Computational and Structural Biotechnology Journal 21 (2023) 5136–5143

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Survival analysis of patient groups defined by unsupervised machine learning clustering methods based on patient metabolomic data.

Caroline Bailleux [a,b,*], David Chardin [c,b], Jean-Marie Guigonis [b], Jean-Marc Ferrero [a], Yann Chateau [d], Olivier Humbert [c,b], Thierry Pourcher [b], Jocelyn Gal [d]

[a] University Côte d'Azur, Centre Antoine Lacassagne, Medical Oncology Department, Nice F-06189, France
[b] University Côte d'Azur, Commissariat à l'Energie Atomique et aux énergies alternatives, Institut Frédéric Joliot, Service Hospitalier Frédéric Joliot, laboratory Transporters in Oncology and Radiotherapy in Oncology (TIRO), School of medicine, Nice F-06100, France
[c] University Côte d'Azur, Centre Antoine Lacassagne, Nuclear medicine Department, Nice F-06189, France
[d] University Côte d'Azur, Centre Antoine Lacassagne, Epidemiology and Biostatistics Department, Nice F-06189, France

ARTICLE INFO

Keywords:
Unsupervised machine learning
Clustering
Breast cancer
Survival
Untargeted
Proof-of-concept

ABSTRACT

Purpose: Meta-analyses failed to accurately identify patients with non-metastatic breast cancer who are likely to benefit from chemotherapy, and metabolomics could provide new answers. In our previous published work, patients were clustered using five different unsupervised machine learning (ML) methods resulting in the identification of three clusters with distinct clinical and simulated survival data. The objective of this study was to evaluate the survival outcomes, with extended follow-up, using the same 5 different methods of unsupervised machine learning.

Experimental design: Forty-nine patients, diagnosed between 2013 and 2016, with non-metastatic BC were included retrospectively. Median follow-up was extended to 85.8 months. 449 metabolites were extracted from tumor resection samples by combined Liquid chromatography-mass spectrometry (LC–MS). Survival analyses were reported grouping together Cluster 1 and 2 versus cluster 3. Bootstrap optimization was applied.

Results: PCA k-means, K-sparse and Spectral clustering were the most effective methods to predict 2-year progression-free survival with bootstrap optimization (PFSb); as bootstrap example, with PCA k-means method, PFSb were 94% for cluster 1&2 versus 82% for cluster 3 ($p = 0.01$). PCA k-means method performed best, with higher reproducibility (mean HR=2 (95%CI [1.4–2.7]); probability of $p \leq 0.05$ 85%). Cancer-specific survival (CSS) and overall survival (OS) analyses highlighted a discrepancy between the 5 ML unsupervised methods.

Conclusion: Our study is a proof-of-principle that it is possible to use unsupervised ML methods on metabolomic data to predict PFS survival outcomes, with the best performance for PCA k-means. A larger population study is needed to draw conclusions from CSS and OS analyses.

1. Introduction

Worldwide, breast cancer(BC) is the most common cancer in women and the second leading cause of cancer deaths [1]. Breast cancer impacts one in every eight women, contributing to approximately 2.3 million new cases globally each year. According to the 2018 estimates by the International Agency for Research on Cancer, the annual age-adjusted incidence rate per 100,000 women nears 100. This disease can occur at any age, but the average age of diagnosis is 62 years. Half of breast cancer diagnoses occur between the ages of 50 and 69, while 20% are identified before the age of 50%, and 10% before the age of 40 [2–4]. Currently, approximately 60% of breast cancer cases are discovered at the localized stage (stage I), 30% at a locally advanced stage (stage II-III), and 10% at a metastatic stage (stage IV) [3].

For non-metastatic high-risk breast cancer, chemotherapy is proposed to reduce the risk of relapse. A significant portion of breast cancer patients may not derive substantial benefits from adjuvant or neo-adjuvant chemotherapy. Nevertheless, chemotherapy comes with short and long-term risks, including immediate side effects such as nausea, vomiting, alopecia, myelosuppression, early cognitive impairments,

* Correspondence to: Medical Oncology Department, Centre Antoine Lacassagne, University Côte d'Azur, 33 avenue de Valombrose, 06189 Nice, France.
E-mail address: caroline.bailleux@nice.unicancer.fr (C. Bailleux).

https://doi.org/10.1016/j.csbj.2023.10.033
Received 30 April 2023; Received in revised form 16 October 2023; Accepted 16 October 2023
Available online 19 October 2023

fertility loss, infectious risk, and neuropathy. In some instances, these neuropathies persist over the long term, leading to lasting consequences [5]. Long-term toxicities also comprise potential cardiotoxicity associated with anthracyclines and the rare but noteworthy risk of secondary leukemia linked to chemotherapy [6]. The decision to include CT in the treatment regimen is based on clinicopathological criteria associated with BC prognosis. Meta-analyses failed to accurately identify the characteristics of patients who are likely to benefit from adjuvant chemotherapy. However, the decision to forgo chemotherapy is a challenging one. Therefore, following surgical intervention, genomic tests are recommended for individuals with intermediate clinical risk, hormone-dependent breast cancer. No genomic test exists for triple negative of HER2-enriched breast cancer.

Metabolic pathway alterations associated with BC tumors and disease progression have been widely explored at the genomic level [7–9]. Proteomics studies have also revealed alterations in metabolism-associated protein expression in BC tumors with a correlation with overall and recurrence-free survival[10]. Metabolomics is a new, rapidly developing field of investigation dedicated to the study of metabolism in tissues and fluids. There are two distinct approaches to metabolomics: a targeted approach aiming to precisely quantify a limited number of predefined metabolites of interest [11] and a non-targeted approach aiming to objectively measure the largest possible number of metabolites in a sample [12,13]. Metabolomics can generate a large amount of data, which can make their analysis difficult, hence the usefulness of machine learning (ML) methods to extract useful information. In the case of metabolomics, ML involve supervised or unsupervised methods. Supervised methods can be used to predict metabolites or biomarkers associated with a particular disease from labeled metabolomic data. Unsupervised learning can be used to identify patterns or groups of patients and metabolites that may be associated with specific diseases or phenotypes from unlabeled metabolomic data. The unsupervised algorithm takes a dataset and attempts to find a structure in the data by grouping or clustering the data points [14,15].

