



HAL
open science

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

Sebastien Benzekry

► **To cite this version:**

Sebastien Benzekry. Machine learning versus mechanistic modeling for prediction of metastatic relapse in breast cancer. 2019 CSBC-PSON Mathematical Oncology Meeting, May 2019, Portland, United States. hal-02424419

HAL Id: hal-02424419

<https://inria.hal.science/hal-02424419>

Submitted on 27 Dec 2019

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.

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

S. Benzekry

Inria MONC

Bordeaux, France

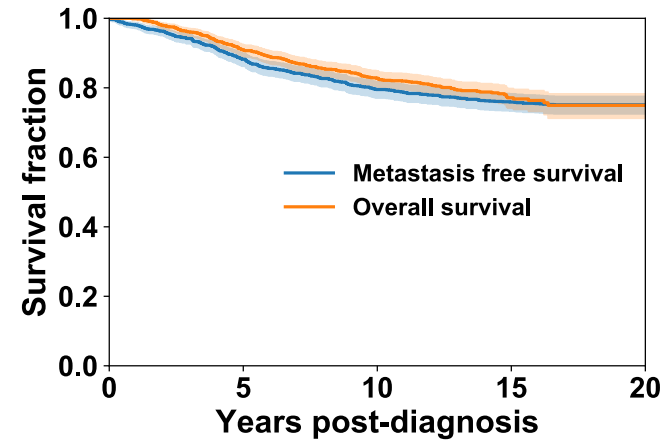
Inria

MONC

Modeling in ONCology

Breast cancer

- 94% of cases are diagnosed as local or regional but ~30% will relapse
- Estimation of the metastatic risk is key to **individualize adjuvant therapy**
- Reduce the number of chemo cycles for patients with low risk



Objectives

- Use a **mechanistic model** to predict metastasis
- Compare to standard survival methods and **machine learning** algorithms

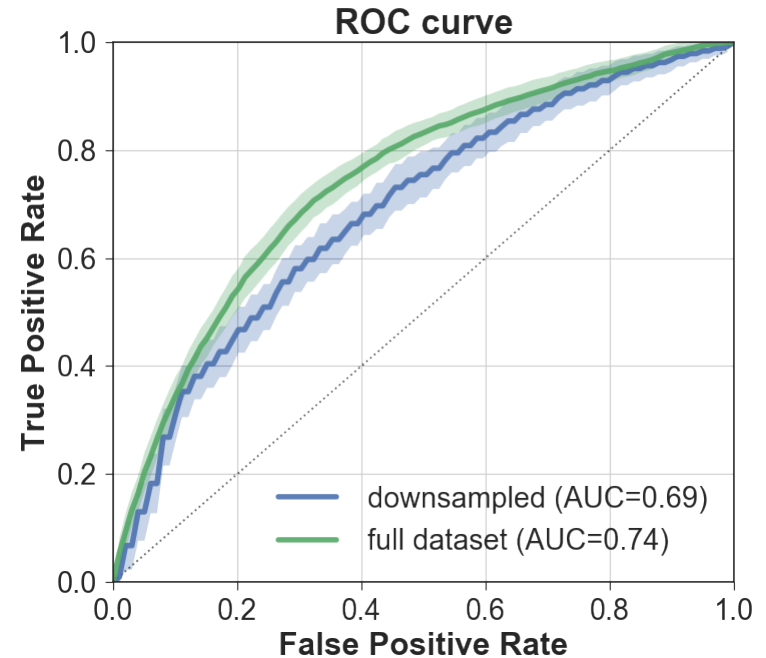
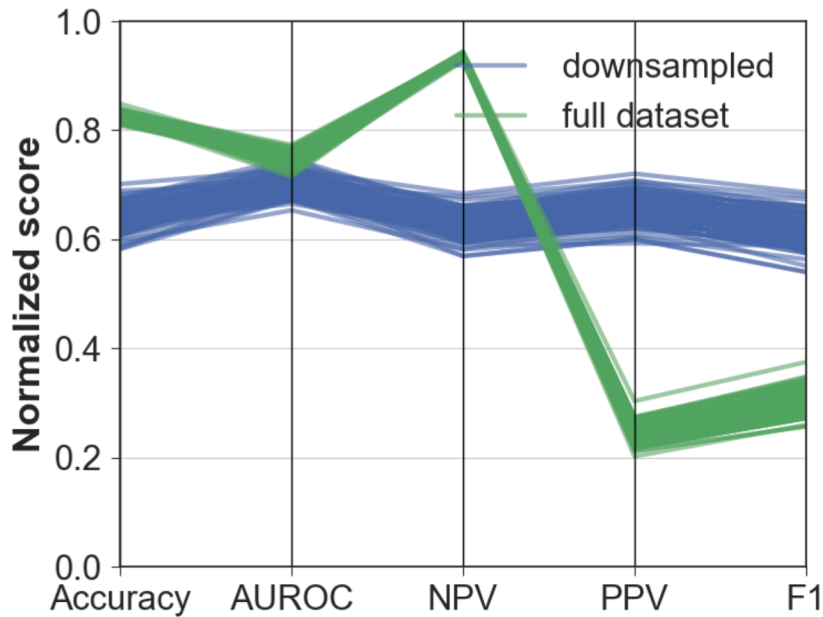
n = 1057 patients (642 w/o adj)

K = 25 features

menopausal_status	ER	PR	Ki67	HER2	HER2_intensity	CK56	EGFR	VIM	ALDH1
Post-ménopause	20	0	0	0	0	0	0	0	0
Ménopause	40	95	8	0	0	0	0	0	0
Activité génitale	87	10	26	0	0	0	0	80	0
Post-ménopause	100	100	8	0	0	0	0	0	0
Post-ménopause	0	0	16	82	+++	0	0	0	0
Activité génitale	100	95	12	0	0	0	0	0	1
Activité génitale	56	100	17	0	0	0	0	0	0
Activité génitale	57	85	23	100	+++	0	0	0	0
Post-ménopause	80	5	20	0	0	0	0	0	0
Post-ménopause	0	0	15	100	+++	0	5	0	0
Post-ménopause	100	80	10	0	0	0	0	0	0
Post-ménopause	30	0	5	0	0	0	0	0	0
Post-ménopause	0	0	15	40	+++	0	0	0	0
Ménopause	0	80	8	0	0	0	0	0	0
Post-ménopause	0	0	27	0	0	0	30	0	1
Post-ménopause	0	0	56	0	0	80	60	100	0
Activité génitale	50	92	2	1	+	0	0	0	0
Post-ménopause	0	47	5	0	0	0	0	80	0
Post-ménopause	65	0	10	0	0	0	0	60	0
Post-ménopause	100	50	11	0	0	0	0	0	0
Ménopause	20	100	0	0	0	0	0	0	0
Activité génitale	90	6	5	0	0	0	0	0	0
Post-ménopause	100	3	5	0	0	0	0	0	0
Activité génitale	0	0	6	0	0	0	0	0	0
Ménopause	80	100	5	0	0	0	0	0	0
Post-ménopause	100	85	25	0	0	0	0	0	0
Post-ménopause	10	45	11	13	+++	0	0	0	0
Post-ménopause	66	1	2	40	++	0	0	0	0

Machine learning for 5-years relapse

Random Forest

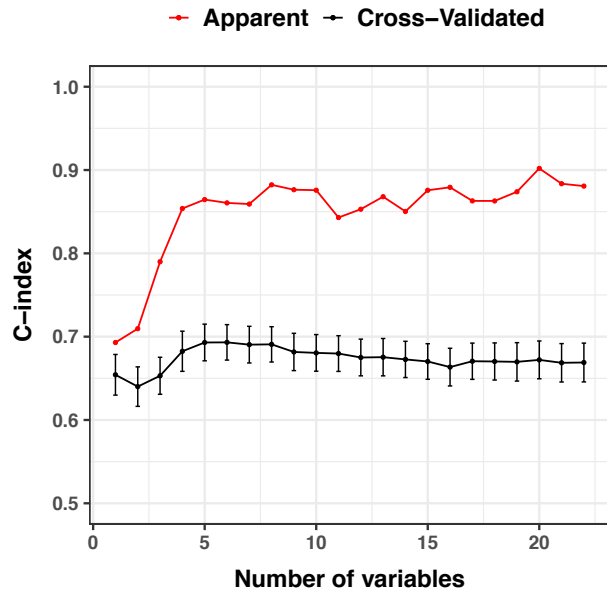


- Classification problem
- Imbalanced data
- Algorithms tested: logistic regression, NN, naïve Bayes, gradient boosting, SVM,...

