Mechanistic modeling of metastatic relapse in early breast cancer to investigate the biological impact of prognostic biomarkers

Célestin BIGARRÉ, François Bertucci, Pascal Finetti, Gaëtan Macgrogan, Xavier Muracciole, Sébastien Benzekry

▶ **To cite this version:**

Célestin BIGARRÉ, François Bertucci, Pascal Finetti, Gaëtan Macgrogan, Xavier Muracciole, et al.. Mechanistic modeling of metastatic relapse in early breast cancer to investigate the biological impact of prognostic biomarkers. 2022. hal-03936594

HAL Id: hal-03936594
https://inria.hal.science/hal-03936594v1

Preprint submitted on 12 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mechanistic modeling of metastatic relapse in early breast cancer to investigate the biological impact of prognostic biomarkers

Célestin BIGARRÉ [1] ✉, **François BERTUCCI** [2, 3], **Pascal FINETTI** [2], **Gaëtan MACGROGAN** [4, 5], **Xavier MURACCIOLE** [1, 6 †], **Sébastien BENZEKRY** [1 †]

[1]COMPutational pharmacology and clinical Oncology Department, Inria Sophia Antipolis – Méditerranée; [2]Predictive Oncology laboratory, Marseille Cancer Research Centre (CRCM), Inserm U1068, CNRS UMR7258, Institut Paoli-Calmettes, Aix-Marseille University, Marseille, France. Equipe labellisée Ligue Nationale Contre Le Cancer.; [3]Department of Medical Oncology, CRCM, Institut Paoli-Calmettes, Aix-Marseille University, CNRS, Inserm, Marseille, France; [4]Department of Biopathology, Institut Bergonié, Regional Comprehensive Cancer Centre, Bordeaux, France; [5]Inserm U1218, Bordeaux Public Health, University of Bordeaux, Bordeaux, France; [6]Department of Radiotherapy, Assistance Publique - Hôpitaux de Marseille, Aix Marseille University, Marseille, France.

✉ **For correspondence:** celestin.bigarre@inria.fr (Célestin BIGARRÉ)

[†]Senior authors have contributed equally to this work

Present address: COMPO team, Faculté de pharamcie, 27 bd Jean Moulin, 13005 Marseille, FRANCE

Funding: Inria - Inserm "digital heath" Phd Grant (C. BIGARRÉ and S. BENZEKRY)

Competing interests: The authors declare no competing interests.

Abstract

Background and Objective

Estimating the risk of metastatic relapse is a major challenge to decide adjuvant treatment options in early-stage breast cancer (eBC). To date, distant metastasis-free survival (DMFS) analysis mainly relies on classical, agnostic, statistical models (e.g., Cox regression). Instead, we propose here to derive mechanistic models of DMFS.

Methods

The present series consisted of eBC patients who did not receive adjuvant systemic therapy from three datasets, composed respectively of 692 (Bergonié Institute), 591 (Paoli-Calmettes Institute, IPC), and 163 (Public Hospital Marseille, AP-HM) patients with routine clinical annotations. The last dataset also contained expression of three non-routine biomarkers. Using a mechanistic model of DMFS, we define two mathematical parameters of growth (α) and dissemination (μ). We identified their population distributions using the mixed-effects modeling. Critically, we propose a novel variable selection procedure allowing to: (i) identify the association of biological

parameters with either α, μ or both and (ii) generate an optimal candidate model for DMFS prediction.

Results

We found that Ki67 and Thymidine Kinase-1 were associated with α and nodal status and Plasminogen Activator Inhibitor-1 with μ. The predictive performances of the models were excellent in calibration but moderate in discrimination, with c-indices of 0.71 (95% confidence interval $[0.42, 0.99]$, AP-HM), 0.63 ($[0.44, 0.83]$, Bergonié) and 0.60 (95% CI $[0.54, 0.80]$).

Conclusions

Overall, we demonstrate that our novel method combining mechanistic and advanced statistical modeling is able to unravel the biological roles of clinic-pathological parameters from DMFS data.

Introduction

Breast cancer is the most common cancer amongst women and has a high survival probability at 5 years [1]. However, 15% of the patients with early-stage breast cancer (eBC) will suffer from distant metastatic relapse after surgery, with limited treatment options [2, 3]. Prevention of metastatic relapse is the purpose of adjuvant (post-operative) systemic therapies designed to eradicate the minimal residual disease. Such therapies, which include chemotherapy, and/or hormone therapy in hormone receptor-positive tumor and/or trastuzumab in human epidermal growth factor 2 (HER2)-positive tumors[4], have substantially improved the metastasis-free and overall survivals[5–7]. Nevertheless, the clinical outcome of eBC patients is heterogeneous. Current routine prognostic features are mainly age, lymph node status, tumor size and grade, and HR and HER2 statuses.

However, several critical issues remain such as the identification of patients who would have been cured by surgery and radiotherapy alone, thus avoiding the use of toxic chemotherapy[8]. The current relapse risk assessment models are simple regressions based on the above-cited biological parameters (BP). Examples are the Nottingham Prognostic Index [9] and the PREDICT score [10, 11]. As of today, multiparameter genomic tests with elaborate gene expression signatures (e.g. MammaPrint [12, 13], Oncotype DX [14] or Endopredict [15]) can be used in clinical practice for predicting the clinical course. However, their use remains limited, in part due to their expensive price. More recently, machine learning algorithms have being developed for prognosis [16], but few of them focused on the prediction of breast cancer relapse [17]. In addition, these approaches are agnostic and do not rely on biological knowledge.

In contrast, mechanistic models of metastatic development have been developed during the last decades, integrating the pathophysiology of the metastatic process [18]. They have been used to estimate the occult metastatic burden at diagnosis after the resection of the primary tumor [19–21], to predict the impact of individual treatments in pancreatic cancer [22] or to describe brain metastasis in lung cancer [23]. Following pre-clinical validation [24, 25], we showed in a previous work [26], that a mechanistic approach based on simulation of the metastatic disease could be used to create a predictive tool of breast cancer relapse. Building upon this work, we here tested our mechanistic models on three

datasets of eBC patients who did not receive any adjuvant systemic therapy. This allowed us to calibrate the models using mixed-effects modeling from data of the natural history of the disease. We show — through a careful univariate analysis — that our model is able to describe the biological links between BP and processes of the metastatic disease. Then we propose a model selection procedure to establish the best covariate structures to use in prediction. Eventually, we establish the predictive performances of the selected models for each dataset.

Materials and methods

Patient datasets

The data consisted of distant metastasis-free survival (DMFS) information and main prognostic variables for patients with operated eBC from three databases. Inclusion criteria were: invasive breast carcinoma, early-stage, treated with primary surgery followed or not by adjuvant radiotherapy, without any adjuvant systemic therapy (hormone therapy, chemotherapy, trastuzumab), with clinicopathological data and follow-up available for DMFS. The patients who did not experience distant metastatic relapse were censored at the time of death or last follow-up.

The first dataset contained data from 591 women who were treated at the Bergonié institute (Bordeaux, France) between 1989 and 1993. The clinicopathological parameters were: age at the time of diagnosis, pathological tumor size, axillary lymph node status, tumor grade, and expression of estrogen (ER) and progesterone (PR) receptors, HER2 and Ki67, based on immunohistochemistry (IHC) assays. The tumors were considered ER- or PR-positive when more than 1% of the cells showed expression of the corresponding receptor in IHC, Ki67 high when 14% or more of the cells expressed the marker (and Ki67 low otherwise), and HER2-positive when the IHC score was 3+ or 2+ with 60% or more of the cells expressing HER2.

The second dataset included data from 676 patients extracted from our clinically annotated database ($8,982$ invasive breast cancer samples) made from aggregation of 36 public gene expression datasets [27]. This set included the same clinicopathological annotations as the Bergonié set.

The third dataset was composed of 167 patients treated between 1980 and 1990 at the public hospital of Marseille (AP-HM), France. Information on individual DMFS, age status, pathological tumor size, axillary lymph node status and grade were available with the same definition as for the other data sets. Protein dosage information, based on biochemical assays, was available for ER, PR, Urokinase Plasminogen Activator (UPA) and Plasminogen Activator Inhibitor-1 (PAI-1) as well as the enzymatic activity for Thymidine Kinase 1 (TK). Tumors were considered ER- or PR-positive when the quantity of respective proteins was greater than 15 femtomoles per milligram of protein.