We previously published [16] a comparison of 5 different unsupervised machine learning methods Principal Component Analysis k-means (PCA K-means), Sparse k-means, Spectral clustering, Single-cell Interpretation via Multi-kernel LeaRning (SIMLR), and k-sparse, to establish a metabolomic signature of breast cancer (BC). In-silico survival analysis based on survival data simulated by predict tool (https://breast.predict.nhs.uk/tool)[17,18] revealed a significant difference for 5-year predicted overall survival (OS) and cancer-specific survival (CSS) between the 3 clusters [16]. However, these simulated data may also be biased. Only few studies have reported associations between metabolic alterations and early BC patient survival outcomes based on serum analyses [19,20]. At present, only one article has looked at the analysis metabolomics data from breast cancer tumors using unsupervised machine learning [21]. Alakwaa *and al.* have identified signatures associated with metabolomics using unsupervised methods. They proposed a bioinformatics pipeline that analysed metabolomic data from breast cancer tumours, highlighting subgroups. The major limitation of this work was that the authors are not interested in the clinical relevance in terms of patient characteristics or survival. As the event occurs many years after the initial diagnosis, unlike certain other types of cancer [22], the data were not initially mature enough to allow survival analyses to be carried out with real data.

Therefore, with extended follow-up of an additional three years, we analyzed the real survival data to provide the first real survival data derived from clusters identified through machine learning. The objective of this study was to evaluate the results of 5 different methods of unsupervised machine learning (PCA k-means, Sparse k-means, Spectral clustering, SimLR and K-sparse) to predict progression free survival (PFS), CSS and OS.

2. Material and methods

2.1. Selection and data collection of patients

A cohort of patients treated in our institution between March 2013 and September 2016 for a clinical stage I to III_B biopsy-proven BC, with an indication for adjuvant therapy after surgery, was included retrospectively in the study. Compared to the first publication[16], 3 metastatic patients were excluded from the survival analysis. A patient was considered de novo metastatic if metastatic diagnosis occurred within the first two months of treatment. Patients were treated according to national guidelines. Clinical, histological, radiological, and therapeutic data were retrospectively extracted from our facility's digital records or collected by a clinical data monitor. Follow up data was either extracted from our facility's digital records or retrieved by telephone if patients had changed facilities during surveillance. The date of the latest news was updated at the time of the final survival analysis, on December 2022. Written informed consent was obtained from all study participants. All procedures performed in studies involving tissue collection and analyses were in accordance with the ethical standards of the institutional and/or national research committee (French National Commission for Informatics and Liberties N°17003 and National Institute Health data N° 1515251018).

2.2. Metabolomic analysis and clustering algorithms

Sample collection and preparation, details of LC-MS analysis, data preprocessing using MZmine have been already reported previously [16](in Supplementary Material S1 and Supplementary Fig. 1). Metabolites from both positive and negative ionization modes were merged. Only metabolites without any missing values after pre-processing were chosen for analysis. In cases where a metabolite was detected in both positive and negative modes, only the mode with the highest average intensity was taken into account. Following these procedures, 1271 metabolites were identified. Prior to statistical analysis, a filtering function was applied to remove noisy data. Finally, statistical analysis was performed on 449 metabolites.

Five unsupervised clustering methods were then analyzed: Principal Component Analysis (PCA) k-means, Sparse k-means, Single-cell Interpretation via Multi-kernel LeaRning (SIMLR), k-sparse and Spectral clustering. In order to apply these five unsupervised clustering methods, the optimal number of clusters, $k = 3$, was determined [16].

2.3. Statistical analysis

Relevance of the discovered clusters was assessed by comparing the clinical and survival characteristics between clusters using χ^2 or Fisher's exact tests for categorical data, analysis of variance or Mann-Whitney's test for continuous variables and log-rank test for censored data. *P*-values below 0.05 (two-sided) were considered statistically significant. Overall survival (OS) was defined as the time between diagnosis and death due to any cause. Cancer-specific survival (CSS) was defined as the time between diagnosis and death due to breast cancer. Progression-Free Survival (PFS) was defined as the time between diagnosis and the first progression (local, regional and metastasis). Patients showing no event (death or recurrence) or lost to follow-up were censored at the date of their last contact. OS, CSS and PFS were estimated using the Kaplan-Meier method. Median follow-up with a 95% confidence interval (95%CI) was calculated by reverse Kaplan-Meier method. Median follow-up and survival curves were compared using the log-rank test. Cox proportional hazards models were used to estimate hazard ratios (HR) and 95% CIs for the relation between treatment and survival. 2-year outcomes were detailed, and bootstrap optimization was applied to these results to simulate the effect in a larger study population and highlight first trends: 200 patients were randomly sampled to assess effectiveness without overpowering; 1000 patients were randomly

Table 1
Clinical comparison of 49 patients between clusters.