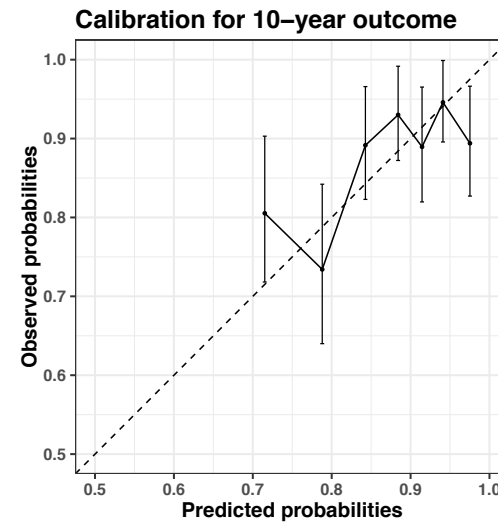
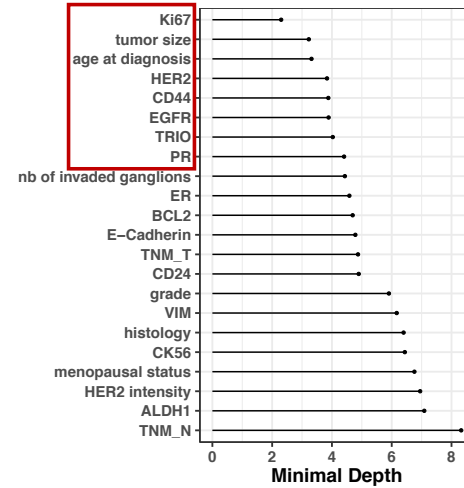
Does not account for full survival nor censorship ☹️

Random survival forests

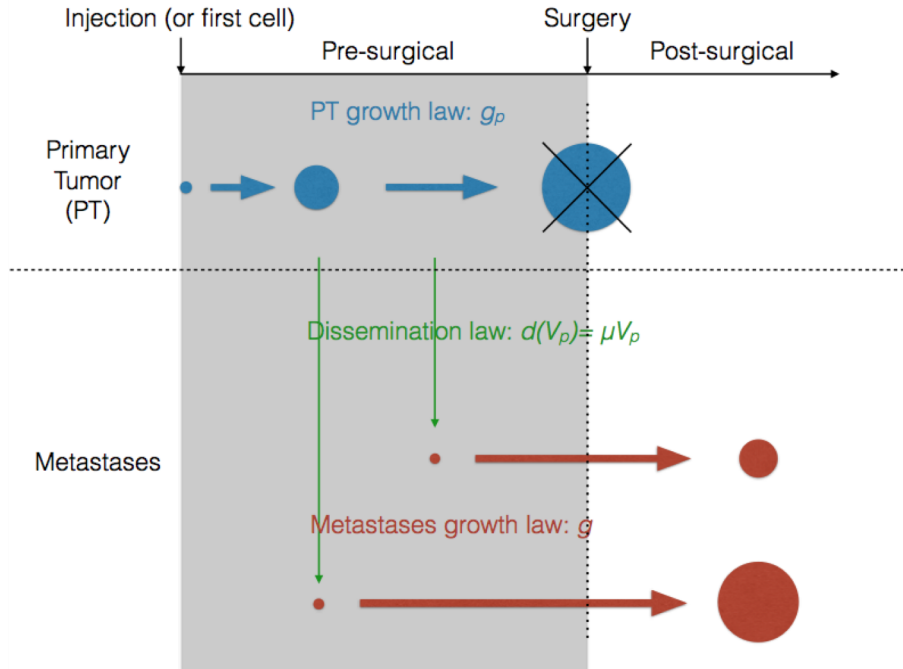
Accounts for censoring 😊



c-index = 0.69



Mechanistic model of metastatic dissemination and growth



Growth rates of primary and secondary tumors g_p and g

$$\frac{dV_p}{dt} = (\alpha_p - \beta_p \ln(V_p)) V_p \quad (\text{Gompertz})$$

$$g(v) = (\alpha - \beta \ln(v)) v$$

Dissemination rate

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

Size distribution of the metastases $\rho(t, v)$

$$\begin{cases} \partial_t \rho(t, v) + \partial_v (g(v) \rho(t, v)) = 0 \\ g(V_0) \rho(t, V_0) = d(V_p(t)) \\ \rho(0, v) = \rho^0 \end{cases}$$

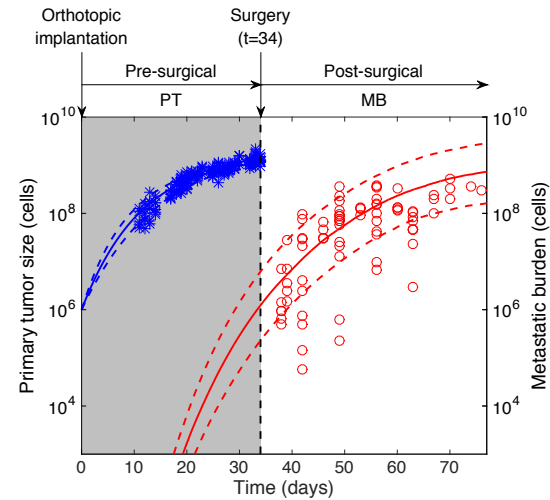
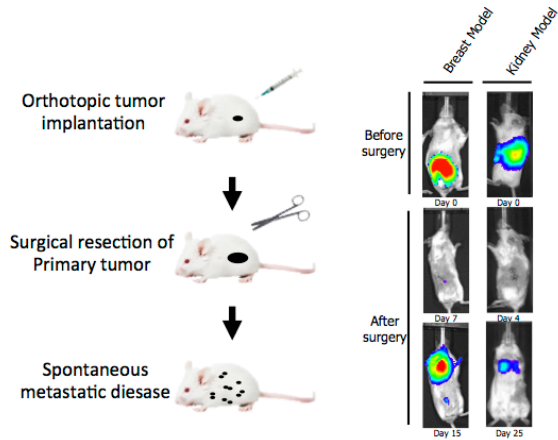
Metastatic burden (total number of metastatic cells)

$$M(t) = \int_{V_0}^{+\infty} v \rho(t, v) dv = \int_0^t d(V_p(t-s)) V(s) ds$$



Ebos lab
Roswell Park Cancer Institute

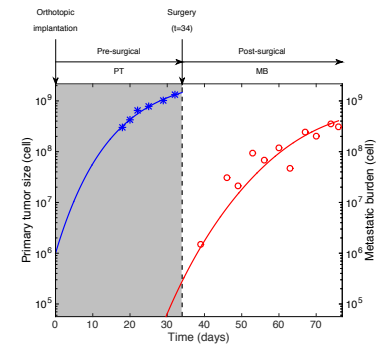
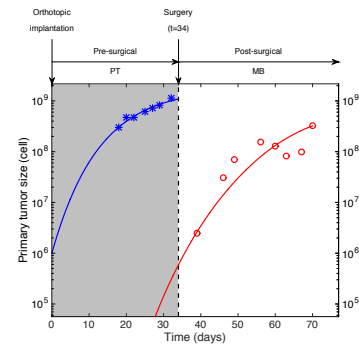
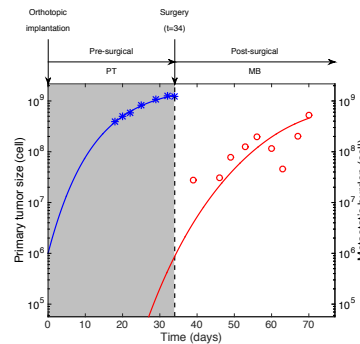
Validation on animal data



