Application d'algorithmes de machine learning pour l'exploitation de données omiques en oncologie

Jocelyn Gal

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES DE LA VIE ET DE LA SANTÉ

THÈSE DE DOCTORAT

Application d'algorithmes de machine learning pour l'exploitation de données omiques en oncologie

Jocelyn Gal

Unité de Pharmacogénétique et radiogénétique des cancers (UPRC) EA7497
Centre Antoine Lacassagne, F-06189, Nice, France

Présentée en vue de l'obtention du grade de docteur en Sciences de la vie et de la Santé
Mention : Recherche clinique et thérapeutique
et de l'Université Côte d'Azur
Dirigée par : Emmanuel Chamorey
Soutenue le : 28/11/2019

Devant le jury, composé de :
Pascal Staccini, PU-PH, CHU de Nice
Simone Mathoulin-Pelissier, PU-PH, UFR médecine, Bordeaux
Thomas Filleron, Dr, Institut Claudius-Regaud, Toulouse
Jean-Marc Ferrero, PU-PH, Centre Antoine Lacassagne, Nice
Joseph Ciccolini, Professeur, UMR U1068, Aix-Marseille Univ
Gérard Milano, Dr, Centre Antoine Lacassagne, Nice

Application d'algorithmes de machine learning pour l'exploitation de données omiques en oncologie

Jury :

Président du jury

Pascal Staccini, PU-PH, CHU de Nice

Rapporteurs

Simone Mathoulin-Pelissier, PU-PH, UFR médecine, Bordeaux
Thomas Filleron, Dr, Institut Claudius-Regaud, Toulouse

Examinateurs

Jean-Marc Ferrero, PU-PH, Centre Antoine Lacassagne, Nice
Joseph Ciccolini, Professeur, UMR U1068, Aix-Marseille Université

Invités

Gérard Milano, Dr, Centre Antoine Lacassagne, Nice

Résumé

Le développement de l'informatique en médecine et en biologie a permis de générer un grand volume de données. La complexité et la quantité d'informations à intégrer lors d'une prise de décision médicale ont largement dépassé les capacités humaines. Ces informations comprennent des variables démographiques, cliniques ou radiologiques mais également des variables biologiques et en particulier omiques (génomique, protéomique, transcriptomique et métabolomique) caractérisées par un grand nombre de variables mesurées relativement au faible nombre de patients. Leur analyse représente un véritable défi dans la mesure où elles sont fréquemment « bruitées » et associées à des situations de multi-colinéarité. De nos jours, la puissance de calcul permet d'identifier des modèles cliniquement pertinents parmi cet ensemble de données en utilisant des algorithmes d'apprentissage automatique. A travers cette thèse, notre objectif est d'appliquer des méthodes d'apprentissage supervisé et non supervisé, à des données biologiques de grande dimension, dans le but de participer à l'optimisation de la classification et de la prise en charge thérapeutique des patients atteints de cancers. La première partie de ce travail consiste à appliquer une méthode d'apprentissage supervisé à des données d'immunogénétique germinale pour prédire l'efficacité thérapeutique et la toxicité d'un traitement par inhibiteur de point de contrôle immunitaire. La deuxième partie compare différentes méthodes d'apprentissage non supervisé permettant d'évaluer l'apport de la métabolomique dans le diagnostic et la prise en charge des cancers du sein en situation adjuvante. Enfin la troisième partie de ce travail a pour but d'exposer l'apport que peuvent présenter les essais thérapeutiques simulés en recherche biomédicale. L'application des méthodes d'apprentissage automatique en oncologie offre de nouvelles perspectives aux cliniciens leur permettant ainsi de poser des diagnostics plus rapidement et plus précisément, ou encore d'optimiser la prise en charge thérapeutique en termes d'efficacité et de toxicité.

Mots-clés Intelligence artificielle, Apprentissage automatique non supervisé, Apprentissage automatique supervisé, Médecine de précision, Oncologie, Métabolomique, Génomique.