			K − sparse				SpectralClustering				PCA − K − means				SIMLR				SparseK − means				
Characteristics		All (n = 49)	Cluster 1&2 (n = 30)	Cluster 3 (n = 19)	p-value		Cluster 1&2 (n = 30)	Cluster 3 (n = 19)	p-value		Cluster 1&2 (n = 30)	Cluster 3 (n = 19)	p-value		Cluster 1&2 (n = 28)	Cluster 3 (n = 21)	p-value		Cluster 1&2 (n = 31)	Cluster 3 (n = 18)	p-value	
		Nb of patients (%)	Nb of patients (%)	Nb of patients (%)			Nb of patients (%)	Nb of patients (%)			Nb of patients (%)	Nb of patients (%)			Nb of patients (%)	Nb of patients (%)			Nb of patients (%)	Nb of patients (%)		
Age (median min-max)*		65 (37–88)	65 (38–88)	63 (37–84)	0.552		65 (38–88)	63 (37–84)	0.552		65 (38–88)	65 (37–84)	0.547		65 (38–88)	63 (37–84)	0.424		65 (38–88)	65.5 (37–82)	0.881	t-test
Histology type£					0.144				0.144				0.727				0.756				0.378	
	Invasive ductal carcinoma	45 (91.8)	27 (90.0)	18 (94.7)			27 (90.0)	18 (94.7)			28 (93.3)	17 (89.5)			26 (92.9)	19 (90.5)			27 (87.1)	18 (100.0)		Fisher
	Invasive lobular carcinoma	3 (6.1)	3 (10.0)	0 (0.0)			3 (10.0)	0 (0.0)			2 (6.7)	1 (5.3)			2 (7.1)	1 (4.8)			3 (9.7)	0 (0.0)		
	Other	1 (2.0)	0 (0.0)	1 (5.3)			0 (0.0)	1 (5.3)			0 (0.0)	1 (5.3)			0 (0.0)	1 (4.8)			1 (3.2)	0 (0.0)		
Tumor stage£					0.073				0.073				0.074				0.080				0.237	
	T1	20 (40.8)	16 (53.3)	4 (21.1)			16 (53.3)	4 (21.1)			16 (53.3)	4 (21.1)			15 (53.6)	5 (23.8)			15 (48.4)	5 (27.8)		Fisher
	T2	22 (44.9)	10 (33.3)	12 (63.2)			10 (33.3)	12 (63.2)			10 (33.3)	12 (63.2)			9 (32.1)	13 (61.9)			11 (35.5)	11 (61.1)		
	T3	7 (14.3)	4 (13.3)	3 (15.8)			4 (13.3)	3 (15.8)			4 (13.3)	3 (15.8)			4 (14.3)	3 (14.3)			5 (16.1)	2 (11.1)		
Axillary lymph node status#					0.090				0.090				0.090				0.243				0.441	
	N0	28 (57.1)	20 (66.7)	8 (42.1)			20 (66.7)	8 (42.1)			20 (66.7)	8 (42.1)			18 (64.3)	10 (47.6)			19 (61.3)	9 (50.0)		Chi
	N +	21 (42.9)	10 (33.3)	11 (57.9)			10 (33.3)	11 (57.9)			10 (33.3)	11 (57.9)			10 (35.7)	11 (52.4)			12 (38.7)	9 (50.0)		
Histological grade£					**0.023**				**0.023**				**0.049**				**0.018**				**0.049**	
	I	5 (10.2)	5 (16.7)	0 (0.0)			5 (16.7)	0 (0.0)			5 (16.7)	0 (0.0)			5 (17.9)	0 (0.0)			5 (16.1)	0 0		Fisher
	II	20 (40.8)	15 (50.0)	5 (26.3)			15 (50.0)	5 (26.3)			14 (46.7)	6 (31.6)			14 (50.0)	6 (28.6)			14 (45.2)	6 (33.3)		
	III	23 (46.9)	10 (33.3)	13 (68.4)			10 (33.3)	13 (68.4)			11 (36.7)	12 (63.2)			9 (32.1)	14 (66.7)			11 (35.5)	12 (66.7)		
Hormonal status#					**0.016**				**0.016**				0.070				**0.017**				**0.035**	
	Negatif	23 (46.9)	10 (33.3)	13 (68.4)			10 (33.3)	13 (68.4)			11 (36.7)	12 (63.2)			9 (32.1)	14 (66.7)			11 (35.5)	12 (66.7)		Chi
	Positif	26 (53.1)	20 (66.7)	6 (31.6)			20 (66.7)	6 (31.6)			19 (63.3)	7 (36.8)			19 (67.9)	7 (33.3)			20 (64.5)	6 (33.3)		
Her-2 status£					0.282				0.282				0.282				0.470				0.708	
	Non-over-expressed	40 (81.6)	26 (86.7)	14 (73.7)			26 (86.7)	14 (73.7)			26 (86.7)	14 (73.7)			24 (85.7)	16 (76.2)			26 (83.9)	14 (77.8)		Fisher
	Over-expressed	9 (18.4)	4 (13.3)	5 (26.3)			4 (13.3)	5 (26.3)			4 (13.3)	5 (26.3)			4 (14.3)	5 (23.8)			5 (16.1)	4 (22.2)		
Triple-negatif status#					**0.043**				**0.043**				0.165				**0.025**				**0.025**	
	No	34 (69.4)	24 (80.0)	10 (52.6)			24 (80.0)	10 (52.6)			23 (76.7)	11 (57.9)			23 (82.1)	11 (52.4)			25 (80.6)	9 (50.0)		Chi
	Yes	15 (30.6)	6 (20.0)	9 (47.4)			6 (20.0)	9 (47.4)			7 (23.3)	8 (42.1)			5 (17.9)	10 (47.6)			6 (19.4)	9 (50.0)		
Luminal#					**0.006**				**0.006**				**0.030**				**0.006**				**0.013**	
	No	24 (49.0)	10 (33.3)	14 (73.7)			10 (33.3)	14 (73.7)			11 (36.7)	13 (68.4)			9 (32.1)	15 (71.4)			11 (35.5)	13 (72.2)		Chi
	Yes	25 (51.0)	20 (66.7)	5 (26.3)			20 (66.7)	5 (26.3)			19 (63.3)	6 (31.6)			19 (67.9)	6 (28.6)			20 (64.5)	5 (27.8)		
Adjuvant chemotherapy£					0.323				0.323				0.323				0.192				1	
	No	12 (24.5)	9 (30.0)	3 (15.8)			9 (30.0)	3 (15.8)			9 (30.0)	3 (15.8)			9 (32.1)	3 (14.3)			8 (25.8)	4 (22.2)		Fisher
	Yes	37 (75.5)	21 (70.0)	16 (84.2)			21 (70.0)	16 (84.2)			21 (70.0)	16 (84.2)			19 (67.9)	18 (85.7)			23 (74.2)	14 (77.8)		
Adjuvant radiotherapy£					1				1				1				1				1	
	No	5 (10.2)	3 (10.0)	2 (10.5)			3 (10.0)	2 (10.5)			3 (10.0)	2 (10.5)			3 (10.7)	2 (9.5)			3 (9.7)	2 (11.1)		Fisher
	Yes	44 (89.8)	27 (90.0)	17 (89.5)			27 (90.0)	17 (89.5)			27 (90.0)	17 (89.5)			25 (89.3)	19 (90.5)			28 (90.3)	16 (88.9)		
Adjuvant hormone therapy#					0.181				0.181				0.181				0.154				0.319	
	No	20 (40.8)	10 (33.3)	10 (52.6)			10 (33.3)	10 (52.6)			10 (33.3)	10 (52.6)			9 (32.1)	11 (52.4)			11 (35.5)	9 (50.0)		Chi
	Yes	29 (59.2)	20 (66.7)	9 (47.4)			20 (66.7)	9 (47.4)			20 (66.7)	9 (47.4)			19 (67.9)	10 (47.6)			20 (64.5)	9 (50.0)		

£ : Fisher's exact test; #: Chi2-test; * : student t-test

Table 2
Survival outcomes with 5 different methods of unsupervised machine learning.