Missing values could not be imputed uniformly across all three datasets and were thus removed from each dataset: 55 patients were removed from the Bergonié dataset (646 patients initially), 16 from the IPC dataset (692 patients initially), and 7 from the AP HM dataset (174 patients initially).

Mechanistic modeling of the metastatic process

The model has been introduced and extensively described previously[26] but we include here a brief description for self-consistency. We consider a simple description of the natural history of eBC, starting at time $t = 0$ with one cell. The primary tumor (PT) grows following a Gompertz law [28]:

$$V_p(t) = \exp\left(\frac{\alpha}{b}(1 - e^{-bt})\right)$$

where V_p is the number of cells in the tumor at time t, α is the specific growth rate (i.e. $V_p^{-1} \cdot dV_p/dt$, expressed in d^{-1}) at 1 cell and b is the exponential decay parameter of the initial growth rate (unitless) [29]. These assumptions implicate that the tumor size converges to a theoretical limit $K = \exp(\alpha b)$ cells when $t \to +\infty$. To avoid over-parametrization and based on biological evidence[21, 30–32], we fixed K to 10^{12} cells and considered α as the only free parameter for growth (b being computed using the previous equation). At the time of diagnosis (t_{diag}), the patient undergoes a surgery and the PT is removed.

During the course of the pre-surgical period, we assume that all cells from the primary tumor have an instantaneous probability of dissemination of μ (expressed in $cell^{-1}\ d^{-1}$). The dissemination rate of the tumor is then:

$$d(V_p) = \mu V_p,$$

leading to the following continuous expression for the total number of metastases at time t:

$$N_{cont}(t) = \int_0^t d\left(V_p(s)\right)\, ds = \mu \int_0^t V_p(s)\, ds$$

To be consistent with the biological reality, the numerical implementation considers that the number of metastases is an integer, given by:

$$N(t) = \lfloor N_{cont}(t) \rfloor,$$

where $\lfloor x \rfloor$ is the integer part of x. The metastases are assumed to also follow a Gompertz growth law with the same parameters as for the PT[25].

To define the time to distant metastatic relapse (TTR), we considered that metastases are detected as soon as they reach a detectability threshold V_{detect} taken to correspond to a tumor of 5 mm (detecting limit in imaging) [33, 34]. From the size V_{diag} of the PT at diagnosis and the growth parameters, we can compute t_{diag}, the time between the initiation of the disease and the diagnosis and $\tau_v is$, the time needed for a metastasis to reach V_{detect} [26].

Since the first metastasis emitted will be the first to reach the visibility threshold, the time to relapse is given as a function of V_{diag} and the two mathematical parameters (MP) α and μ (see Figure A).

$$TTR(\alpha, \mu, V_{diag}s) = \begin{cases} \tau_{vis} + \arg\min_t\{N(t) \geq 1\} - t_{diag}, & \text{when } N(t_{diag}) \geq 1 \\ +\infty, & \text{otherwise} \end{cases}$$

Statistical mixed-effects population model

Given individual values α^i and μ^i for the i-th patient with observed size of the PT at diagnosis v^i_{diag}, we assumed a log-normal observation error model for the time to metastatic

relapse T^i, to ensure positivity,

$$\log(T^i) = \log\left(TTR(\alpha^i, \mu^i; V^i_{diag})\right) + \varepsilon^i,$$

where $\varepsilon^i \sim \mathcal{N}(0, \sigma^2)$ is the residual error with standard deviation σ.

We also assumed a log-normal distribution of the individual MPs, in the population, with a linear effect of the covariates (BP at diagnosis), denoted by the vector C^i:

$$\begin{cases} \log \alpha^i = \log \alpha_{pop} + \beta_\alpha \cdot C^i + \eta^i_\alpha \\ \log \mu^i = \log \mu_{pop} + \beta_\mu \cdot C^i + \eta^i_\mu \end{cases}$$

where α_{pop} and μ_{pop} are the typical values of α and μ in the population, β_α and β_μ are the vectors of the covariate effects, and $\eta^i = (\eta^i_\alpha, \eta^i_\mu)$ are the random-effects, i.e., independent identically distributed random variables with distribution $\mathcal{N}_2(0, \Omega)$. The latter quantify inter-individual variability.

From the survival function implicitly defined by the structural error model and using a likelihood definition compatible with censored data, we defined a maximum likelihood estimator for the mixed-effect model (see previous work for technical details [26]).

The relative standard-errors of the population-level MPs of the models — α_{pop}, μ_pop, Ω, and σ^2 — were obtained by 100-replicates bootstrap and used to assess parametric identifiability.

Variable selection

The definition of the covariate effects allows for each BP to potentially influence the distribution of α, μ, both, or none, depending on the corresponding coefficients β_α and β_μ. To identify the impact of the BPs on the MPs, we used a two-step approach. First, we performed a univariate analysis in which we tested for significant effects in either α or μ, using models including only one covariate on one MP. Specifically, for a BP C_k, we tested for the null hypotheses $H_0 : \beta_{k,\alpha} = 0$ or $\beta_{k,\mu} = 0$. The univariate models were assessed in 100-samples bootstraps. The standard deviation of the bootstrap distributions was used to evaluate the precision of the MP estimation. For each covariate, we tested if the corresponding coefficient was significantly non null with a Wald test using the bootstrap estimate of the coefficient standard error.

In the second step, the optimal covariate model for each dataset was selected using a backward elimination procedure based on the Bayesian information criteria (BIC) adapted for the selection of covariates in mixed-effects models [35]. Specifically, we started from a model containing all statistically significant covariates for both α and μ. Then we iteratively generated all possible nested models with one covariate less and selected the model with the minimal BIC.

To verify that β was correctly identifiable, the significance of each covariate coefficient in the selected multivariate model was then re-assessed in the multivariate model with a Wald test, based on 100-samples bootstrap estimation of the standard deviation.

Individual predictions

Individual predictions of survival curves $\hat{S}^i$ were obtained by taking the empirical expectation of the survival function with respect to the inter-individual variability over N_sim

replicates,

$$\hat{S}^i(t) = \frac{1}{N_{sim}} \sum_{j=1}^{N_{sim}} S\left(t|\, \alpha_j^i, \mu_j^i;\, V_{diag}^i\right),$$

with α_j^i and μ_j^i sampled from the distribution of α^i and μ^i.

Prediction performance metrics

To assess the prediction performance of the models at a fixed time point, we used calibration plots. These were obtained by predicting the survival probabilities at landmark times t_l for all patients $\hat{S}^i(t_l)$, binning these into 8 quantiles groups and computing the median prediction in each bin. For each bin, this prediction was plotted against the actual data group DMFS at time t_l estimated by the Kaplan-Meier method [36, 37].

Performance of the models for prediction was also assessed by the concordance index using the individual predicted DMFS probability at 5 years to order the comparable pairs [38].

Numerical implementation of the model

The mechanistic model was implemented as an R package with high performance simulation code in C++. All model simulations used a time step of 20 days and were performed with tumor size expressed in number cells. Tumor diameter data were converted into numbers of cells assuming spherical shape and a cell density of 10^6 cells per mm3 [32, 39].

Parameter identification was performed using the stochastic approximation of expectation maximization algorithm [40] implemented in the saemix R package version 3.0 [41]. All computations were performed with R version 4.0.4 [42].

Data availability

The data from Institut Bergonié and AP-HM analyzed in this study are not publicly available due to patient privacy requirements but are available upon reasonable request from the corresponding author. The data from IPC is available online and has been previously described [27].

Results

Biological parameters discrepancies between the datasets

The distributions of the BPs were different in the three datasets (Table 1). There was significantly more Ki67 high patients in the IPC dataset (chi-squared test, $p < 0.001$) and more node-positive patients in the Bergonié dataset when compared to the IPC dataset ($p < 0.001$) and the AP-HM dataset ($p < 0.001$). The distribution of the grade values was significatively different between the Bergonié and IPC datasets ($p < 0.001$), with more low-grade tumors in the Bergonié data and more high-grade tumors in the IPC data. The proportion of ER-positive and PR-positive patients across the three datasets was also significatively different and no statistical difference was found for the distribution of HER2 in the Bergonié and IPC data.

