Distributional Random Forests to predict Oncotype DX scores

Réseau d'Intéractions Bio-Math de Besançon

Zeina Al Masry[1], Romain Pic[2], Clément Dombry[2], Christine Devalland[3]

[1]Institut FEMTO-ST (UBFC/CNRS/SUPMICROTECH-ENSMM)

[2]Laboratoire de Mathématiques de Besançon (CNRS/UBFC)

[3]Service d'anatomie et cytologie pathologiques (Hôpital Nord Franche-Comté)

femto-st SCIENCES & TECHNOLOGIES

(LmB) laboratoiredemathématiquesdebesançon
UNIVERSITÉ DE FRANCHE-COMTÉ • CNRS • UMR 6623

L'HÔPITAL Nord Franche-Comté

Table of Contents

1. Context

2. Method : Distributional Random Forests

3. Results

Context

How to assess the risk of cancer recurrence and potential benefit of adjuvant chemotherapy?

- Oncotype DX (ODX) test : Prognostic and predictive breast cancer information for **hormone positive, HER2-negative patients**
- **How?** — Analysis of 21 genes and give a recurrence score (0-100) : low risk (< 16), intermediate risk ($16 - 25$), high risk (> 25).
- Validated by several studies and recommended by the ASCO and the NCCN.

- High cost ⟶ not used routinely (less than 20% of patients in Europe)
- Current methods : use clinico-pathological features to predict the ODX score or probability of recurrence risk.

Goal — Predict the distribution of the ODX score <u>and</u> make the model explainable and understandable by practitioners.

Cohort

- **Who?** — 333 patients with ER-positive and HER2-negative early breast cancer.
- **Where?** — Three hospitals : Besançon, Belfort and Dijon.
- **When?** — Between 2012 and 2020.

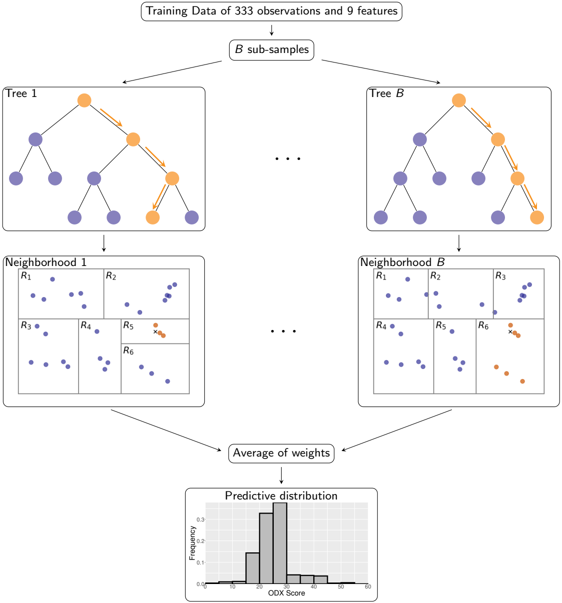
Predictors selected by variable importance and physicians' assessments :

- Age at diagnosis;
- Tumor size;
- Nottingham grade;
- SBR grade;
- ER status;
- PR status;
- Ki67 index proliferation cells;
- Protein p53;
- Lymph node status.