- * Data primary tumor
- Median model primary tumor
- - - 10th and 90th percentiles model primary tumor
- Data metastatic burden
- Median model metastatic burden
- - - 10th and 90th percentiles model metastatic burden

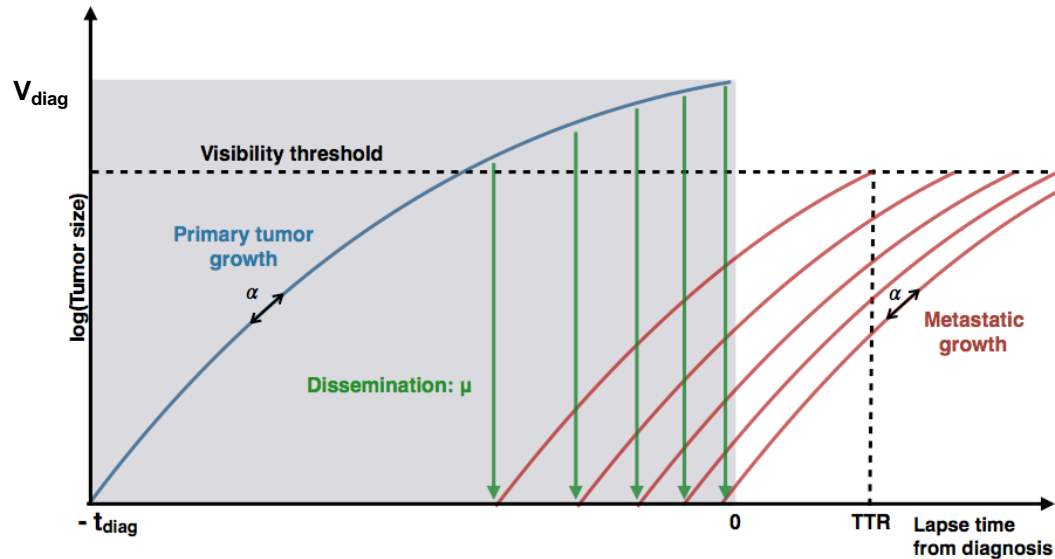
Nonlinear **mixed-effects**
statistical model for inter-
animal variability

$$\theta^i = \theta_{pop} + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \omega^2)$$



⇒ **same growth** for PT and mets: $\alpha_p = \alpha, \beta_p = \beta$

Mechanistic modeling of time to relapse



- Number of metastases with size larger than the **visible size** V_{vis} (= 0.5 cm)

$$N_{vis}(t) = \int_{V_{vis}}^{+\infty} \rho(t, v) dv = \int_0^{t-\tau_{vis}} d(V_p(t)) dt$$

τ_{vis} = time to reach V_{vis}

- **Time to relapse** (TTR) defined as the time elapsed from diagnosis to the appearance of a first visible metastasis

$$TTR = \inf \{t > 0 : N_{vis}(t_{diag} + t) \geq 1\}$$

- Parameter β fixed such that $V_{\infty} = e^{\frac{\alpha}{\beta}} = 10^{12}$ cells

Mixed-effects statistical model

$$\ln(T^i) = \ln(TTR(V_{diag}^i; \alpha^i, \mu^i)) + \varepsilon^i, \quad \varepsilon^i \sim \mathcal{N}(0, \sigma^2)$$

(Observation model)

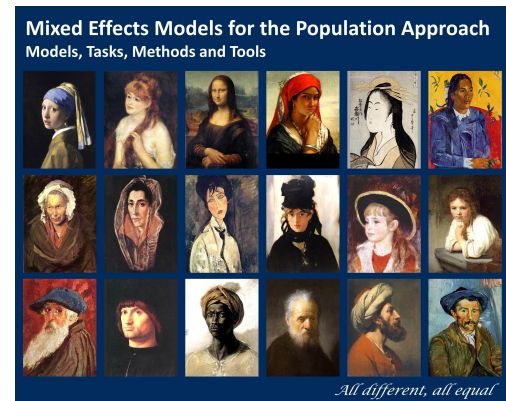
$$S(t|\alpha^i, \mu^i) = \mathbb{P}(T^i > t|\alpha^i, \mu^i)$$

Survival function to account for **censoring** in the likelihood

$$\begin{aligned} \ln(\alpha^i) &= \ln(\alpha_{pop}) + \eta_{\alpha}^i, & \eta_{\alpha}^i &\sim \mathcal{N}(0, \omega_{\alpha}^2) \\ \ln(\mu^i) &= \ln(\mu_{pop}) + \eta_{\mu}^i, & \eta_{\mu}^i &\sim \mathcal{N}(0, \omega_{\mu}^2) \end{aligned}$$

fixed effects

random effects

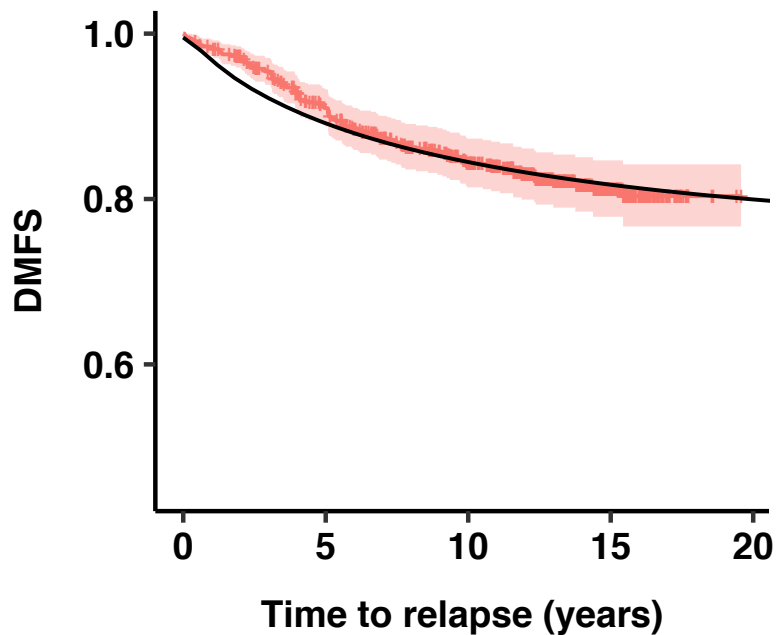


Lavielle, CRC press, 2014

Likelihood maximization performed using the *saemix* R package (SAEM algorithm)

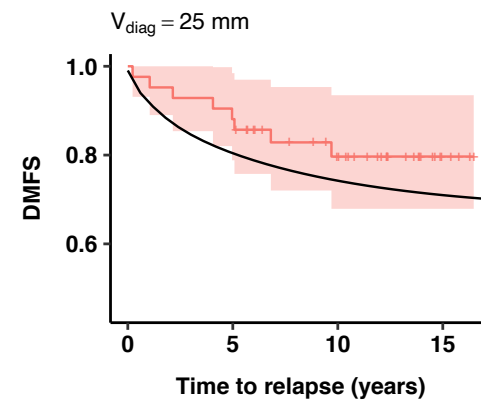
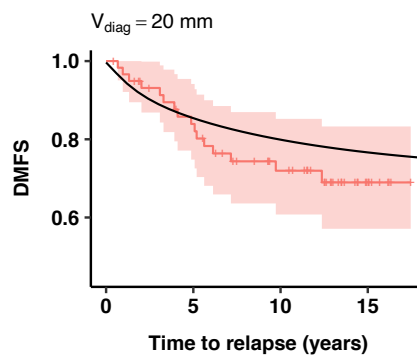
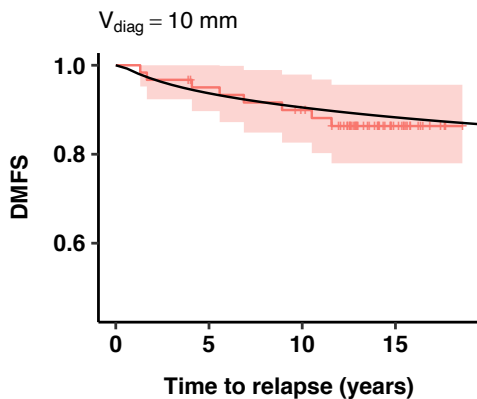
Comets, Lavenu, Lavielle, J Stat Softw, 2017