ML Method		Previous clustering (k = 3)	New clustering (k = 2)	2-year survival outcome (k = 2)
k-sparse	OS	p = 0.5	p = 0.3	p = 0.3
	CSS	p = 0.8	p = 0.5	p = 0.3
	PFS	p = 0.7	p = 0.4	p = 0.5
PCA k-means	OS	p = 0.5	p = 0.3	p = 0.5
	CSS	p = 0.7	p = 0.5	p = 0.5
	PFS	p = 0.5	p = 0.4	p = 0.5
Spectral clustering	OS	p = 0.5	p = 0.3	p = 0.3
	CSS	p = 0.8	p = 0.5	p = 0.4
	PFS	p = 0.7	p = 0.4	p = 0.5
Sparse k-means	OS	p = 0.8	p = 0.7	p = 1
	CSS	p = 0.9	p = 0.7	p = 1
	PFS	p = 0.9	p = 0.6	p = 0.5
SimLR	OS	p = 0.3	p = 0.2	p = 0.3
	CSS	p = 0.7	p = 0.4	p = 0.3
	PFS	p = 0.8	p = 0.5	p = 0.7

ML: machine learning; k: number of clusters; OS: overall survival; CSS: cancer-specific survival; PFS: progression-free survival; k: number of clusters.

sampled to provide reproducibility and performance criteria (*P*-values, 95%CI) in a sample size comparable to similar studies dealing with genomic signature. For each sampling, the survival was compared between each cluster of all 5 methods. The relationship between clusters and the OS, CSS and PFS was analyzed by hazard ratio (95% confidence interval). A total of 1500 replicates were performed. All statistical analyses were performed with R 4.3.1 (R Foundation).

3. Results

3.1. Clinical and tumor characteristics

Forty-nine consecutive patients with non-metastatic breast cancer were analyzed. Tumor and treatment characteristics are described in Table 1. Median age was 65 years (range: 37–8). Main histological type and tumor stage were invasive ductal carcinoma (91.8%), T1 (40.8%) and T2 (44.9%) respectively. Twenty-one patients (42.9%) presented axillary lymph node invasion. Five patients (10.2%) had histological grade I tumors, 20 patients (40.8%) had histological grade II tumors and 23 patients (46.9%) had grade III tumors. Half of the patients' tumors had negative hormone receptor status (46.9%) and 18.4% had a Her-2 overexpression. To study the survival behavior of the supposedly aggressive cancers grouped in cluster 3 and to deal with small population size in each cluster, cluster 1 and 2 (cluster 1&2) were grouped together to be compared to cluster 3. As previously described, patients in cluster 3 were more often those with unfavorable prognostic factors: grade III, non-luminal with negative hormone receptor or triple negative phenotype. On the contrary, patients in cluster 1&2 more often had favorable prognosis factors: tumour stage T1, N0, histological grade I/II, and luminal phenotype. Details of patient characteristics for the five unsupervised machine learning methods cluster 1&2 and cluster 3 are shown in Table 1.

3.2. Survival outcomes for the entire cohort

Median follow–up was extended to 85.8 months (95%CI, [83.6–97.9]). In the entire cohort, 2-year PFS and 5 year PFS were 98% (95%CI [94%−100%]) and 80% (95%CI [69%−92%]) respectively; 2-year CSS and 5 year CSS were 98% (95%CI [94%−100%]) and 85% (95%CI [76%−96%]) respectively; 2-year OS and 5 year OS were 88% (95%CI [79%−97%]) and 79% (95%CI [69%−92%]) respectively (Fig. S1).

3.3. Survival analysis of 2-year PFS with 5 unsupervised ML methods

As shown in Table 2, the survival analysis with the previous clustering (k = 3) did not show a statistical difference in the PFS data. The survival analysis with the new clustering grouping together clusters 1&2 showed a clinical trend, enhanced using the censored PFS at 2 years (Fig. 1 and Fig. S2 a). However, the result was still not statistically significant. We present, in Fig. 1 and Fig. S2, an example of progression-free survival with bootstrap optimization and censored data at 2 years with the 5 unsupervised machine learning methods (n = 200). With n = 200 bootstrap optimization, PCA k-means, k-sparse and spectral clustering were the most effective methods in predicting 2-year progression-free survival with bootstrap optimization (PFSb); PCA k-means 2-year PFSb: 94% (95%CI [90%−98%]) for cluster 1&2 versus 82% (95%CI [75%−91%] for cluster 3 (p = 0.01). K-sparse 2-year PFSb: 94% (95%CI [90%−98%]) for cluster 1&2 versus 82% (95%CI [74%−91%] for cluster 3 (p = 0.01). Spectral clustering also demonstrated significant efficiency for PFSb (p = 0.02) (Fig. 1, Table 3). To evaluate bootstrap reproducibility and performance, we applied a n = 1000 bootstrap optimization. PCA k-means obtained the best performance (mean HR =2 (95%CI [1.4–2.7]); probability of p ≤ 0.05; 85%) followed by k-sparse (mean HR = 1.6 (95%CI [1.1–2.4]); probability of p ≤ 0.05; 83%) and spectral clustering (mean HR = 1.48 (95%CI [1.05–2.1]); probability of p ≤ 0.05; 84%). The results of other methods were less statistically significant (Fig. S2, Tables 2–3).