Abstract

The development of computer science in medicine and biology has generated a large volume of data. The complexity and the amount of information to be integrated for optimal decision-making in medicine have largely exceeded human capacities. These data includes demographic, clinical and radiological variables, but also biological variables and particularly omics (genomics, proteomics, transcriptomics and metabolomics) characterized by a large number of measured variables relatively to a generally small number of patients. Their analysis represents a real challenge as they are frequently "noisy" and associated with situations of multi-colinearity. Nowadays, computational power makes it possible to identify clinically relevant models within these sets of data by using machine learning algorithms. Through this thesis, our goal is to apply supervised and unsupervised learning methods, to large biological data, in order to participate in the optimization of the classification and therapeutic management of patients with various types of cancer. In the first part of this work a supervised learning method is applied to germline immunogenetic data to predict the efficacy and toxicity of immune checkpoint inhibitor therapy. In the second part, different unsupervised learning methods are compared to evaluate the contribution of metabolomics in the diagnosis and management of breast cancer. Finally, the third part of this work aims to expose the contribution that simulated therapeutic trials can make in biomedical research. The application of machine learning methods in oncology offers new perspectives to clinicians allowing them to make diagnostics faster and more accurately, or to optimize therapeutic management in terms of efficacy and toxicity.

Keywords Artificial intelligence, unsupervised machine learning, supervised machine learning, Precision medicine, Oncology, Metabolomics, Genomics.

Remerciements

Tout d'abord, je tiens à remercier très chaleureusement mon directeur de thèse **Emmanuel Chamorey** pour son encadrement, son professionnalisme et les connaissances qu'il a su me transmettre. Merci pour ton temps et toute l'attention que tu m'as consacrée ainsi que la confiance que tu m'as accordée durant toutes ces années. Je ne serais sûrement pas là aujourd'hui sans toi. Merci pour tout.

Je remercie aussi tout particulièrement, **Gérard Milano** pour m'avoir accueilli dans son laboratoire pendant mes trois années de thèse. Réaliser mon doctorat dans son équipe a été une excellente opportunité pour moi et je le remercie plus particulièrement pour ses conseils avisés notamment lors de la rédaction de l'article et de la thèse. J'ai tellement appris à vos côtés.

Je tiens à exprimer toute ma gratitude à Mme **Simone Mathoulin-Pelissier** ainsi qu'à M. **Thomas Filleron** pour avoir accepté de juger mes travaux en tant que rapporteurs.

Mes remerciements les plus chaleureux vont également à M. **Pascal Staccini**, **M. Jean-Marc Ferrero** et M. **Joseph Ciccolini** qui ont accepté de prendre part à mon jury de soutenance de thèse.

Dans le cadre de cette thèse, j'ai eu la chance et le plaisir de collaborer avec des personnes de différents domaines dont notamment **Olivier Humbert** et **David Chardin** pour la partie médicale, **Thierry Pourcher** pour la partie biologique et **Michel Barlaud** pour la partie machine learning. Merci à vous pour votre aide qui a été si précieuse.

Et comment ne pas te remercier **Caroline** pour tout ce que tu as pu m'apporter notamment ta générosité, le partage de tes connaissances, tes nombreuses idées et ton enthousiasme communicatif. Surtout ne change rien, reste comme tu es !

Un grand merci à tous les membres du Département d'Épidémiologie, et des données de la santé qui chacun à leur manière, m'ont aidé à m'affranchir de cette épreuve. **Renaud** tu es quelqu'un d'humain qui, à son tour, va s'élancer dans l'aventure, toi **Yann** sous tes airs sérieux tu es un sacré personnage, et toi **Julien** je te souhaite tous mes vœux de réussite dans ta nouvelle vie Marseillaise.

Merci à toi **Julia** pour toute l'aide que tu m'as donnée. J'espère que nous pourrons collaborer le plus souvent possible car tu es quelqu'un avec une réelle valeur.

J'aimerais remercier, ma compagne **Marjorie** et mon fils **Bastien** pour m'avoir supporté durant ces 3 années, mes parents pour leur très précieux soutien et leur présence au quotidien.

Enfin, je ne pourrais finir sans dire merci à mes amis qui de près ou de loin m'ont toujours soutenu, je veux citer **Horace**, **Thomas**, **Benoit**, **Stéphane**, **Jean-Paul**, **David** et tous mes amis de la caserne de pompiers de Beuil, **Jean-Luc**, **Isabelle, Laurent, Christophe**, **Romain**, **Stéphane** et tous les autres.

A vous tous je vous dédie cette thèse.