The distribution of primary tumor size appeared log-normal for the three datasets, with a smaller median in the Bergonié dataset than in the IPC set ($p < 0.001$, Brown-Mood

median test) whereas no statistical differences in median could be found between the IPC and AP-HM datasets (Supplementary Figure **??**). The three datasets also exhibited differences in DMFS (Supplementary Figure **??**), with lower DMFS in the IPC dataset than in the Bergonié dataset (Cox regression hazard ratio $HR = 2.2$, $p < 0.001$) or in the AP-HM dataset ($HR = 1.8$, $p < 0.001$). No significant difference in DMFS was found between the Bergonié and AP-HM datasets ($HR = 1.3$, $p = 0.157$).

Inter-individual variability of the mathematical parameters accurately describes distant metastasis-free survival curves

We first used our mechanistic model without covariates to see if individual variability in the pathological tumor size at diagnosis (included as a direct parameter of the model), associated to log-normal inter-individual variability of the MPs α and μ, was able to describe the observed TTR in the three datasets (1). Figures 1B, 1C, 1D show the descriptive performances of the mechanistic models. For all three models, the model-based population survival curves correctly described the observed DMFS data and remained within the confidence intervals of the Kaplan-Meier estimators, except for a slight underestimation of the DMFS for small times.

The population values $log\ \alpha_{pop}$ and $log\ \mu_{pop}$ were estimated with good accuracy. The relative standard error (RSE) was 10.7% on $log\ \alpha_{pop}$ in the AP-HM dataset and the RSEs for the remaining parameters were below 10% in all three datasets (Table 2).

Mechanistic modelling yields biological insight on the impact of biological parameters on metastasis

To study the association of the BPs with either growth or dissemination, we tested all univariate models with effect of a BP as a covariate on either α or μ (Table 2). On the AP-HM dataset, the TK concentration had a significant effect on α and the concentrations of PAI 1 and UPA had a significant effect on μ. On the Bergonié and IPC datasets, the ER, PR, HER2, and Ki67 statuses had a significant effect on α and μ, and the lymph node status had a significant effect on μ. The grade of the tumor was also significantly associated with both α and μ on the IPC dataset.

For each dataset, an optimal BP set was selected using a backward elimination procedure starting from the model including all significant BPs (Supplementary Table S1). Model selection was performed based on the BIC to compromise between model performances and number of covariates. For each dataset, the covariate model with the lowest BIC amongst the nested models was kept as the best model (Figure 2). Ties on the minimum (differences on the BIC < 4) were resolved by choosing the larger model, as the BIC is known to be favor smaller models when compared to other criteria [43]. The final models contained the ER and Ki67 statuses on α for the Bergonié dataset (Figure 2A), a more complex model with effects of the grade and PR on α and the HER2 and nodes statuses on μ for the IPC dataset (Figure 2B), and only PAI 1 on μ for the AP-HM dataset (Figure 2C).

Estimation of the population parameters for the selected models is presented in Table 3. The effects of PAI 1 on μ for the AP-HM data and grade 3 for the IPC dataset were estimated with good precision ($RSE < 30\%$), whereas the effects of the other variables were estimated with larger but still acceptable uncertainty ($RSE < 50\%$). Only the effect of the ER status on α for the Bergonié data was estimated with high RSE (52%). In the

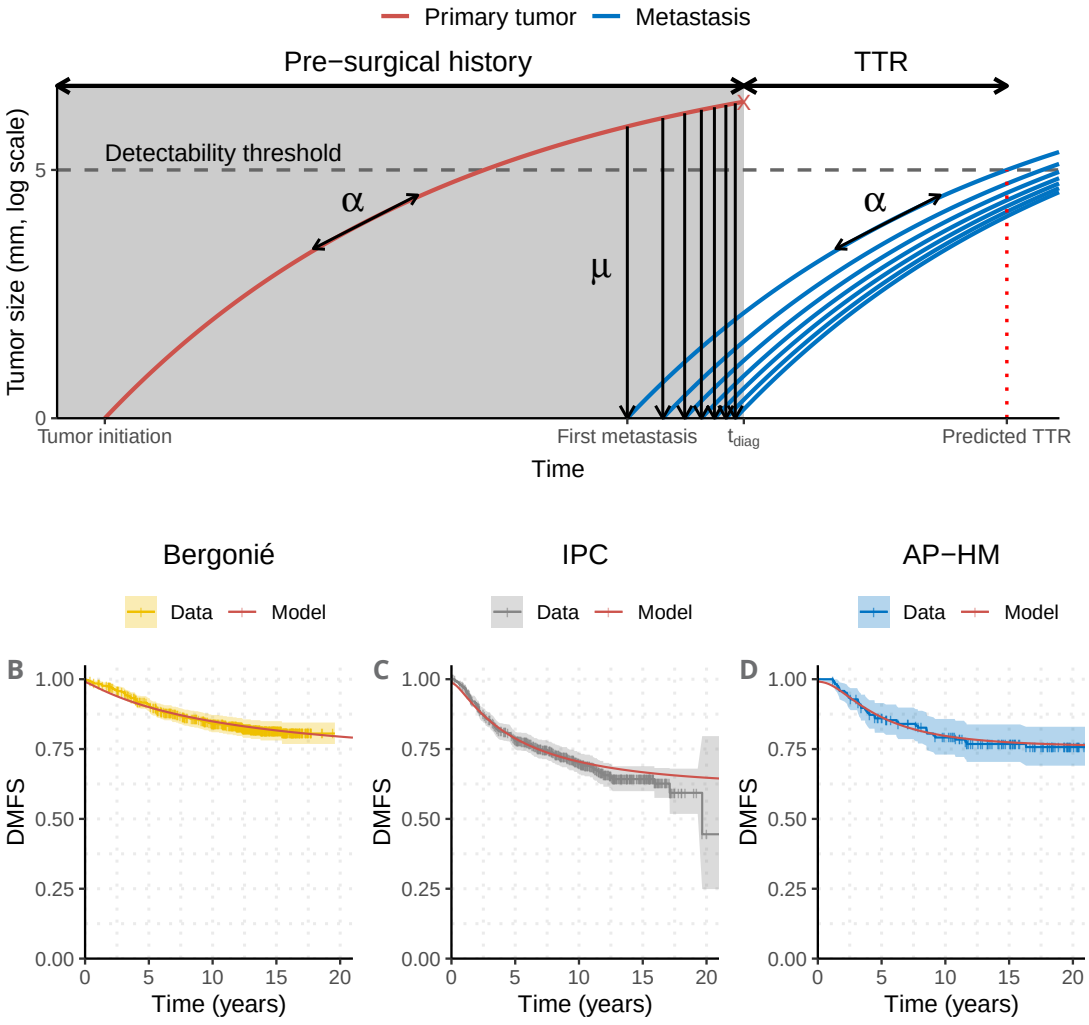


Figure 1. Mechanistic model of the time metastatic relapse.A. Overview of the mechanistic model. The model prediction of the time to metastatic relapse (TTR) is computed from the size of the primary tumor at diagnosis and two mathematical mechanistic parameters, α controlling the growth rate of the primary tumor and metastases, and μ controlling the seeding of new metastases. The scheme shows a unitless simulation of the model. B – D. Fits of the mechanistic model (without covariate effects) on the Bergonié (B), IPC (C) and AP-HM (D) datasets. Each panel presents for one dataset, the DMFS and the model's prediction of the survival function in the population. The model was trained and evaluated on the full dataset.