Table of Contents

1. Context
2. Method : Distributional Random Forests
3. Results

Method : Distributional Random Forests



Training Data of 333 observations and 9 features

B sub-samples

Tree 1

Tree B

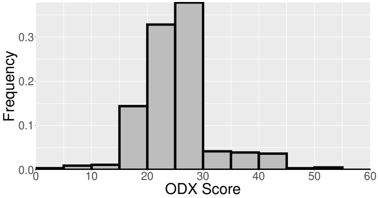
Neighborhood 1

Neighborhood B

Average of weights

Predictive distribution

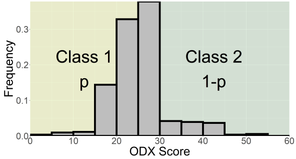
Output/Interpretation of DRF



Mean prediction
Uncertainty assessment

$$\left(\hat{Y},\ \hat{\sigma}_Y\right)$$

Classification



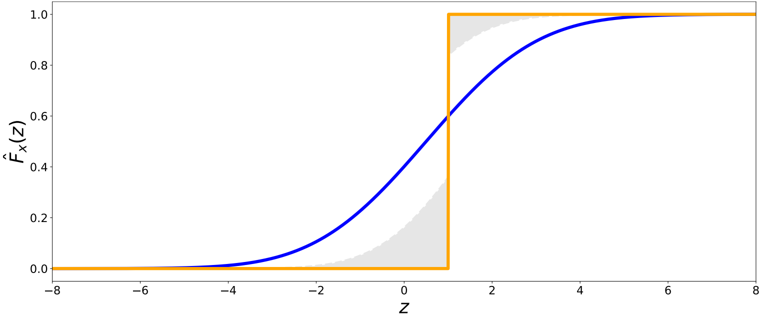
Most similar patients

i	w_i	Ki67	p53	...
1	0.2	17	8	...
2	0.18	20	2	...
...	...	...	...	...

Continuous Ranked Probability Score

- Continuous Ranked Probability Score (CRPS) : [Matheson and Winkler, 1976]

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{1}_{y \leq z}\right)^2 \mathrm{d}z$$



- The CRPS is lower for predictions that are **sharp and accurate**.

Table of Contents

1. Context

2. Method : Distributional Random Forests

3. Results

CRPS for OOB predictions

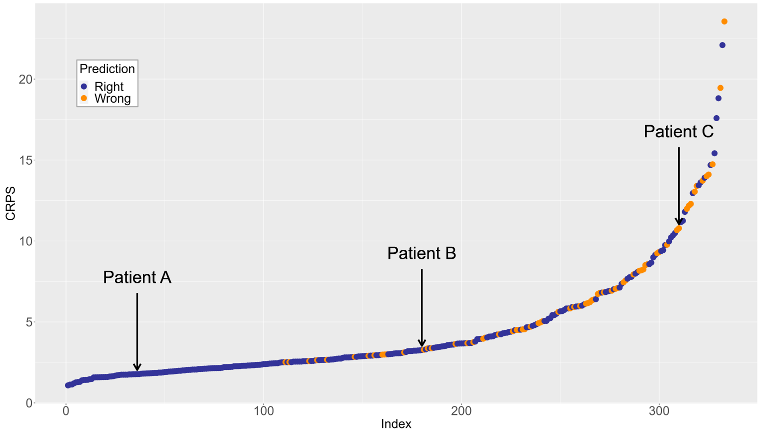


Figure: Sorted CRPS and low risk (≤ 25) and high risk (> 25) prediction.

Representative examples

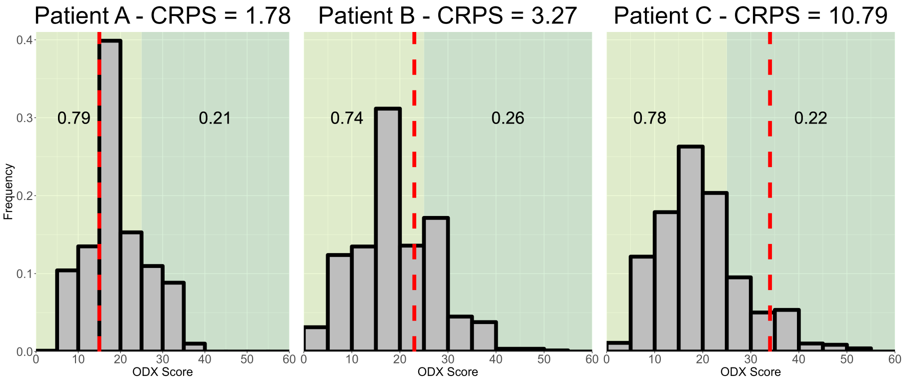


Figure: Three selected patients with a low, medium and high CRPS, respectively.