Table des matières

Production scientifique

Bilan au 22/10/2019

Nombre de Publications : 60
Indice h : 16
Indice h 10 : 21
Nombre de points SIGAPS : 545

En lien avec la thèse

Publications dans des revues avec comité de relecture

1. **Gal J**, Milano G, Ferrero JM, Saada-Bouzid E, Viotti J, Chabaud S, Gougis P, Le Tourneau C, Schiappa R, Paquet A and Chamorey E. Optimizing drug development in oncology by clinical trial simulation: Why and how? Brief Bioinform. 2018; 19:1203-1217.

2. Refae S, **Gal J**, Ebran N, Otto J, Borchiellini D, Peyrade F, Chamorey E, Brest P, Milano G and Saâda-Bouzid E. Germinal Immunogenetics predict treatment outcome for PD-1/PD-L1 checkpoint inhibitors. Invest New Drugs. 2019.

3. **Gal J**, Bailleux C, Chardin D, Pourcher T, Gilhodes J, Jing L, Guigonis JM, Ferrero JM, Milano G, Mograbi B, Brest P, Château Y, Humbert O and Chamorey E. Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer. Brief Bioinform. 2019. *Soumis*

Communications orales

1. **Gal J**, Bailleux C, Chardin D, Pourcher T, Schiappa R, Gilhodes J, Humbert O et Chamorey E. Comparaison de 5 différentes méthodes d'apprentissage non supervisé dans le cas de données de grandes dimensions. Application dans le cancer du sein. **MétaSUD 2019; Journées de Métabolomique en Région Sud – Provence-Alpes-Côte d'Azur,** June 2019, Toulon, France

2. **Gal J**, Bailleux C, Chardin D, Pourcher T, Schiappa R, Gilhodes J, Humbert O et Chamorey E. Comparaison de différentes méthodes d'apprentissage non-supervisé dans le cas de données de grandes dimensions. Application dans le cancer du sein. **EPICLIN 12 / 25èmes Journées des Statisticiens des CLCC**, May 2019, Toulouse, France.

3. **Gal J**, Milano G, Bailleux C, Ettaiche M, Paquet A, Gougis P, Borchiellini D, Ferrero JM et Chamorey E. Optimisation du processus de développement d'un nouveau médicament par modélisation et simulation: état des lieux et enjeux. **EPICLIN 10 / 23èmes Journées des Statisticiens des CLCC**, May 2017, Strasbourg, France.

Communications affichées

1. **Gal J**, Bailleux C, Chardin D, Pourcher T, Ferrero JM, Barranger E, Humbert O et Chamorey E. Comparaison de 5 méthodes de machine learning non supervisé appliquées à des données de métabolomique chez des patientes atteintes d'un cancer du sein. **41èmes Journées de la Société Française de Sénologie et de Pathologie Mammaire**, November, 2019, Marseille, France.

2. **Gal J**, Bailleux C, Chardin D, Pourcher T, Schiappa R, Gilhodes J, Humbert O et Chamorey E. Unsupervised machine learning methods reveal metabolomic based clusters in breast cancer

patients. **Séminaire annuel du Cancéropole Provence-Alpes-Côte-d'Azur**, July 2019, St-Raphaël, France

3. **Gal J**, Bailleux C, Chardin D, Pourcher T, Jing L, Guigonis JM, Ferrero JM, Schiappa R, Chamorey E and Humbert O. Unsupervised machine learning methods reveal metabolomic based clusters in breast cancer patients. **American Association for Cancer Research (AACR)**, April, 2019, Atlanta, United states of America.

4. **Gal J**, Barlaud M, Pourcher T, Bailleux C, Jing L, Chamorey E and Humbert O. Étude des marqueurs métaboliques du cancer du sein adjuvant: comparaison de différentes méthodes de clustering. **EPICLIN 11 / 24èmes Journées des Statisticiens des CLCC**, June 2018, Nice, France.

5. **Gal J**, Milano G, Viotti J, Schiappa R, Dugué A, Paquet A, Chabaud S, Ferrero JM and Chamorey E. Optimizing drug regimens in oncology by clinical trial simulations: Why and how. **American Association for Cancer Research (AACR),** April, 2017**,** Washington, United states of America

Ouvrages

1. **Gal J** et Gilhodes J. Ecriture du chapitre: « Analyse statistique des données d'expression de gènes issues de puces à ADN » du livre Methodes biostatistiques appliquées à la recherche clinique en cancérologie: John Libbey Eurotext; 2011.