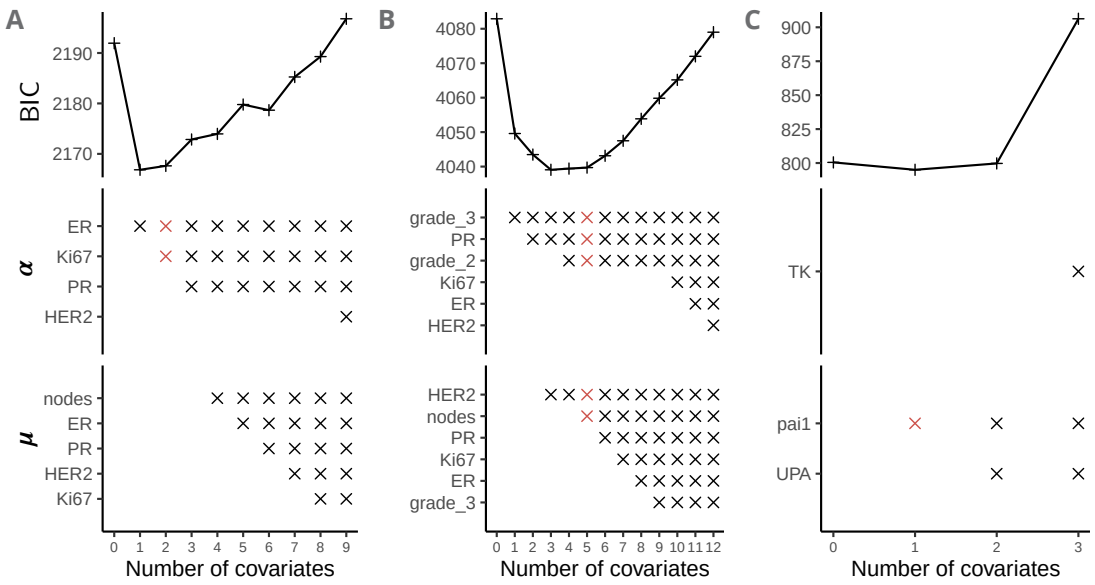


Figure 2. Biological parameters selection For each panel, the top line presents the evolution of the Bayesian information criteria (BIC) as a function of the number of parameters during the backward elimination procedure. Starting from the model containing all the covariates with a significant effect in univariate analysis, the selection procedure iteratively tested all models with one parameter less, keeping at each step the one with minimal BIC. The bottom lines show the evolution of the covariate model during the selection process. Backward selection on the Bergonié (A), IPC (B) and AP-HM, (C) data.

final models, all BPs were significant in multivariate analysis as covariates (Wald test). The coefficients of all BPs were in the same order of magnitude as the corresponding inter-individual variability standard deviation, confirming that the estimated coefficients have significant impact on the individual parameter distributions.

Covariate models accurately describe DMFS curves

To assess the ability of the models to describe different subpopulations of patients in each dataset, we analyzed the model predicted DMFS for different subgroups of patients (Figure 3). On the Bergonié dataset, the model captured well the DMFS difference between the ER-positive and ER-negative patients and between the Ki67 high and Ki67 low patients (Figure 3A-3B). On the IPC dataset, the model was able to correctly describe the differences of DMFS when comparing patients with different grades, HER2, based on the presence of invaded lymph nodes, or between PR-positive and PR-negative patients (Figure 3C-F). For the AP-HM dataset, the model captured very well the DMFS for patient in the upper PAI 1 tercile, was adequate for the second tercile but seemed to slightly overestimate DMFS for larger times, and the DMFS for the lower tercile of PAI 1 was underestimated by the model (Figure 3G).

Predictive performances

Next, we evaluated the prediction performances of the models. Calibration curves indicated excellent individual predictive power for all three models at the 5-, 10- and 15-years landmark times (Figure 4). Calibrations at 5 years (Figure 4A-C) and 10 years (Figure

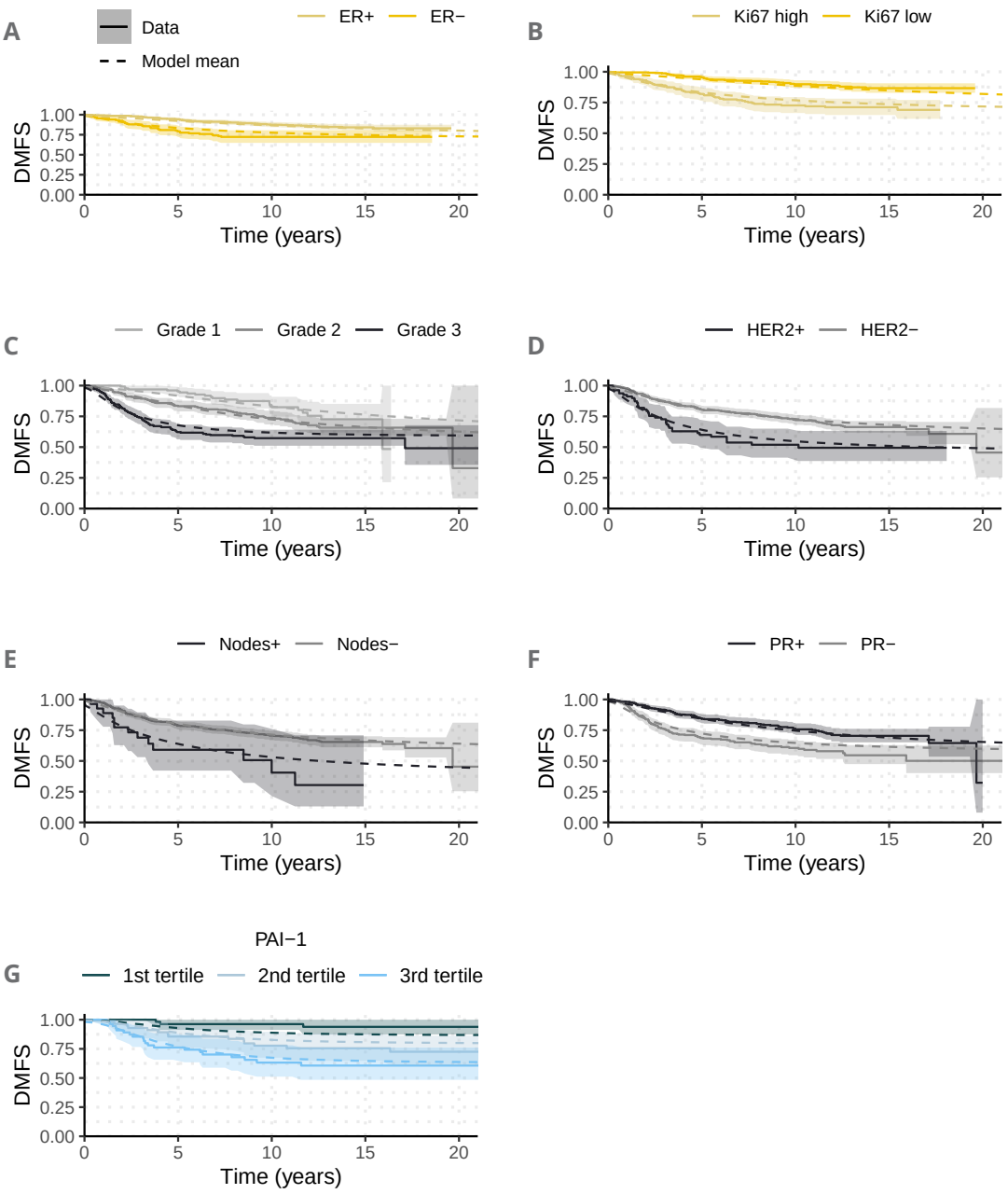


Figure 3. Model predictions in stratified groups. Group-comparison of the distant metastasis-free survival data (Kaplan-Meier estimate, solid line and 95% confidence interval, colored band) and the model mean prediction of the metastasis-free survival function. Bergonié dataset, patients stratified on the ER (A) or Ki67 (B) status. Paoli-Calmettes Institutes (IPC) dataset, patients stratified by grade (C), HER2 (D) status, invaded lymph nodes (E) or PR (F) status. G. APHM dataset, patients stratified by PAI-1 tertiles.