Descriptive power: fit to the data



—+— Kaplan–Meier estimate
— Model fit

Parameter	Estimate	r.s.e. (%)
$\log \alpha_{pop}$	-6.337	12.635
$\log \mu_{pop}$	-26.814	3.683
σ	0.542	28.409
ω_{α}	3.373	36.435
ω_{μ}	3.780	15.876



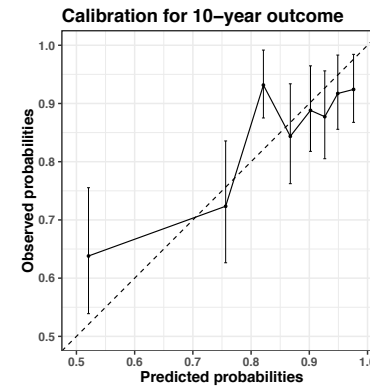
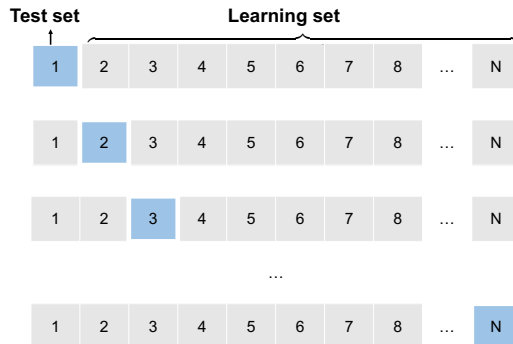
Predictive power: covariates

$$\ln(\mu^i) = \ln(\mu_{pop}) + \beta_{\mu}^T \mathbf{x}_{\mu}^i + \eta_{\mu}^i, \quad \eta_{\mu}^i \sim \mathcal{N}(0, \omega_{\mu}^2)$$

$$\ln(\alpha^i) = \ln(\alpha_{pop}) + \beta_{\alpha}^T \mathbf{x}_{\alpha}^i + \eta_{\alpha}^i, \quad \eta_{\alpha}^i \sim \mathcal{N}(0, \omega_{\alpha}^2)$$

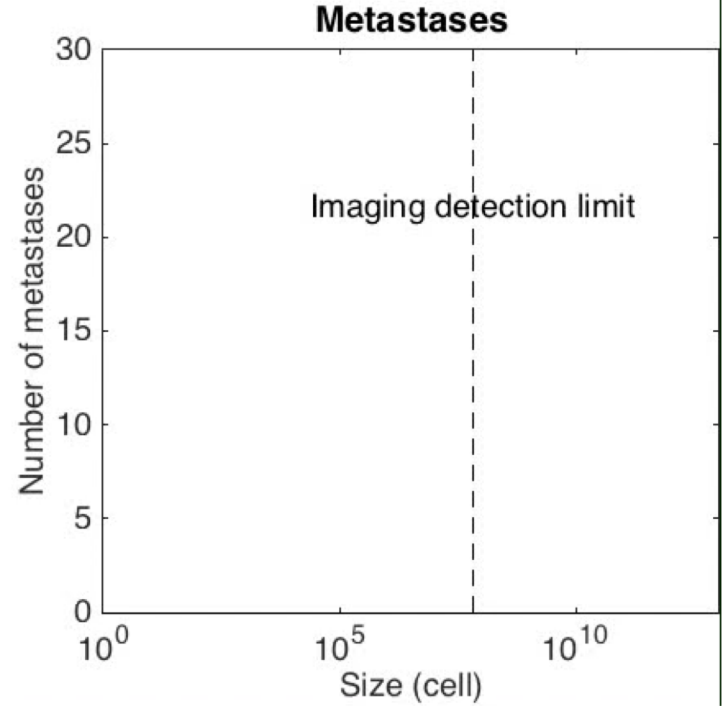
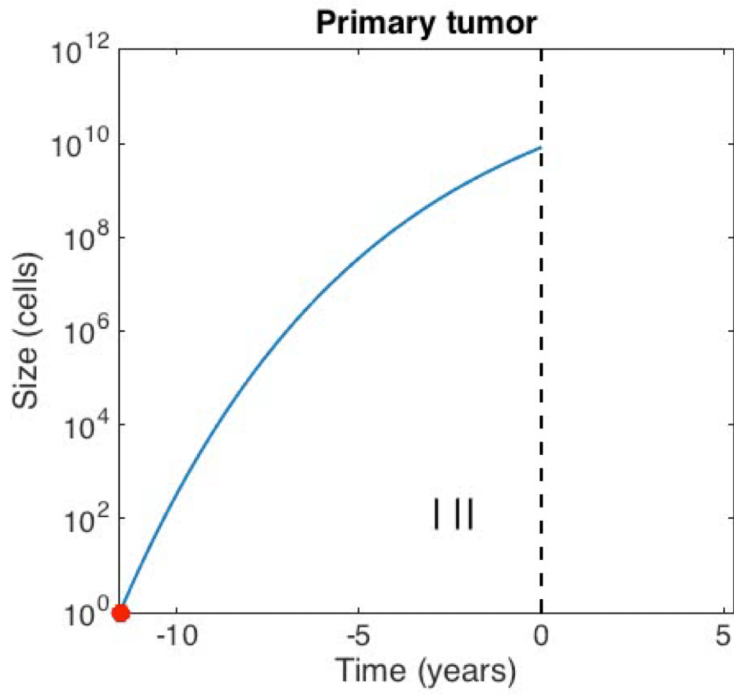
Parameter	Estimate	r.s.e. (%)	p-value
$\log \alpha_{pop}$	-8.883	10.151	
$\beta_{Ki67, \alpha}$	0.086	27.376	$2.59 \cdot 10^{-4}$
$\beta_{HER2, \alpha}$	0.029	42.833	0.020
$\beta_{CD44, \alpha}$	0.011	60.816	0.1
$\beta_{TRIO, \alpha}$	0.016	58.119	0.085
$\log \mu_{pop}$	-26.342	3.696	
$\beta_{EGFR, \mu}$	0.039	47.527	0.035
σ	0.606	23.104	
ω_{α}	2.062	22.715	
ω_{μ}	3.563	16.759	

c-index = 0.62 (10-folds cross-validation)

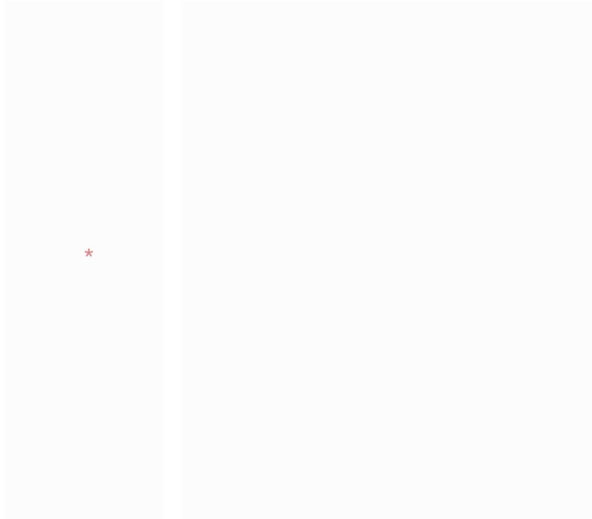


Patient ID	Tumor size (mm)	Ki67	HER2	CD44	TRIO	EGFR	Observed TTR (cens)	Predicted TTR	Prediction error (days)
47	20	32	100	0	0	50	739 (1)	447	292
255	25	1	60	90	60	0	1812 (1)	1609	203
143	18	60	0	50	0	0	2798 (1)	434	2364
12	10	20	0	23	0	0	5970 (0)	$+\infty$	-