Autres

Publications dans des revues avec comité de relecture

i. Refae S, **Gal J**, Brest P and Milano G. Germinal immunogenetics as a predictive factor for immunotherapy. *Critical reviews in oncology/hematology*. 2019.

ii. Refae S, **Gal J**, Ebran N, Brest P, Peyrade F, Guigay J, Chateau Y, Milano G and Saada-bouzid E. Predicting checkpoint inhibitor treatment outcome in head and neck cancer patients: a potential role for host immunogenetics. *Head and Neck*. 2019;*Under review*.

iii. Refae S, **Gal J**, Brest P, Giacchero D, Borchiellini D, Ebran N, Peyrade F, Guigay J, Milano G and Saâda-Bouzid E. Hyperprogression under Immune Checkpoint Inhibitor: a potential role for germinal immunogenetics. *Scientific Reports*. 2019 ; *Under review*

iv. Chardin D, Pourcher T, **Gal J**, Guigonis J, Bailleux C, Darcourt J, Humbert O and Arnould L. Métabolomique et imagerie TEP-FDG des cancers du sein. *Medecine Nucléaire*. 2019. *Under review*

Communications affichées

i. Milano G, Refae S, **Gal J**, Ebran N, Otto J, Shell S, Everts R, Chamorey E and Saâda-Bouzid E. A SNP germinal signature for predicting checkpoint inhibitor treatment outcome: **American Association for Cancer Research (AACR)**, April, 2019, Atlanta, United States of America.

ii. Saâda-Bouzid E, Refae S, Ebran N, **Gal J**, Peyrade F, Guigay J and Milano G. Variations in PD1, PD-L1, IDO1 and VEGR2 genes and association with outcomes in advanced head and neck squamous cell carcinoma (HNSCC) patients treated with anti-PD1/PD-L1 based immunotherapy. **American Society of Clinical Oncology (ASCO) Annual Meeting,** June, 2018, Chicago, United States of America

iii. Refae S, Ebran N, **Gal J**, Otto J, Giacchero D, Borchiellini D, Guigay J, Peyrade F, Milano G and Saâda E. Host immunogenetics and hyperprogression under PD1/PD-L1 checkpoint inhibitors. **American Association for Cancer Research (AACR)**, April, 2018, Atlanta, United States of America.

Liste des abréviations

Les abréviations indiquées ci-dessous sont en anglais, car ce sont celles communément admise par la communauté scientifique.

AI: Artificial Intelligence
ANOVA: Analysis Of Variance
ANN: Artificial Neural Network
AUC: Area Under the Curve
CNN: Convolutional Neural Network
CTS: Clinical Trial Simulation
DL: Depp Learning
GC: Gas chromatography
GBM: Gradient Boosting Machine
HCA: Hierarchical Clustering Analysis
HMDB: Human Metabolom Database
HPLC: High Performance Liquid Chromatography
KNN: K-Nearest Neighbor
LC: Liquid chromatography
NMR: Nuclear magnetic resonance
ML: Machine learning
MS: Mass Spectrometry
PCA: Principal Components Analysis
PCR: Principal Components Regression
PLS-DA: Partial Least Squares regression
RF: Random Forest
ROC: Receiver Operating Characteristic
SVM: Support Vector Machine
SIMLR: Single-cell Interpretation via Multikernel Learning
SK-Sparse: Supervised K-sparse

Liste des figures

Partie I : Contexte général

I. Partie I : Contexte général

En règle générale en oncologie, pour une même localisation tumorale, le traitement appliqué diffère peu entre les patients. Cependant, chaque patient est unique d'un point de vue génétique et il existe une multitude de sous-types pour un même cancer. Avec l'avènement des technologies omiques dont les plus connues sont la génomique, la protéomique, la transcriptomique et la métabolomique, la médecine personnalisée est devenue une application majeure en oncologie [1]. Elle consiste en l'attribution du traitement le plus approprié en fonction du profil du patient lui-même et de son cancer. Les progrès technologiques, réalisés par exemple dans le domaine du séquençage haut-débit, permettent de générer d'importants volumes de données, et ceci à différentes échelles du vivant. Les données ainsi obtenues, appelées données omiques sont caractérisées comme des données dites de grande dimension. Ce type de données, souvent associées à un faible nombre de patients [2], pose néanmoins un certain nombre de problèmes aux statistiques classiques, comme ceux de multi-colinéarité, d'inférence statistique ou de biais des estimateurs. On parle alors de « fléau de la dimension », terme introduit en 1956 par Bellman [3]. Dans un tel contexte, des méthodes statistiques capables de pallier à ces problèmes ont été développées ces dernières années. Ces dernières peuvent être regroupées sous le terme d'apprentissage artificiel ou plus communément d'intelligence artificielle (IA).