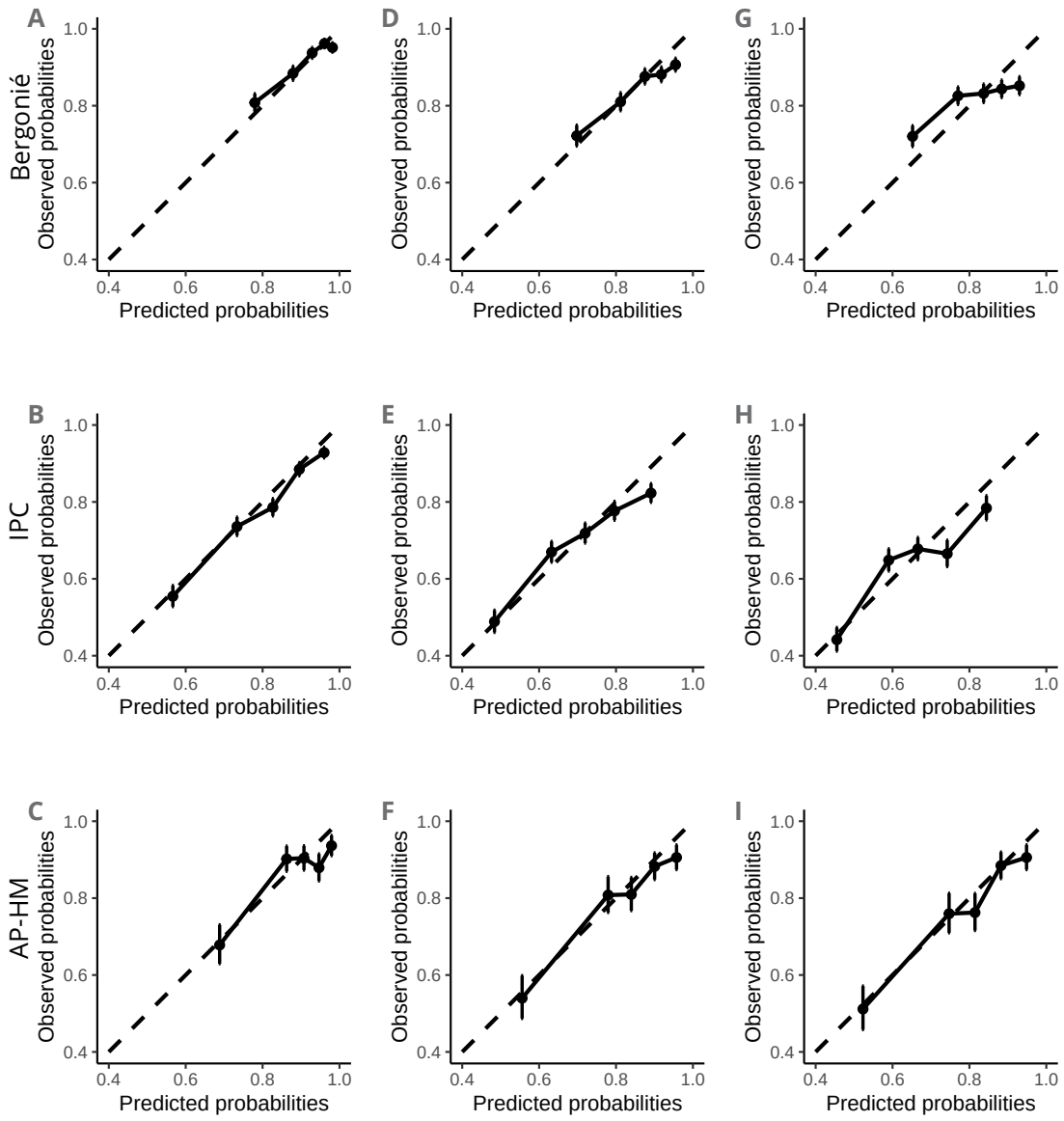


Figure 4. Calibration curves. At a fixed time-point, cross-validations predictions of the distant metastasis-free survival (DMFS) were binned into 5 quantile groups. The median prediction and 95% confidence interval of each group is compared to the Kaplan-Meier estimate of the group DMFS at the specified time. The identity (dashed line) is indicated for comparison to perfect prediction. (A-C): calibration curves at 5 years for the selected model on the respectively the Bergonié, IPC and AP-HM datasets (D-F): calibration curves at 10 years for the selected models (G-I), calibration curves at 15 years for the selected models.

4D-F) were good for all three models and for all DMFS probabilities, with a slight trend to overestimate DMFS probability at 10 years for the higher probability group. In the Bergonié dataset, the model slightly underestimated DMFS at 15 years for smaller probabilities while the over-estimation of higher probabilities was more pronounced (Figure 4G). Whereas for the IPC (4H) and AP-HM (Figure 4I) datasets, the calibration at 15 years was still good.

We then computed Harrell's concordance index (c-index) as another measure of the prediction performance [38]. The standard deviation for the c-index was computed in 10-folds (8-folds for the AP-HM dataset) cross-validation. The performances were modest with a c-index of 0.63 (95% confidence interval (CI) $[0.44-0.83]$) in the Bergonié dataset, 0.71 (95% CI $[0.42, 0.99]$) in the AP-HM dataset and 0.60 (95% CI $[0.54, 0.80]$) in the IPC dataset.

Discussion

Classical statistical models of metastatic risk, although able to detect correlation between biomarkers and outcome, fail to give causal insights about the mechanisms at stake. Several genomics-based prognostic tools are commercially available (Oncotype DX Recurrence Score, Prosigna Risk of Recurrence score, EndoPredict, Breast Cancer Index) for estimation of the recurrence risk in HR-positive and HER2-negative eBC. However, the cost of these tests limits their clinical use. Our approach helps to provide mechanistic information from routine clinical markers. We used a simple mechanistic model of metastatic development based on two processes – growth and dissemination – to analyze DMFS data from three different datasets of eBC patients, two of which contained only routinely available data, whereas the third contained also non-routine markers (UPA, PAI-1 and TK), known to have biological roles in the metastatic process [44]. We not only correctly described the DMFS in the three populations, but also showed that our method could be used to link biological features with specific parts of the metastatic process.

Studying the effect of the biological parameters on the population distribution of the growth parameter α and the dissemination parameter μ allowed us to associate each predictive feature with one (or both) aspect(s) of the metastatic process. Specifically, based on the data from the AP-HM cohort, our model supports the association of protease UPA and its inhibitor PAI 1 with metastatic dissemination potential. This is consistent with previous pre-clinical studies [45]. The association of Thymidine kinase 1, an enzyme involved in DNA synthesis allowing cell division [46] known to be associated with larger tumors at diagnosis in eBC [47], was found to impact the growth parameter α, again consistently with the biology. On the two other datasets (Bergonié and IPC), the model specifically associated the presence of invaded lymph nodes with μ. Overall, these findings support the ability of our mechanistic model to identify the biological role of specific markers.

Based on the first step of our univariate analysis, we proposed a variable selection method to establish the best combination of biological parameters to include in a predictive model for each dataset. Using the BIC, we selected the best compromise between goodness-of-fit and the number of biological parameters. The selected models also improved the insights given by the univariate approach. In the Bergonié dataset, the model pinpointed the Ki67 marker as an important predictor of the growth parameter α and as the least important predictor of the dissemination parameter μ, in accordance with the established biological role of Ki67 as a marker of cell division and tumor proliferation [48]. Similarly, in the IPC dataset, the lymph node status was selected in μ (and was the vari-

able to be eliminated in μ in the Bergonié dataset). The hormone receptor statuses were relevant in α for both the IPC (where PR status was selected) and the Bergonié (where ER status was selected) datasets. Their effect on μ was less clear, since no hormone receptor status was selected on μ but PR status persisted in μ up until late steps of the elimination process.

The prediction performances of the best model for each dataset were mitigated, with very good performances in calibration at various time points, but surprisingly low c-index values. The c-index computation may not be very accurate considering that the cross-validation sampling was not stratified on the event indicator variable, meaning that the number of events and thus the number of comparable pairs for the c-index computation was probably different between the cross-validation folds.

The differences in DMFS within each dataset (Supplementary Figure 2), the measurement methods (IHC staining for Bergonié, mRNA expression for IPC, and protein dosage for AP-HM) or variable availability prevented the possibility to properly compare the results across datasets. We chose to study routine clinical and biological data to gather as many patients as possible and to match as closely as possible the information available in the clinic.