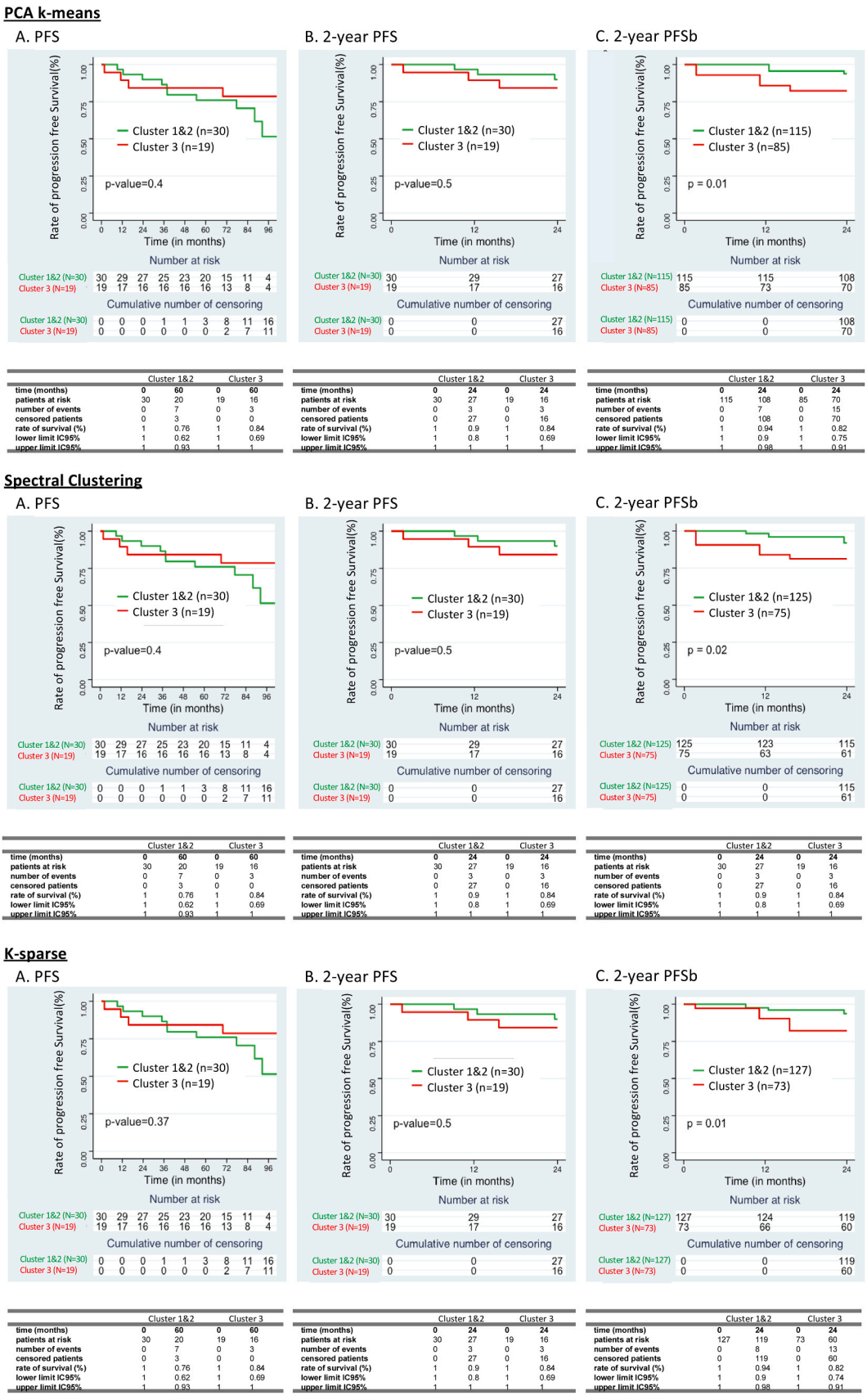
3.4. Survival Analysis of 5-year survival outcomes with 5 unsupervised ML methods

Progression free survival curves were consistent between 4 ML methods and different with SimLR clustering. In the first 2 years, cluster 3 showed lower PFS than clusters 1&2. After 2 years, events in cluster 3 became rarer while events in cluster 1&2 were consistent over time, becoming progressively numerically higher. At the end of 5 years, PFS was lower in clusters 1&2 than in cluster 3 (Fig. 2A, Table 3). With SimLR clustering, the switch occurred earlier, at 1 year (Fig. S2A). Concerning OS and CSS, results were homogenous between 4 ML methods and different with Sparse K-means. Cluster 3 had better survival outcomes, except for Sparse K-means clustering, where the trend was in disfavor of cluster 3, but only for OS (no significance reached for CSS) (Fig. 2B-C, Table 3). Only Sparse K-means OS results were consistent with *in silico* survival analysis previously performed with PREDICT Tool [16], although the difference found was not statistically significant. With a n = 1000 bootstrap optimization, Sparse K-means obtained a mean HR 1.6 (95%CI [1.2–2]) and a probability of p ≤ 0.05 81% for OS prediction.

4. Discussion

To the best of our knowledge, this proof-of-concept study is the first to compare different unsupervised methods to identify metabolomics-based prognostics signatures in BC with survival analysis. We demonstrated that K-sparse, Spectral clustering and PCA k-means methods are better at predicting 2-year PFS after bootstrap optimization than the other two ML methods. However, for CSS and OS analyses, results were not consistent with *in silico* survival analyses previously performed with PREDICT Tool, except for Sparse K-means method, and only for OS.

From a clinical point of view, the ML methods were able to identify a distinct group of patients with a poor prognosis and a high risk for early recurrence (cohort 3). The PFS behavior switch at 2-years between cluster 3 and cluster 1&2 could be explained by the heterogenicity of the entire population. Patients in cluster 3 more often had triple-negative or HER-2 overexpressed tumors, which are known to be aggressive and relapse mainly in the first 2 years. In contrast, patients in cluster 1&2 were more likely to have HR+ tumors, which are less aggressive, but present a consistent risk of relapse over time. Indeed, for patients with



	Cluster 1&2		Cluster 3	
time (months)	0	60	0	60
patients at risk	30	20	19	16
number of events	0	7	0	3
censored patients	0	3	0	0
rate of survival (%)	1	0.76	1	0.84
lower limit IC95%	1	0.62	1	0.69
upper limit IC95%	1	0.93	1	1

	Cluster 1&2		Cluster 3	
time (months)	0	24	0	24
patients at risk	30	27	19	16
number of events	0	3	0	3
censored patients	0	27	0	16
rate of survival (%)	1	0.9	1	0.84
lower limit IC95%	1	0.8	1	0.69
upper limit IC95%	1	1	1	1

	Cluster 1&2		Cluster 3	
time (months)	0	24	0	24
patients at risk	115	108	85	70
number of events	0	7	0	15
censored patients	0	108	0	70
rate of survival (%)	1	0.94	1	0.82
lower limit IC95%	1	0.9	1	0.75
upper limit IC95%	1	0.98	1	0.91

	Cluster 1&2		Cluster 3	
time (months)	0	60	0	60
patients at risk	30	20	19	16
number of events	0	7	0	3
censored patients	0	3	0	0
rate of survival (%)	1	0.76	1	0.84
lower limit IC95%	1	0.62	1	0.69
upper limit IC95%	1	0.93	1	1

	Cluster 1&2		Cluster 3	
time (months)	0	24	0	24
patients at risk	30	27	19	16
number of events	0	3	0	3
censored patients	0	27	0	16
rate of survival (%)	1	0.9	1	0.84
lower limit IC95%	1	0.8	1	0.69
upper limit IC95%	1	1	1	1

	Cluster 1&2		Cluster 3	
time (months)	0	24	0	24
patients at risk	125	123	75	63
number of events	0	3	0	3
censored patients	0	27	0	16
rate of survival (%)	1	0.9	1	0.84
lower limit IC95%	1	0.8	1	0.69
upper limit IC95%	1	1	1	1

	Cluster 1&2		Cluster 3	
time (months)	0	60	0	60
patients at risk	30	20	19	16
number of events	0	7	0	3
censored patients	0	3	0	0
rate of survival (%)	1	0.76	1	0.84
lower limit IC95%	1	0.62	1	0.69
upper limit IC95%	1	0.93	1	1