Dans le cadre de nos travaux, l'IA a été choisie comme outil pour le traitement de données omiques.

I.1. Intelligence artificielle

I.1.1. Historique

Le concept d'intelligence artificielle (IA) est né en 1950 avec la parution d'un article intitulé «Computing Machinery and Intelligence» [4] publié par le mathématicien Alan Turing dans lequel il explore le problème de définir si une machine est consciente ou non. Il résultera de cet article le Test de Turing qui consiste à évaluer les capacités d'un ordinateur à imiter les réponses humaines. Le terme d'IA est officialisé en 1956 lors d'une conférence aux États-Unis qui s'est tenue au Dartmouth College [5]. L'intelligence artificielle a ensuite été définie par *Marvin Lee Minsky [6]*, comme: «*la construction de programmes informatiques qui s'adonnent à des tâches qui sont pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que: l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critiqué*». Entre les années 1990 et les années 2000, avec le développement croissant de l'informatique et la modernisation de la loi de Moore [7] qui traite de l'évolution de la puissance de calcul des ordinateurs et de la complexité du matériel informatique [8], l'IA commence à être introduite dans des domaines de pointe comme l'aéronautique ou la médecine. De 2000 à 2010, internet se déploie, les téléphones portables voient le jour et les ordinateurs deviennent de plus en plus accessibles. C'est à partir de cette période que la quantité de données créées et stockées ne cessera de croître de façon exponentielle nous faisant rentrer dans le monde du «big data». L'IA s'illustre grâce aux performances d'IBM Watson en 2011 avec un ordinateur qui bat les deux plus grands champions de *Jeopardy!* (à partir d'indices, le but est de trouver la question correspondante). La combinaison simultanée de 3 facteurs va permettre à l'IA de franchir une nouvelle étape en résolvant des problèmes jusqu'alors inconcevables pour un être humain : 1° L'introduction d'une nouvelle catégorie d'algorithmes (les réseaux de neurones convolutionnels [9]); 2° L'arrivée sur le marché de processeurs graphiques capables d'effectuer d'énormes quantités de calculs en un minimum de temps; 3° La disponibilité de très grandes bases de données permettant un apprentissage plus fin (par exemple ImageNet qui est la plus grande base de données annotées [10]). Les premiers réseaux de neurones partaient de données numériques, qui pouvaient être une image, et essayaient de transformer ces données en une information, par exemple «c'est l'image d'un chien», via des algorithmes qui s'inspiraient du fonctionnement des neurones du cerveau. Ce sont ces algorithmes qui sont aujourd'hui utilisés dans les grandes applications de l'IA, comme la reconnaissance de la parole de Siri, la voiture autonome de Google ou bien la reconnaissance d'images de Facebook.

I.1.2. Les différentes techniques d'intelligence artificielle

La prise de décision médicale à partir d'informations tirées de données historiques est la nature même de la médecine factuelle. Traditionnellement, les méthodes statistiques «classiques» ont abordé cette tâche en caractérisant les structures des données médicales sous la forme de modèles, comme par exemple les régressions linéaires ou les régressions logistiques. Avec l'augmentation croissante du volume des données médicales, l'IA fournit des techniques qui permettent de découvrir des associations complexes qui ne peuvent pas être facilement réduites en une « simple » régression. De plus, contrairement à un seul clinicien, les techniques d'IA peuvent simultanément observer et traiter rapidement un nombre presque illimité de données. Récemment, une application basée sur l'IA a été capable de surpasser les dermatologues en matière de classification correcte des lésions cutanées suspectes [11]. Un article paru en avril 2019 dans le New England Journal of Medecine [12] expliquait, par un exemple fictif, ce que pourrait apporter l'utilisation de l'IA par rapport à une prise en charge standard d'un patient en oncologie.

L'IA constitue un très vaste champ d'applications et cette thèse se concentre exclusivement sur les principales méthodes d'apprentissage automatique (*figure 1*) en raison de leur utilisation omniprésente dans d'importantes applications cliniques.