The prediction at the individual level still needs further investigation to be up-part with the existing agnostic models. In particular a larger and more homogenous dataset with more patients and raw values of the markers instead of dichotomized categories could give better results. The next step of the development of our model should be to integrate the impact of adjuvant treatment on the individual risk of metastatic recurrence. With a better identification of the individual values of the MP in a mechanistic model of the metastatic process under the course of adjuvant therapies would give a relevant framework to tackle the problem of identifying patients who could avoid cytotoxic chemotherapy, or limit its extent to a minimal number of cycles.

References

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2022. *CA: A Cancer Journal for Clinicians* **72,** 7–33. ISSN: 1542-4863. https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21708 (2022) (2022).
2. Holleczek, B., Stegmaier, C., Radosa, J. C., Solomayer, E.-F. & Brenner, H. Risk of Loco-Regional Recurrence and Distant Metastases of Patients with Invasive Breast Cancer up to Ten Years after Diagnosis – Results from a Registry-Based Study from Germany. *BMC Cancer* **19,** 520. pmid: 31146706 (2019).
3. Steeg, P. S. Targeting Metastasis. *Nature Reviews Cancer* **16,** 201–218. ISSN: 1474-175X. pmid: 27009393 (2016).
4. Cardoso, F. *et al.* Early Breast Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Annals of Oncology* **30,** 1194–1220. ISSN: 0923-7534. pmid: 31161190 (2019).
5. Cameron, D. *et al.* 11 Years' Follow-up of Trastuzumab after Adjuvant Chemotherapy in HER2-positive Early Breast Cancer: Final Analysis of the HERceptin Adjuvant (HERA) Trial. *The Lancet* **389,** 1195–1205. ISSN: 0140-6736. pmid: 28215665 (2017).

6. (EBCTCG), E. B. C. T. C. G. *et al.* Comparisons between Different Polychemotherapy Regimens for Early Breast Cancer: Meta-Analyses of Long-Term Outcome among 100 000 Women in 123 Randomised Trials. *Lancet* **379,** 432–444. ISSN: 0140-6736. pmid: 22152853 (2012).
7. Organisation, N. A. T. Controlled Trial of Tamoxifen as Adjuvant Agent in Management of Early Breast Cancer Interim Analysis at Four Years. *The Lancet* **321,** 257–261. ISSN: 0140-6736 (1983).
8. Pondé, N. F., Zardavas, D. & Piccart, M. Progress in Adjuvant Systemic Therapy for Breast Cancer. *Nature Reviews Clinical Oncology* **16,** 27–44. ISSN: 1759-4774. pmid: 30206303 (2019).
9. Galea, M. H., Blamey, R. W., Elston, C. E. & Ellis, I. O. The Nottingham Prognostic Index in Primary Breast Cancer. *Breast Cancer Research and Treatment* **22,** 207–219. ISSN: 0167-6806. pmid: 1391987 (1992).
10. Wishart, G. *et al.* A Population-Based Validation of the Prognostic Model PREDICT for Early Breast Cancer. *European Journal of Surgical Oncology (EJSO)* **37,** 411–417. ISSN: 0748-7983. pmid: 21371853 (2011).
11. Wishart, G. C. *et al.* PREDICT: A New UK Prognostic Model That Predicts Survival Following Surgery for Invasive Breast Cancer. *Breast Cancer Research* **12,** R1. ISSN: 1465-5411. pmid: 20053270 (2010).
12. Buyse, M. *et al.* Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. *JNCI: Journal of the National Cancer Institute* **98,** 1183–1192. ISSN: 0027-8874. pmid: 16954471 (2006).
13. Van 't Veer, L. J. *et al.* Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* **415,** 530–536. ISSN: 0028-0836. pmid: 11823860 (2002).
14. Sparano, J. A. Gene Expression Assays in Early-Stage Breast Cancer. *Oncology (Williston Park, N.Y.)* **32.** ISSN: 0890-9091. pmid: 30334241 (2018).
15. Gradishar, W. J. *et al.* Predicting Expected Absolute Chemotherapy Treatment Benefit in Women with Early-Stage Breast Cancer Using a 12-Gene Expression Assay. *Journal of Clinical Oncology* **36,** 525–525. ISSN: 0732-183X (15_suppl 2018).
16. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine Learning Applications in Cancer Prognosis and Prediction. *Computational and Structural Biotechnology Journal* **13,** 8–17. ISSN: 2001-0370. pmid: 25750696 (2015).
17. Kim, W. *et al.* Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine. *Journal of Breast Cancer* **15,** 230–238. ISSN: 1738-6756. pmid: 22807942 (2012).
18. Benzekry, S. Artificial Intelligence and Mechanistic Modeling for Clinical Decision Making in Oncology. *Clinical Pharmacology & Therapeutics* **108,** 471–486. ISSN: 0009-9236. pmid: 32557598 (2020).
19. Hanin, L. & Rose, J. Suppression of Metastasis by Primary Tumor and Acceleration of Metastasis Following Primary Tumor Resection: A Natural Law? *Bulletin of Mathematical Biology* **80,** 519–539. ISSN: 0092-8240. pmid: 29302774 (2018).

20. Iwata, K., Kawasaki, K. & Shigesada, N. A Dynamical Model for the Growth and Size Distribution of Multiple Metastatic Tumors. *Journal of Theoretical Biology* **203,** 177–186. ISSN: 0022-5193. pmid: 10704301 (2000).

21. Koscielny, S., Tubiana, M. & Valleron, A. J. A Simulation Model of the Natural History of Human Breast Cancer. *British Journal of Cancer* **52,** 515–524. ISSN: 1532-1827. https://doi.org/10.1038/bjc.1985.222 (1985).

22. Haeno, H. *et al.* Computational Modeling of Pancreatic Cancer Reveals Kinetics of Metastasis Suggesting Optimum Treatment Strategies. *Cell* **148,** 362–375. ISSN: 0092-8674. pmid: 22265421 (2012).

23. Bilous, M. *et al.* Quantitative Mathematical Modeling of Clinical Brain Metastasis Dynamics in Non-Small Cell Lung Cancer. *Scientific Reports* **9,** 13018. ISSN: 2045-2322. https://www.nature.com/articles/s41598-019-49407-3 (2022) (1 2019).

24. Baratchart, E. *et al.* Computational Modelling of Metastasis Development in Renal Cell Carcinoma. *PLoS Computational Biology* **11,** e1004626. ISSN: 1553-734X. pmid: 26599078 (2015).

25. Benzekry, S. *et al.* Modeling Spontaneous Metastasis Following Surgery: An In Vivo-In Silico Approach. *Cancer Research* **76,** 535–547. ISSN: 0008-5472. pmid: 26511632 (2016).

26. Nicolò, C. *et al.* Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer. *JCO Clinical Cancer Informatics* **4,** 259–274. pmid: 32213092 (2020).

27. De Nonneville, A., Finetti, P., Mamessier, E. & Bertucci, F. RE: NDRG1 in Aggressive Breast Cancer Progression and Brain Metastasis. *JNCI: Journal of the National Cancer Institute,* djac031–. ISSN: 0027-8874. pmid: 35148398 (2022).

28. Norton, L. A Gompertzian Model of Human Breast Cancer Growth1. *Cancer Research* **48,** 7067–7071. ISSN: 0008-5472 (24_Part_1 1988).

29. Vaghi, C. *et al.* Population Modeling of Tumor Growth Curves and the Reduced Gompertz Model Improve Prediction of the Age of Experimental Tumors. *PLoS Computational Biology* **16,** e1007178. ISSN: 1553-734X. pmid: 32097421 (2020).

30. Coumans, F. A., Siesling, S. & Terstappen, L. W. Detection of Cancer before Distant Metastasis. *BMC Cancer* **13,** 283. ISSN: 1471-2407. https://doi.org/10.1186/1471-2407-13-283 (2013).

31. Klein, C. A. Parallel Progression of Primary Tumours and Metastases. *Nature Reviews Cancer* **9,** 302–312. ISSN: 1474-1768. https://doi.org/10.1038/nrc2627 (2009).