	Cluster 1&2		Cluster 3	
time (months)	0	24	0	24
patients at risk	30	27	19	16
number of events	0	3	0	3
censored patients	0	27	0	16
rate of survival (%)	1	0.9	1	0.84
lower limit IC95%	1	0.8	1	0.69
upper limit IC95%	1	1	1	1

	Cluster 1&2		Cluster 3	
time (months)	0	24	0	24
patients at risk	127	119	73	60
number of events	0	8	0	13
censored patients	0	119	0	60
rate of survival (%)	1	0.94	1	0.82
lower limit IC95%	1	0.9	1	0.74
upper limit IC95%	1	0.98	1	0.91

Fig. 1. Survival Analysis of PFS with PCA k-means, Spectral clustering and K-sparse unsupervised machine learning methods. Cluster 1 and cluster 2 were regrouped in Cluster1&2 and compared to cluster 3. (A) PFS: progression-free survival; (B) 2-year PFS: censored data at 2-year; (C) 2-year PFSb: example of progression-free survival with bootstrap optimization and censored data at 2-year. Bootstrap optimization performance details are exposed in Table 3.

Table 3
Survival outcomes with 5 different methods of unsupervised machine learning and Bootstrap optimization (k = 2).

ML Method		mean HR [95%CI]	probability of $p \leq 0.05$
k-sparse	OS	0.53 [0.4–0.7]*	82%
	CSS	0.6 [0.4–0.9]*	83%
	2-years PFS	1.6 [1.1–2.4]*	83%
PCA k-means	OS	0.5 [0.3–0.6]*	85%
	CSS	0.6 [0.4–0.9]*	85%
	2-years PFS	2 [1.4–2.7]*	85%
Spectral clustering	OS	0.5 [0.35–0.65]*	85%
	CSS	0.6 [0.4–0.8]*	85%
	2-years PFS	1.48 [1.05–2.1]*	84%
Sparse k-means	OS	1.6 [1.2–2.0]*	81%
	CSS	1.1 [0.8–1.5]	80%
	2-years PFS	1.3 [0.9–1.9]	82%
SimLR	OS	0.35 [0.25–0.45]*	83%
	CSS	0.5 [0.35–0.7]*	84%
	2-years PFS	0.65 [0.4–0.9]*	83%

ML: machine learning; k: number of clusters; OS: overall survival; CSS: cancer-specific survival; PFS: progression-free survival; HR: hazard ratio; 95CI%: 95% confidence interval; k: number of clusters; * : statistically significant.

aggressive tumors, PFS is lower the first 2 years, but for patients without relapse at 2-years, the risk of late relapse decreases compared to the risk for patients with HR+ tumors[23]. With SimLR clustering, the switch occurred earlier, at 1 year, and some late relapses were observed. This may reflect a less strict selection in cluster 3 for aggressiveness, but better performance in clustering patients with relapse overall. However, even with a median follow-up of only 85.8 months, the analyses failed to find a significant difference in terms of OS and CSS, contrasting with previously published *in silico* analyses[16]. Only Sparse K-means method yielded the expected trend, but only for OS. This result is not sufficiently consistent to recommend the use of Sparse K-means method for survival analyses. The failure of the analysis is probably due to the limited sample size and the dearth of reported events. In addition, the retrospective nature of our study may interfere with long-term follow-up and survival analyses. To finish, survival outcomes are largely dependent on histological subtype and treatment received. Therefore, future studies should analyze a specific subtype of breast cancer with a homogenous clinical setting and treatment in order to be able to study long-term outcomes.

From a methodological perspective, new clustering and bootstrap optimization may be a suitable option when the sample size is too small for significant statistical analysis. The latest genomic signature trials have examined several thousand patients to show a difference of a few percent [24–26]. For example, the RxPonder trial dealing with Oncotype DX signature randomized a total of 5083 women and 5018 participated in the trial. Among postmenopausal women, invasive disease–free survival at 5 years was 91.9% in the endocrine-only group and 91.3% in the chemoendocrine group, with no chemotherapy benefit. Among premenopausal women, invasive disease–free survival at 5 years was 89.0% with endocrine-only therapy and 93.9% with chemoendocrine therapy (hazard ratio, 0.60; 95% CI, 0.43–0.83; P = 0.002)[26]. It is therefore possible that our study size is too small to show a significant difference. Bootstrap optimization was thus applied to simulate a larger study (n = 200 and n = 1000) and see if such a study would be worthwhile to conduct, as a proof-of-principle. Therefore, in a very pragmatic manner, we compare the survival of three prognostic groups identified by unsupervised machine learning. Because of the retrospective study design and the small number of patients, no conclusion could be drawn for the prediction of CSS and OS. Future work would involve conducting a new comparison between old and new machine learning methods and deep learning methods to cluster patients based on clinical risk of relapse[27]. The field of unsupervised machine learning in bioinformatics is developing rapidly, with the emergence of new methods such as model-based clustering[28], bi-clustering [29] and deep learning. Karim M. *and al.* have published a recent review that evaluates different deep learning-based unsupervised machine learning methods for solving emerging problems in bioinformatics research[30]. Yet it is worth noting that, even though deep learning methods are of particular interest in many fields, they require a very large number of patients to be efficiently trained and may therefore not be suitable for small metabolomics datasets obtained on real life patients, such as the one we have used. While obtaining imaging or clinical data concerning several thousands of patients seems achievable, obtaining metabolomics data for that many patients is currently much more complicated. Furthermore, even though some efforts are being made to tackle this issue[30], it is currently impossible to understand which features are responsible for the outcome when using deep-learning clustering techniques. It would therefore be impossible to understand the metabolic differences underlying different patient clusters if deep learning clustering was used. A supervised analysis could also be worthwhile but would require a more homogeneous, larger population.