t = -11.6 years

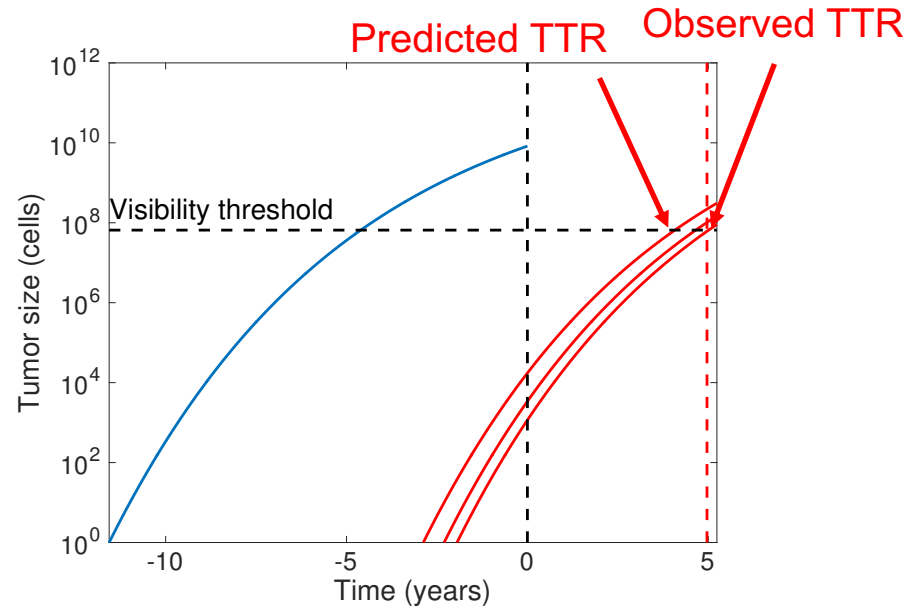


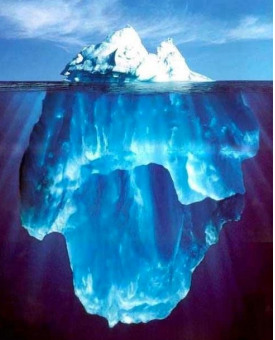
t = -141 months
— 10 mm



Primary tumor

Metastases



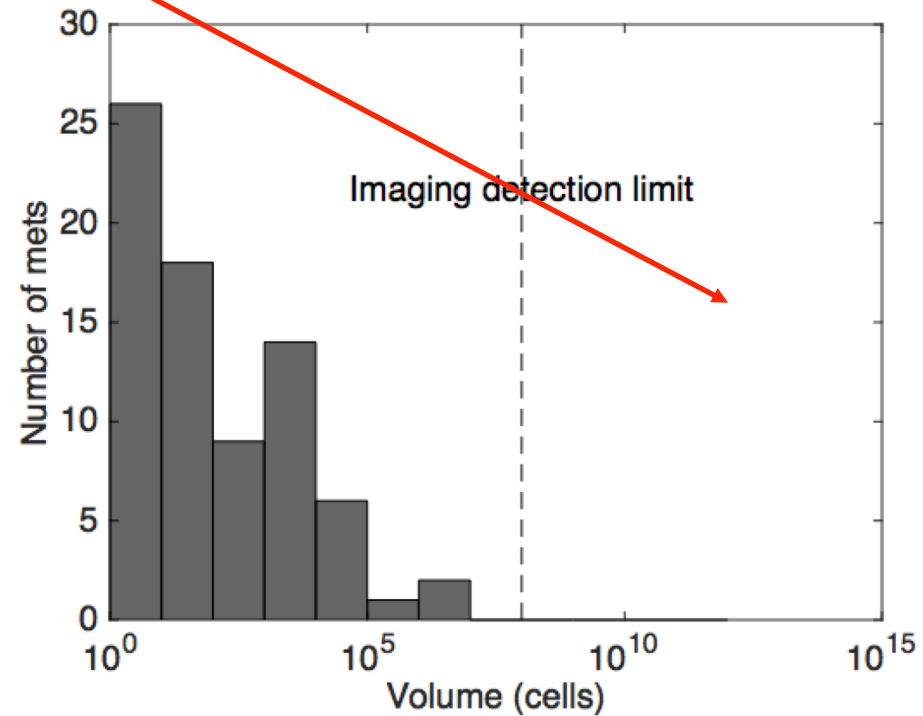
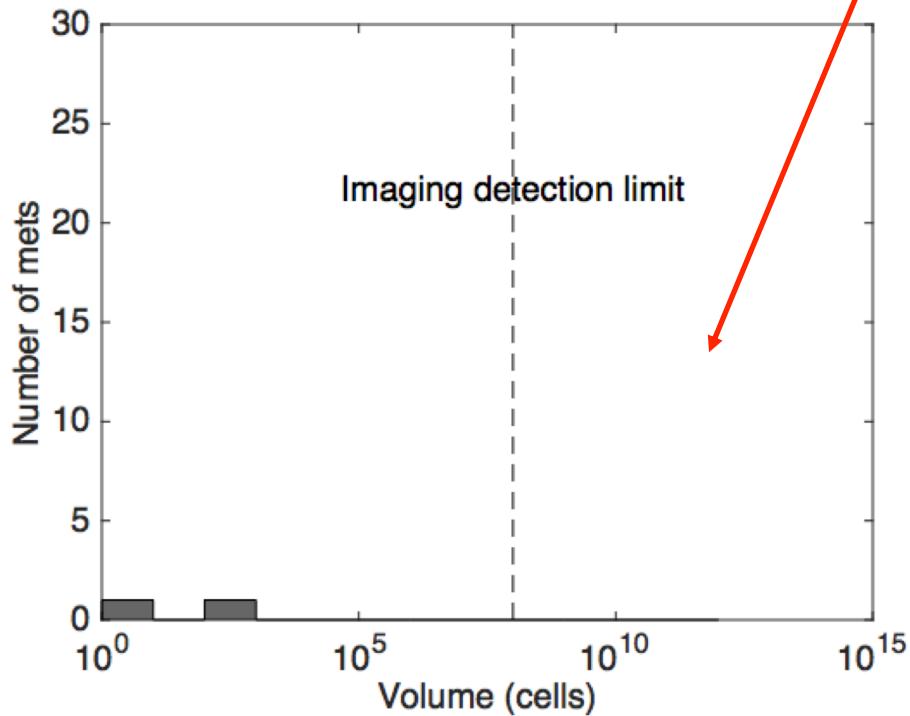


Model-based prediction of metastatic state at diagnosis

Median μ

Large μ (90th prct)

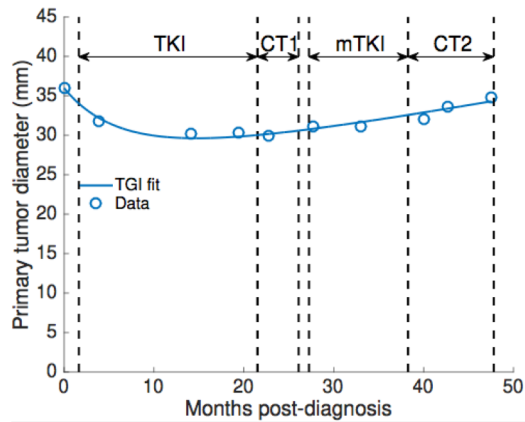
Nothing visible



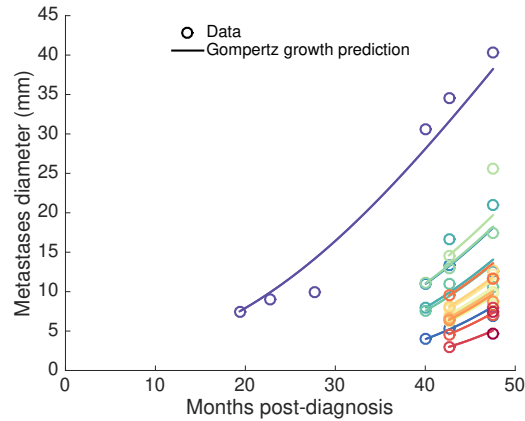
Breast cancer patient with primary tumor of 4.32 cm

