Machine learning versus mechanistic modeling for prediction of metastatic relapse in breast cancer

C. Nicolò
MONC (Modeling in ONCology)
Inria
Bordeaux, France

C. Périer
MONC (Modeling in ONCology)
Inria
Bordeaux, France

M. Prague
SISTM, Inria
Inserm U1219, Bordeaux Public Health
Bordeaux, France

G. MacGrogan
Pathology department
Bergonié Cancer Research Center
Bordeaux, France

O. Saut
MONC (Modeling in ONCology)
Inria
CNRS
Bordeaux, France

S. Benzekry
MONC (Modeling in ONCology)
Inria
Bordeaux, France
sebastien.benzekry@inria.fr

***Abstract*— Predicting the probability of metastatic relapse for patients diagnosed with early-stage breast cancer is essential for decision of adjuvant therapy. Current predictive models rely on biologically-agnostic, statistical models. Using the breast cancer database from the Bergonié Institute (n=1057 patients), we investigated the potential use of machine learning algorithms for predicting 5-years metastatic relapse. Furthermore, we developed a novel prediction tool based on an experimentally validated mechanistic model of metastatic development. Preliminary results indicate a higher positive predictive value when using this biologically-based model when compared to the alternative algorithms.**

Keywords— breast cancer, metastatic relapse, mechanistic modeling, machine learning

I. Introduction

The vast majority of breast cancer patients are diagnosed with localized, resectable disease [1]. However, for 90% of them, metastases are present at diagnosis, although not detectable by current diagnostic modalities. This is the reason why most of them undergo adjuvant therapy with either hormonotherapy, trastuzumab or cytotoxic therapy, depending on the molecular classification of their tumor. On one hand, current treatments still do not prevent relapse for some patients (about 20-30% of them, see Figure 1). On the other hand, a number of patients were probably overtreated, with associated severe toxicities, and might not have required adjuvant treatment, or not as extensive as what they received. A better prediction of the risk of metastatic relapse would therefore be crucial to personalize the adjuvant treatment and adapt the number of cycles required to prevent relapse and avoid unnecessary toxicities.

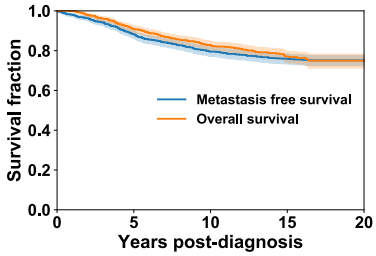

Figure 1: Kaplan-Meier overall and metastasis free survival for breast cancer. Data: Bergonié Institute historical cohort

Current prognostic models mostly rely on statistical models based on proportional hazard Cox regression [2]. The recent success of machine learning algorithms at prediction tasks interrogates whether these could be of use. To date, only few studies have investigated their relevance for survival predictions in oncology [3].

These classical models, either from survival analysis or machine learning theory, are biologically agnostic tools and operate as "black boxes" linking clinico-demographical and biological features of interest to outcome. In parallel to these, quantitative mathematical modeling of metastatic development has made recent progresses. A formalism first initiated for locoregional metastasis of a hepatocellular carcinoma was proposed to model the size distribution of secondary tumors [4]. Although model confrontation was only limited to a single patient in this original study, we have validated variants of this mechanistic model of metastatic dissemination and growth against large data sets from clinically relevant animal models of spontaneous metastasis after resection, including a breast cancer animal model [5]. However, due to experimental limitations, we could measure individual kinetics but only of the total metastatic burden. We have also validated our formalism against longitudinal human data of brain metastasis from non-small cell lung cancer in two patients [6].

The results presented here used this previously developed framework in order to derive a biostatistical model for personalized predictions of metastasis-free survival. We compared this approach to classical Cox regression and machine learning algorithms. The data consisted in the historical cohort of the Bergonié Cancer Research center which contains 25 features for 1057 patients.