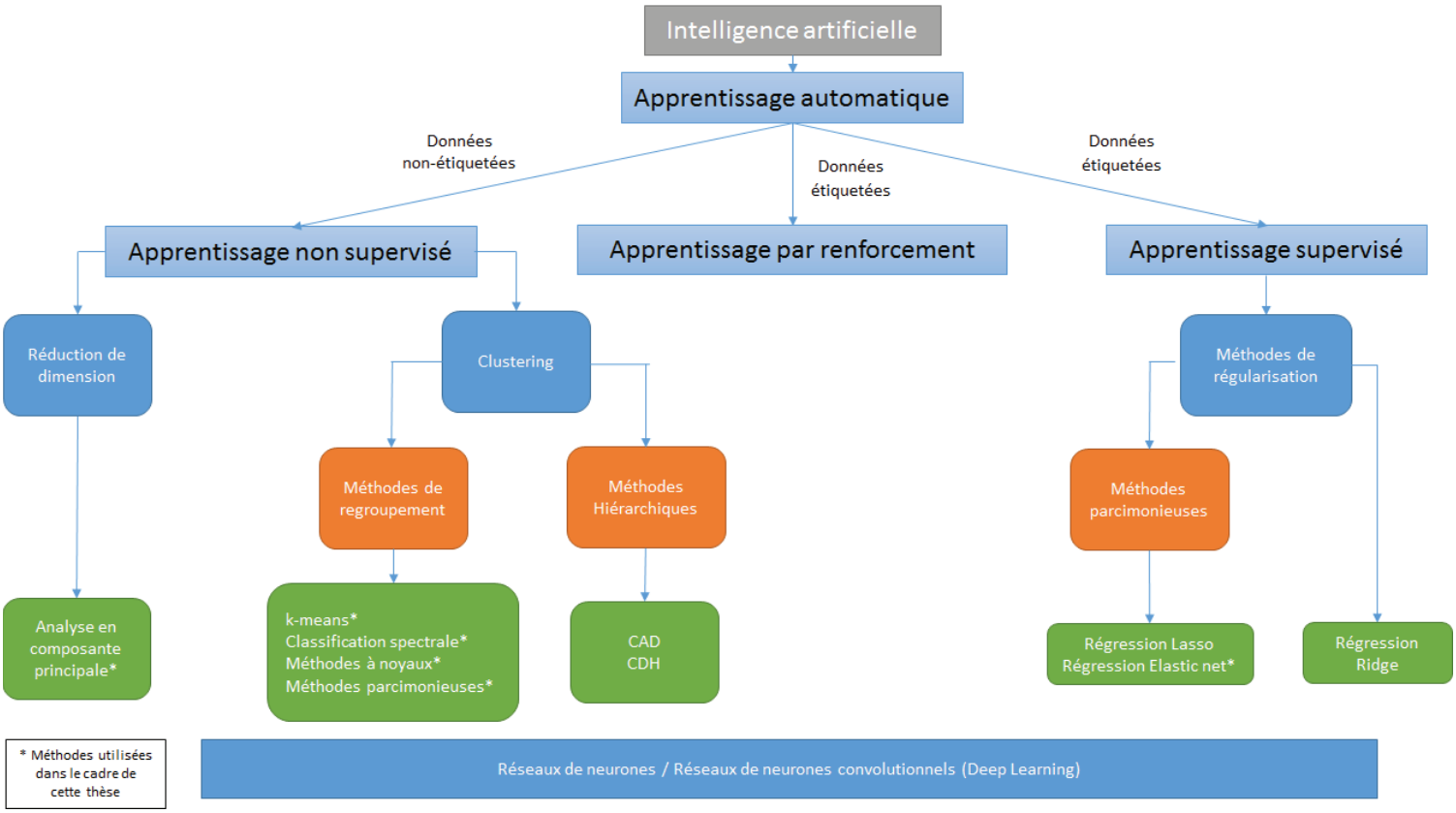


Figure 1 : De l'intelligence artificielle aux réseaux de neurones convolutionnels