Data of a NSCLC patient with brain mets

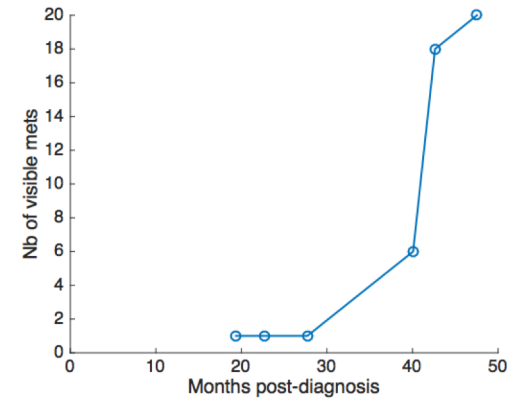
Primary tumor size



Metastases size



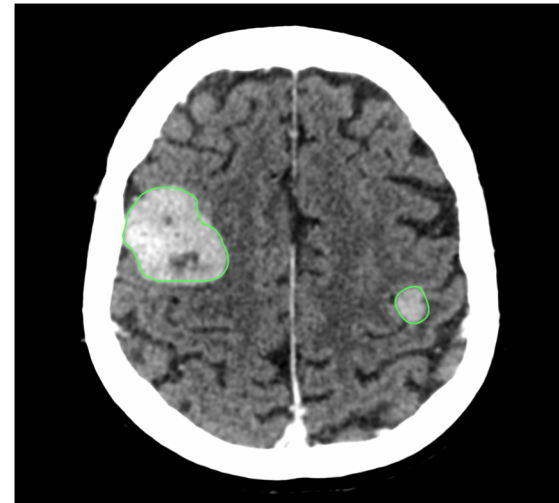
Number of visible mets



Lung CT



Brain CT scan



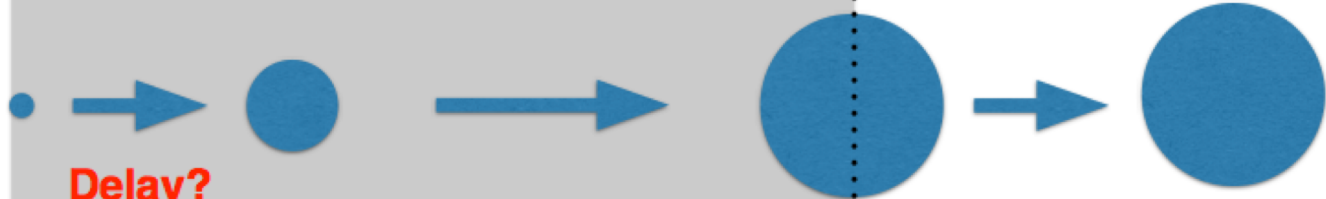
First cancer cell

Diagnosis

treatment

$$\text{Growth law: } g_p(V_p) = V_p(\alpha_p - \beta_p \ln(V_p))$$

Primary Tumor



Delay?

$$\text{Dissemination law: } d(V_p) = \mu(V_p)^\gamma$$

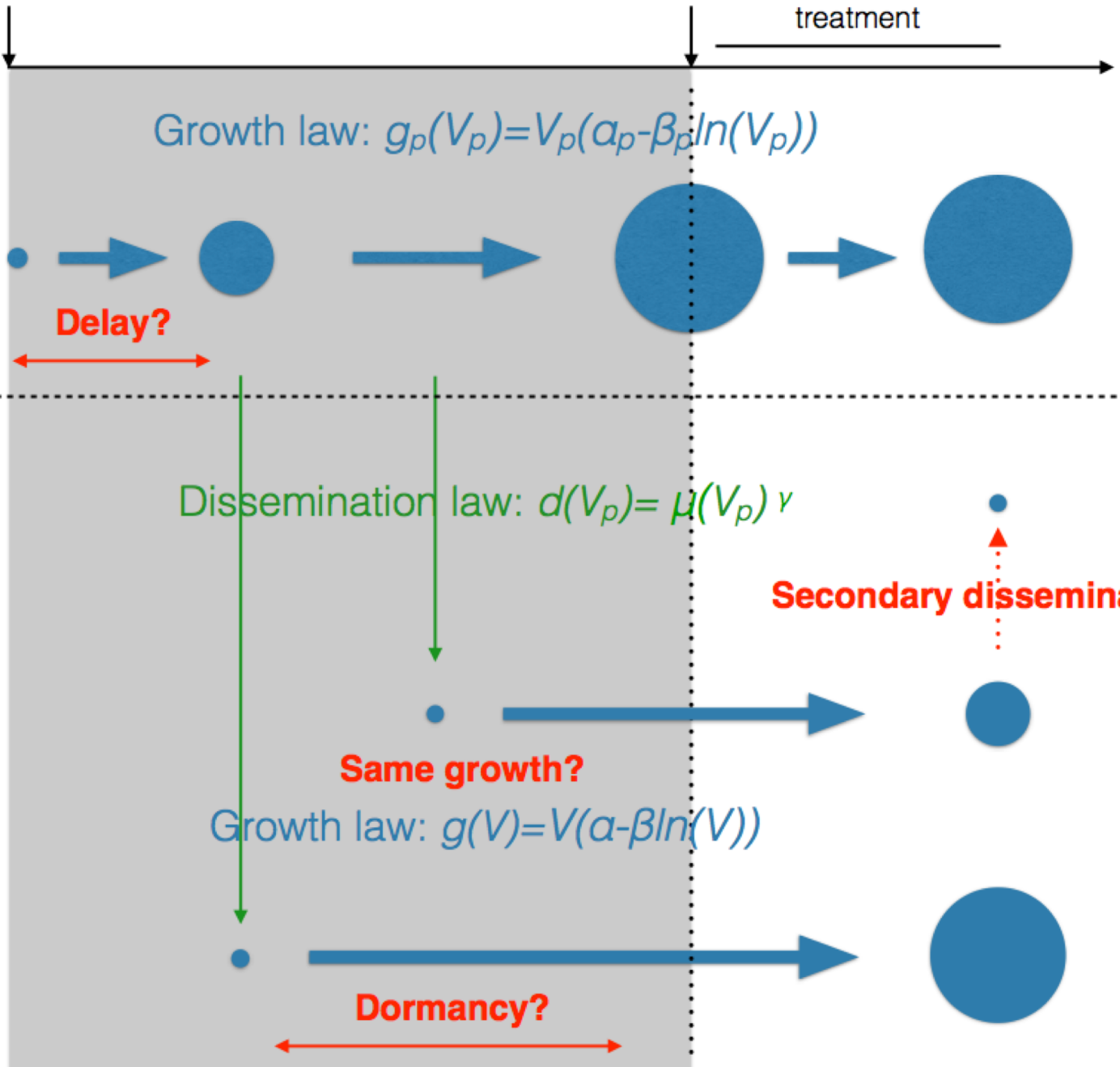
Brain Metastases

Secondary dissemination?

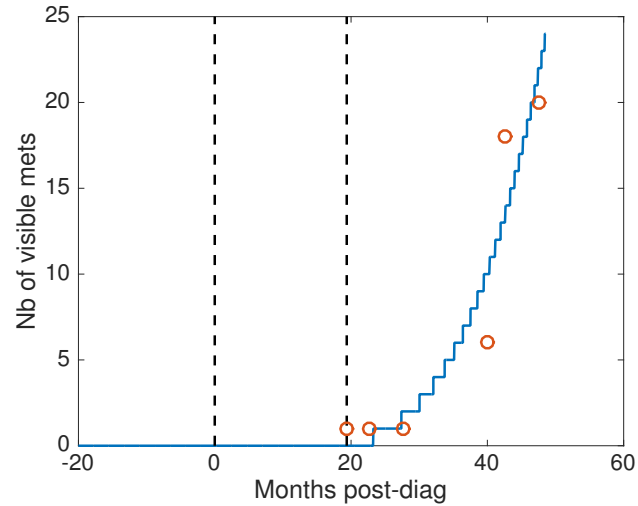
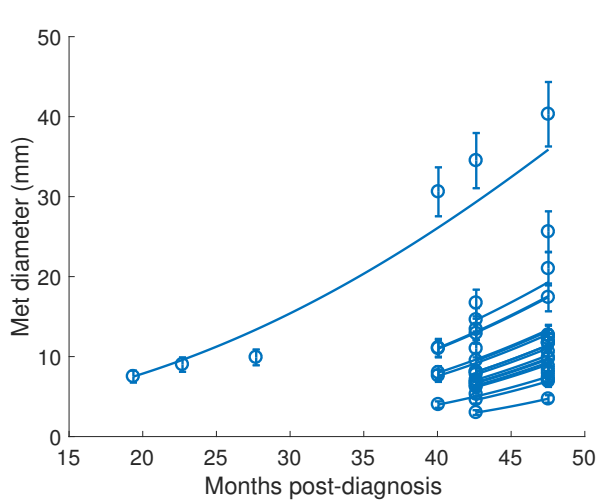
$$\text{Growth law: } g(V) = V(\alpha - \beta \ln(V))$$

Same growth?