II. Model

A. Mechanistic model

From our previously published and validated mechanistic model of metastatic development [5], the time to apparition of the first visible metastasis can be computed in a personalized fashion. Briefly, the time evolution of the population of metastases is modeled by means of a partial differential equation written on a size-structured density of metastases $\rho(t,v)$ (with v denoting the size/volume and t the time) [4]:

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

$$\begin{cases} \partial_t \rho(t,v) + \partial_v\big(\rho(t,v)g(v)\big) = 0 & t \in (0,+\infty),\ \ v \in (V_0,+\infty) \\ g(V_0)\rho(t,V_0) = d\left(V_p(t)\right) & t \in (0,+\infty) \\ \rho(0,v) = 0 & v \in (V_0,+\infty) \end{cases}$$

where g and d are the growth and dissemination rates. Of note, $d(v) = 0$ after surgery of the primary tumor, i.e. for $v \geq V_d$ with V_d the size of the primary tumor at diagnosis. The rates depend on patient specific parameters grouped in a vector $\theta \in \mathbb{R}^p$. V_0 is the minimal size of a metastasis (taken to be one cell). $V_p(t)$ is the volume of the primary tumor which follows the following ordinary differential equation:

$$\frac{dV_p}{dt} = g_p(V_p).$$

with g_p the growth coefficient of the primary tumor, having the same structural shape as g, but possibly different coefficients. From this model, we could mechanistically define the time of relapse as:

$$N_{vis}(t;\theta) = \int_{V_{vis}}^{+\infty} \rho(t,v;\theta)dv$$
$$T(\theta) = \min\{t > 0; N_{vis}(t;\theta) > 1\}$$

where V_{vis} is the minimal visible size with imaging devices and $N_{vis}(t)$ is thus the number of visible metastases at time t. This model has already been shown as able to describe metastatic burden dynamics in a clinically relevant animal model of breast metastasis that includes surgery [5] and we recently demonstrated its applicability to clinical data [6]. Simulations of the model were performed using python 3.7.

B. Statistical model

To fit the model to the training data sets, we used the mixed-effects modeling framework [7]. Briefly, it consisted in the following assumptions

$$T^i = T(\theta^i) + \varepsilon^i, \quad \varepsilon^i \sim \mathcal{N}(0,\sigma)$$
$$\ln(\theta^i) = \ln(\theta_{pop}) + \beta_1 x_1^i + \cdots + \beta_n x_n^i + \eta^i, \quad \eta^i \sim \mathcal{N}(0,\omega)$$

where θ^i is the individual set of parameters in individual i having covariates (features) $x_1^i, \cdots, x_n^i$. The variable ε^i defines the observation model (error). θ_{pop} is the vector of fixed-effects and η^i the vector of random effects specific to individual i, with standard deviation ω. Parameters were fitted using lilkelihood maximization implemented in the R package *saemix* [8]. Once fitted, the model gave personalized distributions of times to relapse in independent patients not used during the training phase, where randomness came from random effects, from which probabilities of 5-years metastatic relapse could be computed.

C. Machine learning algorithms

The machine learning algorithms we employed here could not handle censoring. Therefore, only patients without censoring were kept in the analysis, which only concerned 81 patients. Classification algorithms included logistic regression, support vector classification, k-neighbors, naïve bayes, random forest, gradient boosting and multi-layer perceptron. They were trained using the python package *scikit-learn*. We only report results for the two ones giving the best results (logistic regression and random forests).