Limitations of the cohort

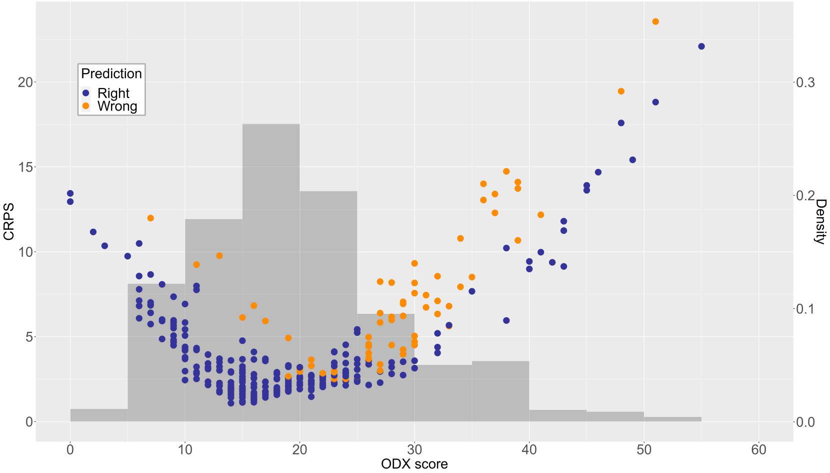


Figure: Comparison between CRPS vs ODX score and the density of ODX score in the cohort.

Comparison with state-of-the-art techniques

		Klein et al. (2013)	Hou et al. (2017)	Kim et al. (2019)	Orucevic et al. (2019)	Baltres et al. (2020)	Pawloski et al. (2021)	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
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

Table: Comparison of our study with six selected published studies to predict the ODX score. For three classes only the sensitivity and specificity of the lower class are given.

Conclusion

- New methodology for Oncotype DX score prediction : Distributional Random Forests.
- Explainability : neighborhood/weights, classification, mean/uncertainty prediction.
- Help oncologists in decision making regarding breast cancer therapy.

- **Perspectives**
  - Study the robustness with respect to noise or missing values.
  - Continue to develop an application to ease the use of DRF.

Preprint : A new methodology to predict the oncotype scores based on clinico-pathological data with similar tumor profiles, Al Masry et al. [HAL:04020992] [arXiv:2303.06966]

References I

- Baltres, Aline et al. (2020). "Prediction of Oncotype DX recurrence score using deep multi-layer perceptrons in estrogen receptor-positive, HER2-negative breast cancer". In: *Breast Cancer* 27.5, pp. 1007–1016. DOI: 10.1007/s12282-020-01100-4.
- Ćevid, Domagoj et al. (2022). "Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression". In: *Journal of Machine Learning Research*. arXiv: 2005.14458 [stat.ML].
- Hou, Yanjun et al. (2017). "Using the Modified Magee Equation to Identify Patients Unlikely to Benefit From the 21-Gene Recurrence Score Assay (Oncotype DX Assay)". In: *American Journal of Clinical Pathology* 147.6, pp. 541–548. DOI: 10.1093/ajcp/aqx008.
- Kim, Isaac et al. (2019). "A predictive model for high/low risk group according to oncotype DX recurrence score using machine learning". In: *European Journal of Surgical Oncology* 45.2, pp. 134–140. DOI: 10.1016/j.ejso.2018.09.011.
- Klein, Molly E et al. (2013). "Prediction of the Oncotype DX recurrence score: use of pathology-generated equations derived by linear regression analysis". In: *Modern Pathology* 26.5, pp. 658–664. DOI: 10.1038/modpathol.2013.36.
- Matheson, James E. and Robert L. Winkler (1976). "Scoring Rules for Continuous Probability Distributions". In: *Management Science* 22 (10). DOI: 10.2307/2629907.

References II

Orucevic, Amila et al. (2019). "Nomogram update based on TAILORx clinical trial results - Oncotype DX breast cancer recurrence score can be predicted using clinicopathologic data". In: *The Breast* 46, pp. 116–125. DOI: 10.1016/j.breast.2019.05.006.

Pawloski, Kate R. et al. (2021). "Supervised machine learning model to predict oncotype DX risk category in patients over age 50". In: *Breast Cancer Research and Treatment* 191.2, pp. 423–430. DOI: 10.1007/s10549-021-06443-w.