Dormancy?



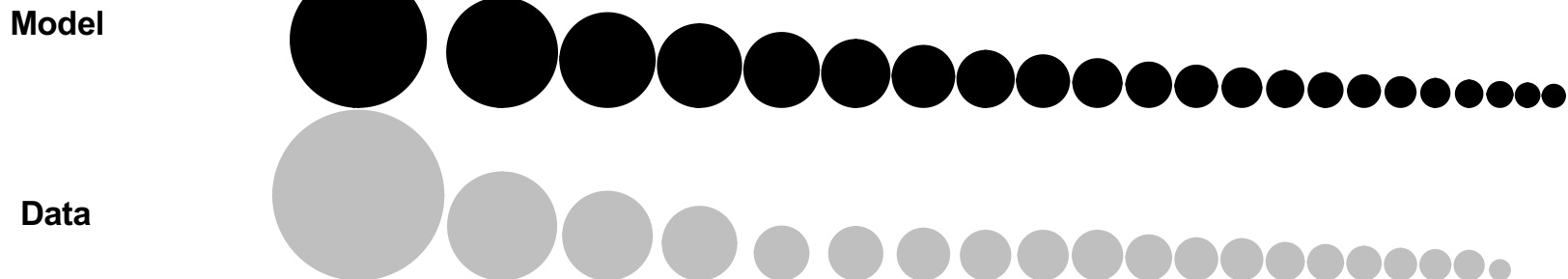
The model with dormancy could describe best the data



Objective function

Model	Patient 1	Patient 2
Base	5.51	2.53
Secondary	5.43	2.3
Delay	5.23	1.53
Dormancy	4.93	1.71
Diff. growth	4.95	1.79

Dormancy estimated to 133 days \pm 4.2%



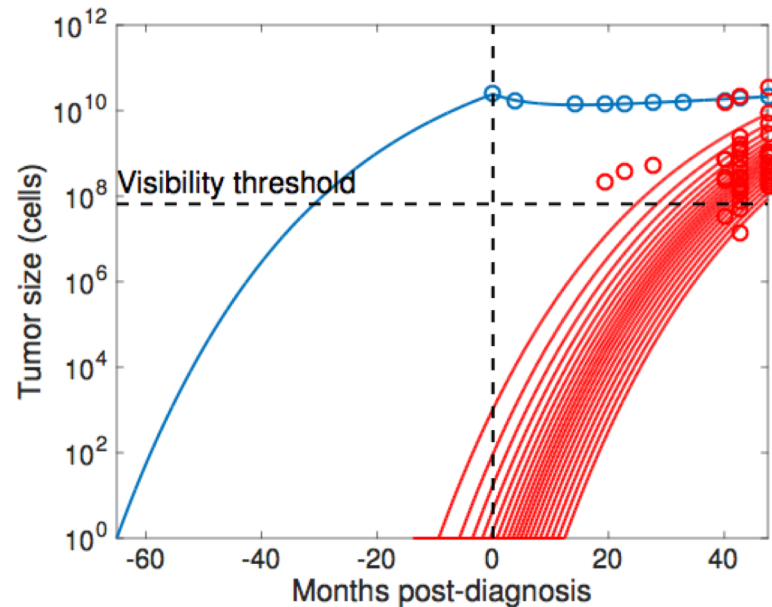
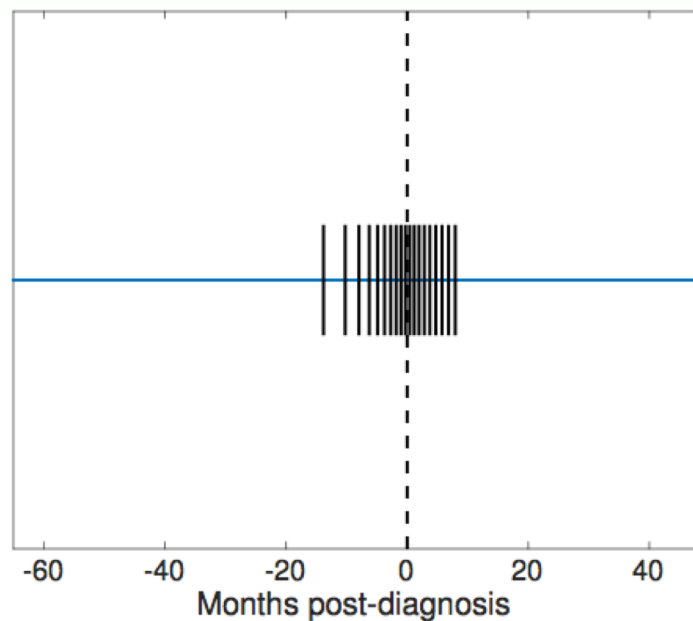
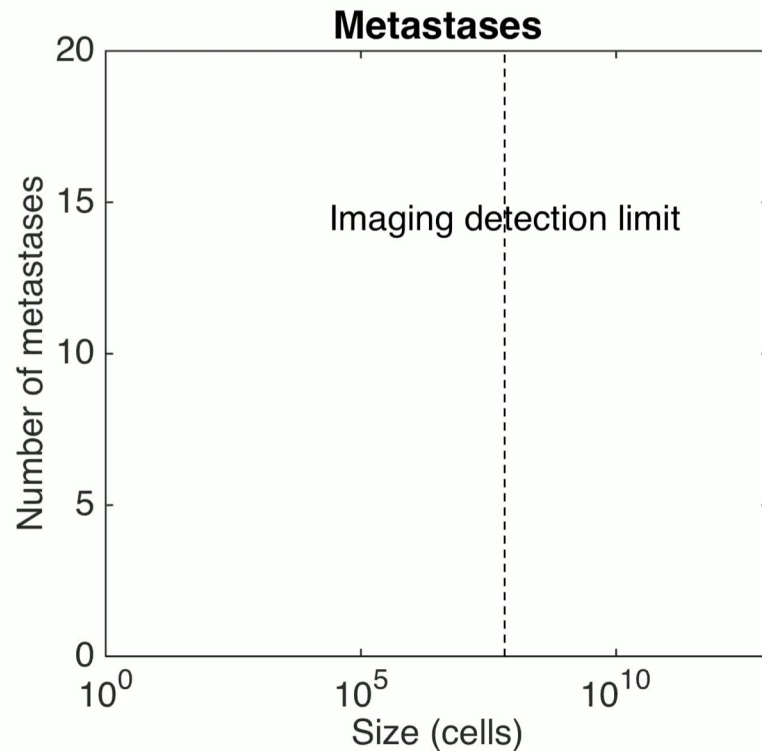
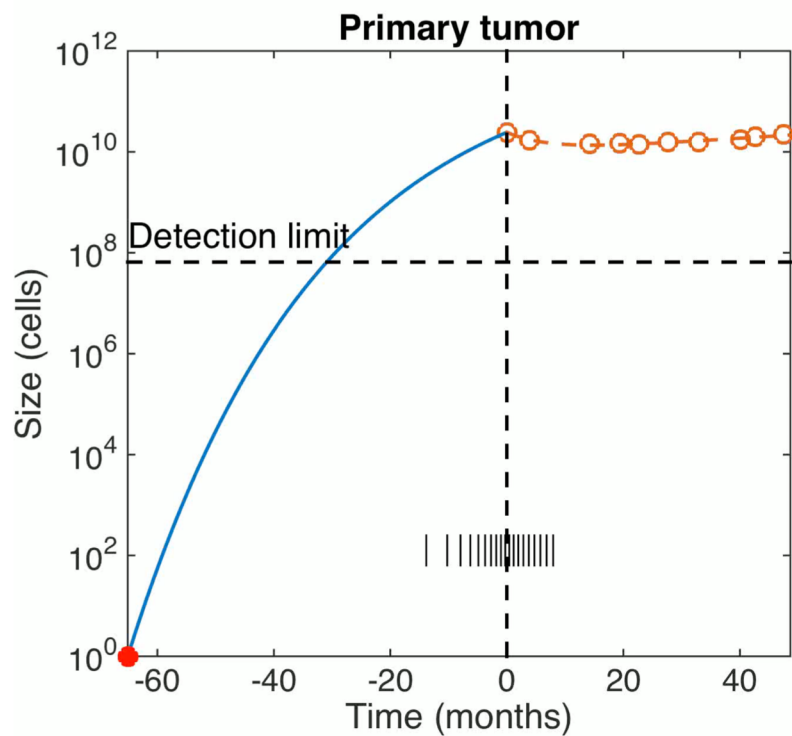
t = -55 months
— 10 mm