III. RESULTS

We first had to identify the most relevant features to be kept for further analysis. This was performed by identifying significant features using multivariate Cox regression ((p<0.05). Among the 25 initial ones, only 13 were thus kept to be including for the machine learning algorithms. Biologically meaningful variables such as the receptor statuses (ER, PR and HER2) were also integrated in the analysis, even if not significant in the Cox regression. Machine learning algorithms integrated in this analysis were specific to classification. Therefore, for comparison purposes, we restricted ourselves to the task of prediction of 5-years metastatic relapse (yes/no). To prevent using the same set for training the models and prediction, we used 10-fold cross-validation. Classification results are reported in the Table 1. In terms of Receiver-Operator Curves (ROC, Figure 2), the logistic regression achieved the best results (AUROC = 0.727), although the other statistical models (Cox regression and random forest) had similar performances. So far, only the primary tumor size at diagnosis (which is the clinical variable most mechanistically linked to outcome in our mathematical model) was used in the reported results of the mechanistic model (no covariates). This might explain why predictive performances were inferior to the other models in terms of AUROC. However, accuracy was better than Cox regression and the random forest algorithm, and similar to the logistic regression. The most clinically relevant metric for practical applications in our case is the positive predictive value, i.e. the probability of developing metastases given that the model says so. Due to the small number of patients with metastatic relapse at 5-years (approximatively 10%), achieving a high score for this task is particularly hard. Notably, our mechanistic model – although including only one patient-specific feature (tumor size) – significantly outperformed the other algorithms with a PPV of 31.2% (Table 1).

Integration of the other covariates into the mechanistic model (in particular, dissemination-specific biomarkers such as the level of cadherins, available in our database) will be crucial to determine the true power of the mechanistic model in comparison to biologically agnostic machine learning and statistical models.

	AUROC	***Accuracy***	***NPV***	***PPV***
Random forest	0.727	87	91.7	22.5
Logistic regression	0.772	90.1	91	25
Cox regression	0.728	87.4	91.1	16.7
Mechanistic model	0.641	89.7	91.3	31.2

Table 1: Classification performances of the various models. AUROC = Area Under the ROC curve. NPV = Negative Predictive Value. PPV = Positive Predictive Value

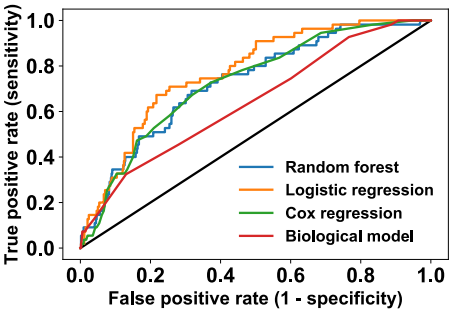

Figure 2: Receiver-Operator Curves for prediction of 5-years metastatic relapse

REFERENCES

[1] AM Noone, N Howlader, M Krapcho, D Miller, A Brest, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, HS Chen, EJ Feuer, KA Cronin (eds). SEER Cancer Statistics Review, 1975-2015, National Cancer Institute. Bethesda, MD, 2018.

[2] Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N., & Parker, H. L. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J Clin Oncol*, 19(4), 980–991.

[3] Gong, X., Hu, M., & Zhao, L. (2018). Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time‐to‐Event Analysis. *Clinical and Translational Science*, 11(3), 305‐311.

[4] Iwata, K., Kawasaki, K., & Shigesada, N. (2000). A Dynamical Model for the Growth and Size Distribution of Multiple Metastatic Tumors. *J Theor Biol*, 203(2), 177–186.

[5] Benzekry, S., Tracz, A., Mastri, M., Corbelli, R., Barbolosi, D., & Ebos, J. M. L. (2016). Modeling Spontaneous Metastasis following Surgery: An In Vivo-In Silico Approach. *Cancer Res*, 76(3), 535–547.

[6] Bilous, M., Serdjebi, C., Boyer, A., Tomasini, P., Pouypoudat, C., Barbolosi, D., et al. (2018). Computational modeling reveals dynamics of brain metastasis in non-small cell lung cancer and provides a tool for personalized therapy. bioRxiv, 448282.

[7] Lavielle, M. (2014). Mixed Effects Models for the Population Approach. CRC Press.

[8] Comets, E., Lavenu, A., & Lavielle, M. (2017). Parameter Estimation in Nonlinear Mixed Effect Models Using saemix, an R Implementation of the SAEM Algorithm. J Stat Softw, 80(1), 1–41.