32. Spratt, J. A., von Fournier, D., Spratt, J. S. & Weber, E. E. Decelerating Growth and Human Breast Cancer. *Cancer* **71,** 2013–2019. ISSN: 0008-543X. https://doi.org/10.1002/1097-0142(19930315)71:6%3C2013::AID-CNCR2820710615%3E3.0.CO;2-V (2022) (1993).

33. Kundel, H. L. Predictive Value and Threshold Detectability of Lung Tumors. *Radiology* **139,** 25–29. ISSN: 0033-8419. pmid: 7208937 (1981).

34. MacMahon, H. *et al.* Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society. *Radiology* **237,** 395–400. ISSN: 0033-8419. https://pubs.rsna.org/doi/10.1148/radiol.2372041887 (2022) (2005).

35. Delattre, M., Lavielle, M. & Poursat, M.-A. A Note on BIC in Mixed-Effects Models. *Electronic Journal of Statistics* **8.** ISSN: 1935-7524 (2014).

36. Austin, P. C., Harrell, F. E. & Klaveren, D. Graphical Calibration Curves and the Integrated Calibration Index (ICI) for Survival Models. *Statistics in Medicine* **39,** 2714–2742. ISSN: 0277-6715. pmid: 32548928 (2020).

37. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53,** 457–481. ISSN: 0162-1459. JSTOR: 2281868 (1958).

38. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the Yield of Medical Tests. *JAMA* **247,** 2543–2546. ISSN: 0098-7484. pmid: 7069920 (1982).

39. Spratt, J. S., Meyer, J. S. & Spratt, J. A. Rates of Growth of Human Solid Neoplasms: Part I. *Journal of Surgical Oncology* **60,** 137–146. ISSN: 0022-4790. https://doi.org/10.1002/jso.2930600216 (2022) (1995).

40. Lavielle, M. *Mixed Effects Models for the Population Approach* (2014).

41. Comets, E., Lavenu, A. & Lavielle, M. Parameter Estimation in Nonlinear Mixed Effect Models Using Saemix, an R Implementation of the SAEM Algorithm. *Journal of Statistical Software* **80,** 1–41 (2017).

42. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2021).

43. Stoica, P. & Selen, Y. Model-Order Selection: A Review of Information Criterion Rules. *IEEE Signal Processing Magazine* **21,** 36–47. ISSN: 1558-0792 (2004).

44. Jänicke, F. *et al.* Randomized Adjuvant Chemotherapy Trial in High-Risk, Lymph Node-Negative Breast Cancer Patients Identified by Urokinase-Type Plasminogen Activator and Plasminogen Activator Inhibitor Type 1. *JNCI: Journal of the National Cancer Institute* **93,** 913–920. ISSN: 0027-8874. pmid: 11416112 (2001).

45. Duffy, M. J., McGowan, P. M., Harbeck, N., Thomssen, C. & Schmitt, M. uPA and PAI-1 as Biomarkers in Breast Cancer: Validated for Clinical Use in Level-of-Evidence-1 Studies. *Breast Cancer Research* **16,** 428. pmid: 25677449 (2014).

46. Sherley, J. L. & Kelly, T. J. Regulation of Human Thymidine Kinase during the Cell Cycle. *Journal of Biological Chemistry* **263,** 8350–8358. ISSN: 0021-9258 (1988).

47. Broët, P. *et al.* Thymidine Kinase as a Proliferative Marker: Clinical Relevance in 1,692 Primary Breast Cancer Patients. *Journal of Clinical Oncology* **19,** 2778–2787. ISSN: 0732-183X. pmid: 11387348 (2001).

48. Dowsett, M. *et al.* Assessment of Ki67 in Breast Cancer: Recommendations from the International Ki67 in Breast Cancer Working Group. *JNCI: Journal of the National Cancer Institute* **103,** 1656–1664. ISSN: 0027-8874. pmid: 21960707 (2011).

	Bergonié, N = 591	**IPC**, N = 676	**APHM**, N = 167	**p-value**
Pathological tumor size	15 (12, 20)	20 (15, 25)	20 (15, 26)	
Age				<0.001
<50	147 (25%)	338 (71%)	35 (21%)	
≥50	444 (75%)	138 (29%)	132 (79%)	
Unknown	0	200	0	
Pathological lymph nodes				<0.001
nodes-	364 (62%)	648 (96%)	147 (88%)	
nodes+	227 (38%)	28 (4.1%)	20 (12%)	
ER status				0.038
ER-	134 (23%)	140 (21%)	50 (30%)	
ER+	457 (77%)	536 (79%)	117 (70%)	
PR status				<0.001
PR-	189 (32%)	271 (40%)	80 (48%)	
PR+	402 (68%)	405 (60%)	87 (52%)	
HER2 status				0.7
HER2-	523 (88%)	594 (88%)		
HER2+	68 (12%)	82 (12%)		
Grade				<0.001
1	182 (31%)	131 (19%)		
2	267 (45%)	298 (44%)		
3	142 (24%)	247 (37%)		
Ki67 status				<0.001
low	390 (66%)	273 (40%)		
high	201 (34%)	403 (60%)		
TK			60 (31, 154)	
UPA			0.79 (0.38, 1.33)	
PAI-1			3.4 (2.0, 5.3)	

Table 1. Patient and disease characteristics. For categorical variables (age, presence of invaded lymph nodes, ER, PR , HER2 and Ki67 statuses, grade), number of patients in each dataset (and proportion of the dataset's values (%)). For continuous variables (pathological tumor size and PAI-1), median value (first quartile – third quartile) in each dataset. P-values correspond to the adequate test of identical distributions in all datasets with available information (chi-squared test for categorical variables, Kruskal-Wallis rank sum test for continuous variables).

Dataset		Base model estimation (RSE)	Full model estimation (RSE)
Bergonié	$\log \alpha_{pop}$	-7.12 (8.69%)	-5.3 (25.4%)
	$\log \mu_{pop}$	-26.1 (6.38%)	-29.7 (1.77%)
	σ	1.02 (18.9%)	0.571 (86.4%)
	ω_α	1.96 (13.3%)	1.05 (257%)
	ω_μ	2.90 (39.7%)	4.94 (1.69%)
	$\beta_{\alpha,\text{ER}}$		-0.99 (53.2%)
	$\beta_{\alpha,\text{Ki67}}$		1.38 (35.8%)
IPC	$\log \alpha_{pop}$	-4.43 (4.43%)	-5.28 (4.43%)
	$\log \mu_{pop}$	-29.1 (1.48%)	-29.2 (0.637%)
	σ	0.47 (20.6%)	0.511 (48.6%)
	ω_α	1.03 (14.2%)	0.734 (36.3%)
	ω_μ	4.23 (10.0%)	4.2 (9.06%)
	$\beta_{\alpha,\text{PR}}$		-0.549 (34%)
	$\beta_{\alpha,\text{grade 2}}$		0.61 (43.7%)
	$\beta_{\alpha,\text{grade 3}}$		1.67 (14.8%)
	$\beta_{\mu,\text{HER2}}$		1.99 (31.6%)
	$\beta_{\mu,\text{nodes}}$		2.72 (47.4%)
AP-HM	$\log \alpha_{pop}$	-4.4 (10.7%)	-4.3 (9.85%)
	$\log \mu_{pop}$	-30.5 (2.7%)	-32 (2.89%)
	σ	0.754 (16.7%)	0.0364 (251%)
	ω_α	0.0905 (368%)	0.722 (10.8%)
	ω_μ	2.97 (21.6%)	3.27 (9.74%)
	$\beta_{\mu,\text{PAI-1}}$		0.30 (30.2%)

Table 2. Values of the Parameters. The mathematical mechanistic parameters (α and μ) were assumed to follow a log-normal distribution such that $\log\ \alpha^i$ and $\log\ \mu^i$ are gaussian with respective mean $\log\ \alpha_{pop}$ and $\log\ \mu_{pop}$, and respective standard deviation ω_α and ω_μ. For each dataset, the base models correspond to the case with no BP effect (aside from the pathological tumor size). The full models correspond to the best models from the selection procedure, where conditionally to the vector of covariates (C^i) included in the model, the mathematical parameters followed a log-normal distribution such that, $\log\ \alpha^i$ and $\log\ \mu^i$ are gaussian with respective mean $\log\ \alpha_{pop} + \beta_\alpha \cdot C^i$ and $\log\ \mu_{pop} + \beta_\mu \cdot C^i$ (where β_α and β_μ are vectors of the BP specific coefficients) and respective standard deviation ω_α and ω_μ. For all models, the log-residual error on time to distant metastatic relapse was assumed to follow a centered gaussian distribution with variance σ^2. Estimation was performed using the stochastic approximation of expectation maximization algorithm. Relative standard errors (RSE) were computed from a 100 replicates bootstrap.