*

Primary
tumor
(lung)

Metastases (brain)

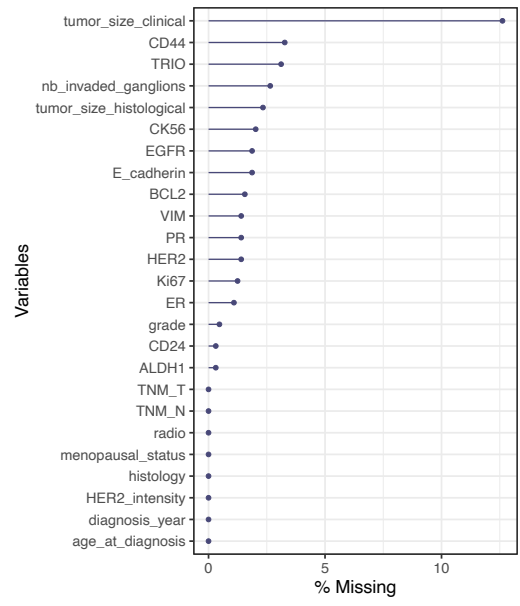
t = -65.1 months



Conclusions and perspectives

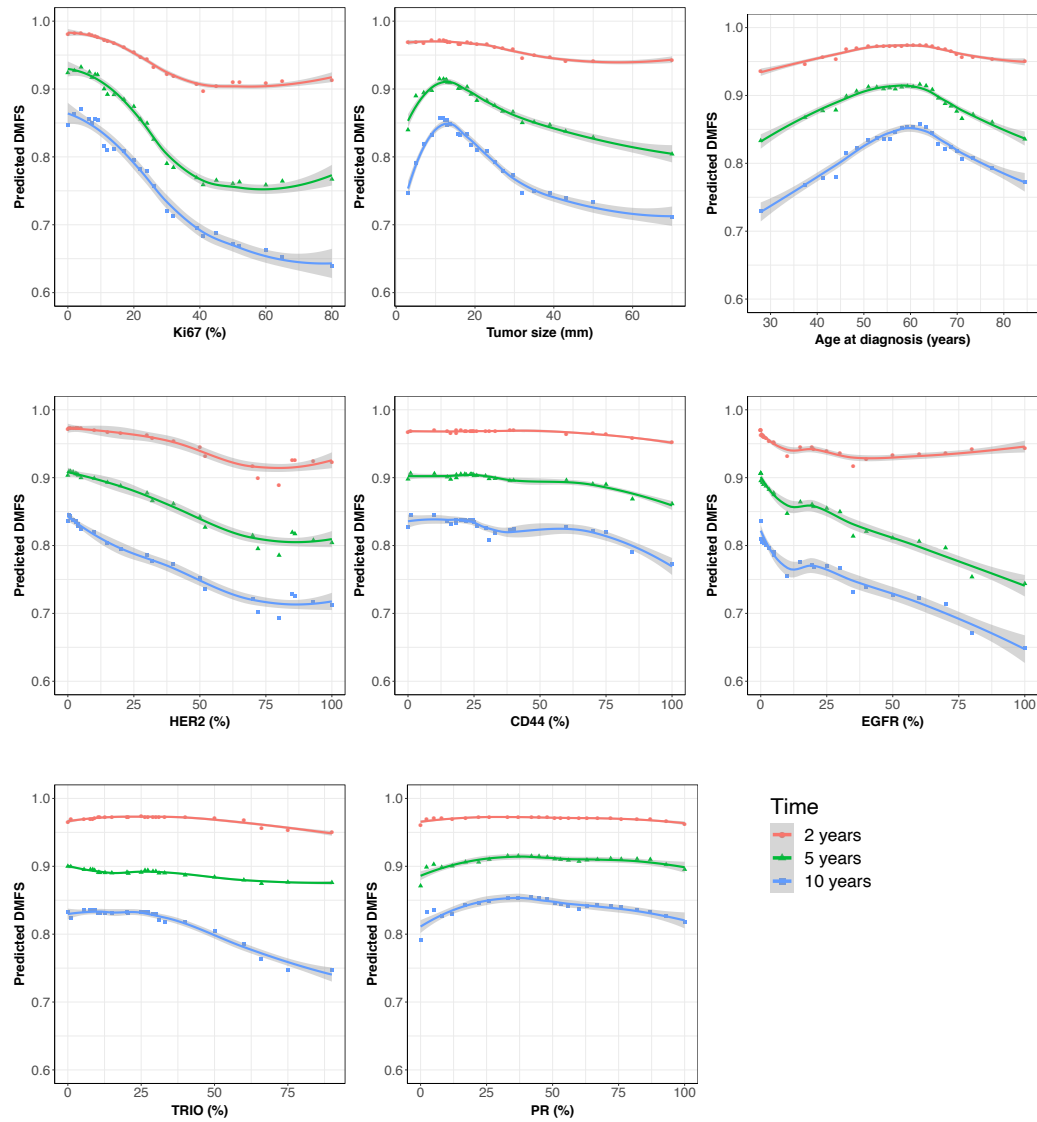
- Machine learning (random survival forest) performed **better than the mechanistic model for pure prediction** (c-index 0.69 vs 0.62), with similar performances as classical Cox regression (not shown)
- But mechanistic modeling provides **biological and clinical insights** that ML does not:
 - Ki67 correlates with proliferation rate α (expected but reassuring)
 - HER2 correlates with α , EGFR with μ (metastatic potential)
 - prediction of the **invisible metastatic state** at diagnosis \Rightarrow potential for **personalized adjuvant therapy**
- This is a first attempt of a **mechanistic, individual-level, predictive metastatic model**. A lot remains to be done!
 - Refinement to well-established breast cancer molecular subtypes
 - Further investigations to **refine the modeling** to improve the predictive power
 - Predictive power to be confirmed in **external data sets**

Supplementary slides



Percentage of missing values in each variable.

Missing covariate values were imputed using an iterative algorithm based on random forests (missForest R package).



Partial dependence plots of the random forest predicted DMFS as a function of the top eight predictors according to the minimal depth ranking.

Results for the Cox regression

	HR	95% CI	p-value
Ki67	1.02	[1.01, 1.03]	$1.71 \cdot 10^{-4}$
Tumor size	1.01	[0.99, 1.03]	0.46
Age	0.99	[0.98, 1.01]	0.49
HER2	1.01	[1.00, 1.01]	0.05
CD44	1.00	[1.00, 1.01]	0.47
EGFR	1.01	[1.00, 1.02]	0.06
TRIO	1.01	[1.00, 1.01]	0.12
PR	1.00	[0.99, 1.00]	0.80

Table S1: Cox regression using the first 8 covariates selected by minimal depth with the random survival forest model.