From a biological point of view, only few studies have reported associations between metabolic alterations and early BC patient survival. To our knowledge, no study has been performed on tumor tissue, but only on serum. Fahrmann et al. reported serum analyses of Diacetylspermine in patients with triple negative breast cancer (TNBC)[31]. Diacetylspermine levels were higher in serum samples from patients with triple-negative breast cancer than in samples from patients without triple-negative breast cancer and from healthy volunteers. In a prospective cohort, the authors observed that serum Diacetylspermine levels were significantly increased in patients with early recurrence (<1 year). Higher serum Diacetylspermine levels were also associated with lower 5-year distant metastasis-free survival and 5-year overall survival. Asiago et al. published very interesting results on early detection of recurrent breast cancer using metabolite profiling with 7 metabolite markers. More than a half of the patients were predicted to have recurrence 13 months (on average) before the recurrence clinical diagnosis. However, this metabolomic signature provides for early detection as opposed to prediction of relapse. Oakman et al. calculated individual early patient 'metabolomic risk' derived from forty-four early breast cancer patients compared with fifty-one metastatic patients who served as control. Metabolomic risk was compared with the Adjuvantionline 10-year mortality estimate. The comparison with Adjuvantionline revealed discordance like in our study. Of 21 patients assessed as high-risk by Adjuvantionline, 10 (48%) and 6 (29%) were at high metabolomic risk pre- and postoperatively, respectively. Of the 23 low-risk patients evaluated by Adjuvantionline, 11 (48%) preoperatively and 20 (87%) postoperatively were at low metabolomic risk. However, these simulated data may also be biased, hence the value of our study and future studies on real survival data to distinguish limitations due to metabolomics from those due to simulated survival data.

To finish, there are several limitations in this work: small number of patients, heterogeneous population, retrospective study design and predicted metabolites. However, the preliminary results obtained despite these limitations clearly highlight the potential contribution of metabolomics, as a proof or principle.

5. Conclusion

The objective of our study was to compare different unsupervised machine learning algorithms on untargeted metabolomics data and to evaluate the performance of these methods in predicting survival outcomes. Our results showed that it is possible to use unsupervised machine learning methods on metabolomic unlabeled data to identify clusters of patients with worse 2-year PFS. Among the 5 unsupervised ML methods reported here, PCA k-means, K-sparse and spectral clustering outperformed the other two unsupervised methods. However, because of the retrospective study design and the small number of patients, no conclusion could be drawn in terms of predicting CSS and OS. Future studies are needed with a larger population of specific

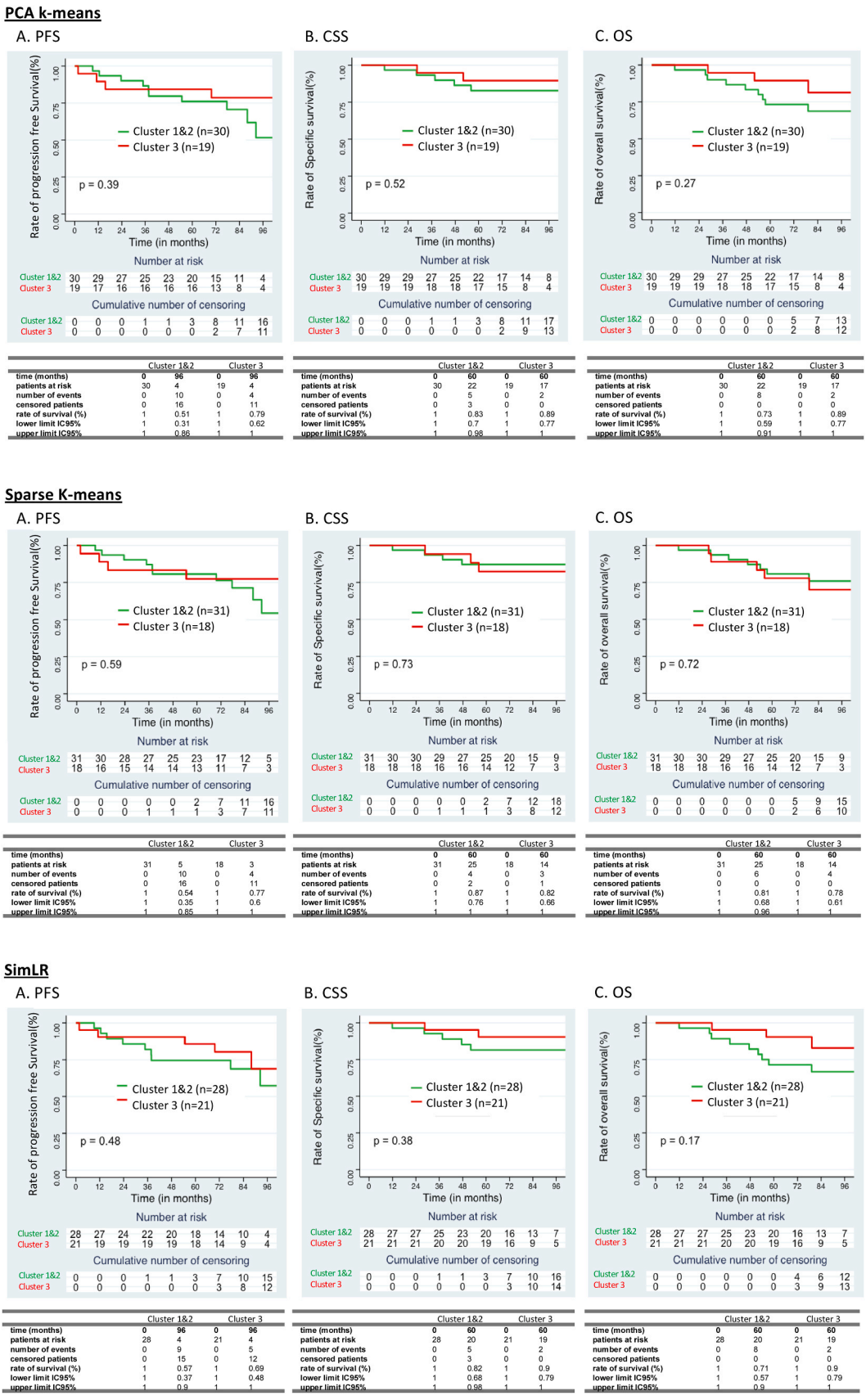


Fig. 2. Survival Analysis with extended follow-up with PCA k-means, Sparse K-means et SimLR. Cluster 1 and cluster 2 were regrouped in Cluster1&2 and compared to cluster 3. (A) PFS: progression free survival; (B) CSS: cancer-specific survival; (C) OS: overall survival. Bootstrap optimization performance details are exposed in Table 3.

histological subtypes.

CRediT authorship contribution statement

Conception and design: **C.B., J.G., D.C., T.P.**; development of methodology: **J.G.**; acquisition of data: **O.H., C.B., J-M.G, Y.C.**; analysis and interpretation of data: **J.G, C.B., T.P.**; writing, review, and/or revision of the manuscript: **C.B., J.G., T.P.** and all the authors; study supervision: **T.P., O.H., J-M.F.**

Grants and Support

Equipment for this study was purchased through grants from the Recherche en Matières de Sûreté Nucléaire et Radioprotection program, from the French National Research Agency and the Conseil Départemental 06.