Appendix 1

Supplementary Tables

The univariate models assumed a log-normal distribution independent of the covariate for one of the two computational biomarkers (α or μ), and a log-normal conditional distribution with respect for the covariate for the other computational biomarker, with a median equal to the sum of a typical population value, and of the weighted covariate value. The estimation of the weight coefficient (beta), the relative standard error (R.S.E.) obtained by bootstrap with 100 repetitions, as well as the p-values for the corresponding Wald test, are presented in the table for all possible univariate models.

		Bergonié		
		β	S.E.	p-value
α	Age ≥ 50	−0.02	0.02	0.5
	nodes+	−0.5	0.4	0.2
	ER+	−2	0.5	0.0002
	PR+	−2	0.5	8×10^{-5}
	HER2+	2	0.5	0.0009
	Grade 2	−0.4	0.4	0.2
	Grade 3	0.7	0.4	0.1
	Ki67 high	2	0.4	3×10^{-7}
μ	Age ≥ 50	−0.02	0.02	0.5
	nodes+	−2	0.7	0.03
	ER+	−4	0.8	2×10^{-6}
	PR+	−3	0.7	8×10^{-5}
	HER2+	4	1	0.0006
	Grade 2	−1	0.8	0.08
	Grade 3	0.7	0.8	0.4
	Ki67 high	3	0.8	2×10^{-5}

Appendix 1—table 1. Univariate effects of the BCP, Bergonié dataset.

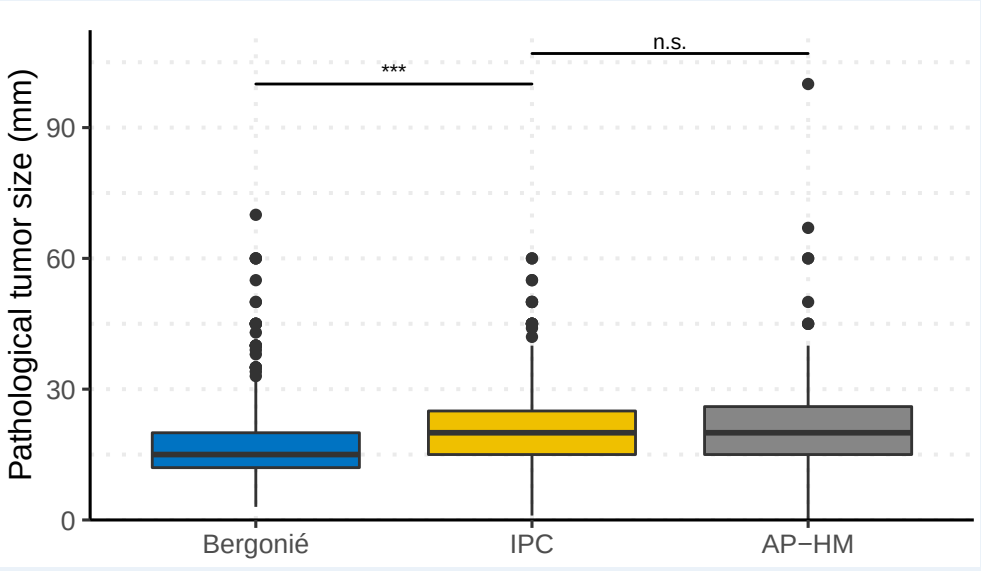
		IPC		
		β	S.E.	p-value
	nodes+	0.4	0.4	0.3
	ER+	–0.8	0.2	0.0003
	PR+	–1	0.5	0.01
	HER2+	0.9	0.4	0.05
	Grade 2	–0.8	0.2	0.001
α	Grade 3	1	0.2	5×10^{-10}
	Ki67 high	0.9	0.3	0.006
	nodes+	3	1	0.02
	ER+	–1	0.5	0.01
	PR+	–2	0.4	1×10^{-5}
	HER2+	3	0.6	1×10^{-6}
	Grade 2	–0.9	0.5	0.06
μ	Grade 3	2	0.5	5×10^{-5}
	Ki67 high	1	0.5	0.007

Appendix 1—table 2. Univariate effects of the BCP, IPC dataset.

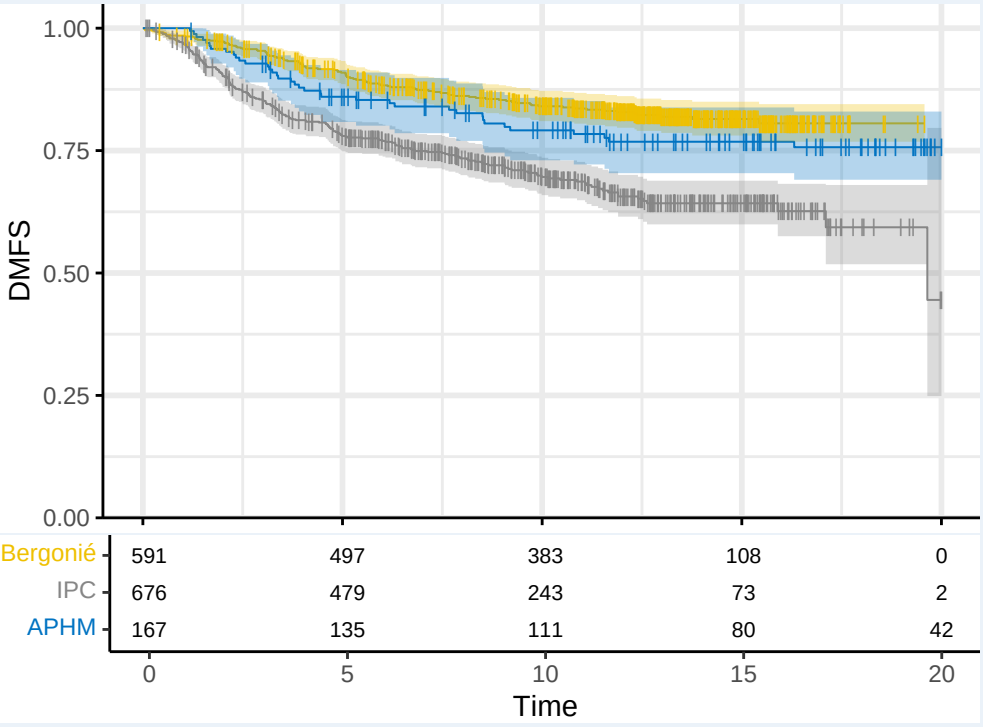
		APHM		
		β	S.E.	p-value
	Age ≥ 50	–0.004	0.02	0.8
	nodes+	–0.6	0.5	0.3
	ER+	–0.6	0.5	0.2
	PR+	–0.4	0.7	0.6
	TK	0.09	0.009	3×10^{-25}
α	UPA	0.1	0.2	0.4
	Pai-1	0.07	0.07	0.3
	Age ≥ 50	–0.05	0.04	0.1
	nodes+	–0.9	1	0.5
	ER+	–0.7	0.8	0.4
	PR+	–1	0.7	0.07
	TK	0.005	0.03	0.9
μ	UPA	0.7	0.3	0.03
	Pai-1	0.3	0.1	0.001

Appendix 1—table 3. Univariate effects of the BCP, AP-HM dataset.

Supplementary Figures



Appendix 1—figure 1. Distribution of the pathological tumor sizes (mm) in each of the three datasets. Statistical analysis was performed with Brown-Mood Median Test (*** $p < 0.001$, n.s. non-significant)



Appendix 1—figure 2. Kaplan-Meier estimate of the distant metastasis-free survival in the datasets.