicted using clinicopathologic data. *The Breast*, 46: 116–125, 2019.
23. Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.
24. Soonmyung Paik, Gong Tang, Steven Shak, Chungyeul Kim, Joffre Baker, Wanseop Kim, Maureen Cronin, Frederick L Baehner, Drew Watson, John Bryant, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *Journal of clinical oncology*, 24(23):3726–3734, 2006.

25. Kate R Pawloski, Mithat Gonen, Hannah Y Wen, Audree B Tadros, Donna Thompson, Kelly Abbate, Monica Morrow, and Mahmoud El-Tamer. Supervised machine learning model to predict oncotype dx risk category in patients over age 50. *Breast cancer research and treatment*, 191(2):423–430, 2022.
26. Joseph A Sparano, Robert J Gray, Della F Makower, Kathleen I Pritchard, Kathy S Albain, Daniel F Hayes, Charles E Geyer Jr, Elizabeth C Dees, Edith A Perez, John A Olson Jr, et al. Prospective validation of a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 373(21):2005–2014, 2015.
27. Joseph A Sparano, Robert J Gray, Della F Makower, Kathleen I Pritchard, Kathy S Albain, Daniel F Hayes, Charles E Geyer Jr, Elizabeth C Dees, Matthew P Goetz, John A Olson Jr, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 379(2):111–121, 2018.
28. Maher Sughayer, Rolla Alaaraj, and Ahmad Alsughayer. Applying new magee equations for predicting the oncotype dx recurrence score. *Breast Cancer*, 25(5):597–604, 2018.
29. Antonio C Wolff, M Elizabeth Hale Hammond, Kimberly H Allison, Brittany E Harvey, Pamela B Mangu, John MS Bartlett, Michael Bilous, Ian O Ellis, Patrick Fitzgibbons, Wedad Hanna, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline focused update. *Archives of pathology & laboratory medicine*, 142(11):1364–1382, 2018.
30. B Yeo, L Zabaglo, M Hills, A Dodson, I Smith, and M Dowsett. Clinical utility of the ihc4+ c score in oestrogen receptor-positive early breast cancer: a prospective decision impact study. *British journal of cancer*, 113(3):390–395, 2015.
31. Alaa Zare, Lynne-Marie Postovit, and John Maringa Githaka. Robust inflammatory breast cancer gene signature using nonparametric random forest analysis. *Breast Cancer Research*, 23(1):1–6, 2021.
32. Domagoj Čevič, Loris Michel, Jeffrey Näf, Nicolai Meinshausen, and Peter Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression, 2021. URL <https://arxiv.org/abs/2005.14458>.