Declaration of Competing Interest

No potential conflicts of interest were disclosed.

Acknowledgements

The authors acknowledge support from Centre Antoine Lacassagne and TIRO Unit, University Côte d′Azur, France.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.033.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J Clin janv 2017; 67(1):7–30.
[2] Gautier Defossez, Sandra Le Guyader-Peyrou, Zoé Uhry, Pascale Grosclaude, Marc Colonna, Emmanuelle Dantony, et al. Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018 - Étude à partir des registres des cancers du réseau Francim. In 2019.
[3] Santé publique, france. Cancer du sein. 2018;
[4] INCA. Panorama des cancers en France. 2022e éd.
[5] Vaz-Luis I, Cottu P, Mesleard C, Martin AL, Dumas A, Dauchy S, et al. UNICANCER: French prospective cohort study of treatment-related chronic toxicity in women with localised breast cancer (CANTO). ESMO Open 2019;4(5):e000562.
[6] Rosenstock AS, Niu J, Giordano SH, Zhao H, Wolff AC, Chavez-MacGregor M. Acute myeloid leukemia and myelodysplastic syndrome after adjuvant chemotherapy: A population-based study among older breast cancer patients: AML/MDS After Adjuvant Chemotherapy. 1 mars Cancer 2018;124(5):899–906.
[7] Gaude E, Frezza C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. 10 oct Nat Commun 2016;7(1):13041.
[8] Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using the cancer genome atlas. 14 déc Nat Commun 2018;9(1):5330.
[9] Wang L, Zhang S, Wang X. The metabolic mechanisms of breast cancer metastasis. 7 janv Front Oncol 2021;10:602416.
[10] Bernhardt S, Bayerlová M, Vetter M, Wachter A, Mitra D, Hanf V, et al. Proteomic profiling of breast cancer metabolism identifies SHMT2 and ASCT2 as prognostic factors (déc) Breast Cancer Res 2017;19(1):112.
[11] Roberts LD, Souza AL, Gerszten RE, Clish CB. Targeted metabolomics (⟨https://onlinelibrary.wiley.com/⟩doi/) Curr Protoc Mol Biol [Internet] 2012;98(1). https://doi.org/10.1002/0471142727.mb3002s98.
[12] Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies—challenges and emerging directions. J Am Soc Mass Spectrom 2016;27(12):1897–905.
[13] Vinayavekhin N, Saghatelian A. Untargeted Metabolomics. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, et al., editors. Current Protocols in Molecular Biology. Hoboken, NJ, USA: John Wiley & Sons Inc; 2010. https://doi.org/10.1002/0471142727.mb3001s90.
[14] Galal A, Talal M, Moustafa A. Applications of machine learning in metabolomics: disease modeling and classification. 24 nov Front Genet 2022;13:1017340.
[15] Dhall D, Kaur R, Juneja M. Machine Learning: A Review of the Algorithms and Its Applications ([Internet]). In: Singh PK, Kar AK, Singh Y, Kolekar MH, Tanwar S, editors. Proceedings of ICRIC 2019, vol. 597. Cham: Springer International Publishing; 2020. p. 47–63 ([Internet]), http://link.springer.com/10.1007/978-3-030-29407-6_5 ([Internet]).
[16] Gal J, Bailleux C, Chardin D, Pourcher T, Gilhodes J, Jing L, et al. Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer. Comput Struct Biotechnol J 2020;18: 1509–24.
[17] Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer (févr) Breast Cancer Res 2010;12(1):R1.
[18] Candido dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation (déc) Breast Cancer Res 2017;19(1): 58.
[19] Asiago VM, Alvarado LZ, Shanaiah N, Gowda GAN, Owusu-Sarfo K, Ballas RA, et al. Early detection of recurrent breast cancer using metabolite profiling. 1 nov Cancer Res 2010;70(21):8309–18.
[20] Oakman C, Tenori L, Claudino WM, Cappadona S, Nepi S, Battaglia A, et al. Identification of a serum-detectable metabolomic fingerprint potentially correlated with the presence of micrometastatic disease in early breast cancer patients at varying risks of disease relapse by traditional prognostic methods (juin) Ann Oncol 2011;22(6):1295–301.
[21] Alakwaa FM, Savelieff MG. Bioinformatics analysis of metabolomics data unveils association of metabolic signatures with methylation in breast cancer. 2 juill J Proteome Res 2020;19(7):2879–89.
[22] Eldridge R, Qin Z, Saba N, Houser M, Hayes D, Miller A, et al. Unsupervised hierarchical clustering of head and neck cancer patients by pre-treatment plasma metabolomics creates prognostic metabolic subtypes. 14 juin Cancers 2023;15(12): 3184.
[23] Darlix A, Louvel G, Fraisse J, Jacot W, Brain E, Debled M, et al. Impact of breast cancer molecular subtypes on the incidence, kinetics and prognosis of central nervous system metastases in a large multicentre real-life cohort. 10 déc Br J Cancer 2019;121(12):991–1000.
[24] Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. 12 juill N Engl J Med 2018;379(2):111–21.
[25] Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. 25 août N Engl J Med 2016;375(8):717–29.
[26] Kalinsky K, Barlow WE, Gralow JR, Meric-Bernstam F, Albain KS, Hayes DF, et al. 21-Gene assay to inform chemotherapy benefit in node-positive breast cancer. 16 déc N Engl J Med 2021;385(25):2336–47.
[27] Kumar I, Singh SP, Shivam. Machine learning in bioinformatics ([Internet]). Bioinformatics. Elsevier; 2022. p. 443–56 ([Internet]), https://linkinghub.elsevier.com/retrieve/pii/B9780323897754000201 ([Internet]).
[28] Gormley IC, Murphy TB, Raftery AE. Model-Based Clustering. 10 mars Annu Rev Stat Its Appl 2023;10(1):573–95.
[29] Ramkumar M, Basker N, Pradeep D, Prajapati R, Yuvaraj N, Arshath Raja R, et al. Healthcare biclustering-based prediction on gene expression dataset. Teekaraman Y, éditeur. 22 févr BioMed Res Int 2022;2022:1–7.
[30] Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, et al. Deep learning-based clustering approaches for bioinformatics. 18 janv Brief Bioinform 2021;22(1):393–415.
[31] Fahrmann JF, Vykoukal J, Fleury A, Tripathi S, Dennison JB, Murage E, et al. Association between plasma diacetylspermine and tumor spermine synthase with outcome in triple-negative breast cancer. 1 juin JNCI J Natl Cancer Inst 2020;112 (6):607–16.