I.1.3. L'apprentissage automatique

L'apprentissage automatique (ou machine learning en anglais) est une branche de l'IA qui consiste à modéliser une fonction à partir de données chiffrées. De nos jours, il joue un rôle prépondérant en médecine et en biologie [13, 14]. Le cancer est une maladie hétérogène constituée de nombreux sous-types histologiques. Le diagnostic précoce et le pronostic d'un type de cancer sont devenus une nécessité en oncologie, car ils peuvent faciliter la prise en charge des patients. L'importance de la classification des patients atteints d'un cancer dans des groupes à risque élevé ou faible a conduit de nombreuses équipes de recherche à appliquer différentes méthodes d'apprentissage automatique. Grâce à l'utilisation de ces méthodes, la performance des prédictions en oncologie a été améliorée de 15% à 20% ces dix dernières années [14]. C'est ainsi que les progrès scientifiques réalisés ont permis d'améliorer les capacités de mesure et de calcul, et il est à présent pratiquement impossible pour un être humain de traiter de façon globale l'ensemble des données générées dans un temps raisonnable. Certaines spécialités médicales comme l'imagerie ou la biologie ont profité d'évolutions technologiques importantes et ont conduit les spécialistes de ces domaines à devoir repenser leur pratique en prenant en compte les informations que ces technologies leur procuraient.

Les techniques d'apprentissage automatique sont dérivées des modélisations statistiques classiques telles que les régressions mais ont également donné naissance au développement de méthodes plus complexes comme les réseaux de neurones convolutionnels [9]. L'apprentissage automatique repose d'une part sur des données à partir desquelles les algorithmes[1] vont apprendre et d'autre part sur l'algorithme utilisé qui correspond à une procédure que l'on exécute sur ces données pour créer un modèle. Le défi en apprentissage automatique est de développer un algorithme capable de résoudre efficacement des problèmes de plus en plus complexes. Chaque algorithme d'apprentissage suggère sa propre stratégie afin de généraliser les données représentant le phénomène à apprendre. Les algorithmes communément utilisés en pratique sont adoptés en vertu de leur efficacité empirique, mais aussi pour leur rapidité d'exécution. Selon l'objectif, il existe trois grandes classes de méthodes d'apprentissage : l'apprentissage non supervisé, l'apprentissage supervisé et l'apprentissage par renforcement.

L'apprentissage profond (deep learning) est une branche particulière de l'apprentissage automatique qui permet aux machines d'assimiler des connaissances. L'apprentissage profond regroupe une famille de modèles d'apprentissage automatique, décrits sous forme de réseaux de neurones organisés en couches [15]. Le terme « profond » se rapporte au nombre de couches cachées au sein du réseau de neurones. Les réseaux de neurones « classiques » comportent entre deux et trois couches cachées, alors que les réseaux profonds peuvent en dénombrer jusqu'à 150. Le réseau de neurones convolutionnels (CNN ou ConvNet) [9] est l'un des algorithmes les plus répandus de l'apprentissage

[1]Un algorithme est un processus ou une succession d'étapes à exécutées dans les calculs

profond. Contrairement à une méthode d'apprentissage « classique » qui a des règles et des conditions d'applications bien définies, l'apprentissage profond permet au système d'être plus flexible.

Deux points différencient l'apprentissage automatique de l'apprentissage profond : 1° l'extraction des caractéristiques (ou features en anglais) se fait manuellement avec un algorithme d'apprentissage automatique alors qu'elle se réalise de manière automatique avec un algorithme d'apprentissage profond. 2° Les techniques d'apprentissage profond ont la capacité de continuer à s'améliorer en même temps que le volume des données augmente alors qu'un algorithme d'apprentissage automatique s'arrêtera à partir d'un certain niveau de performance.

C'est pour cela que la capacité d'extraire de nouvelles informations à partir du volume de données génomiques, en croissance exponentielle, nécessite des modèles d'apprentissage automatique plus performants. L'apprentissage en profondeur est en train de devenir une méthode de choix pour de nombreuses tâches de modélisation génomique, notamment pour la prévision de l'impact de la variation génétique sur les mécanismes de régulation des gènes [16]. Alakwaa *et al.* en 2018 ont comparé la performance prédictive d'une méthode d'apprentissage profond par rapport à six autres méthodes d'apprentissage supervisé chez 271 patientes atteintes d'un cancer du sein à l'aide de données de métabolomique [17]. La méthode d'apprentissage profond a présenté des avantages pour la classification du statut hormonal du cancer basée sur la métabolomique. Cette méthode a montré une performance supérieure en termes de prédiction comparativement aux six autres méthodes supervisées mais aussi une meilleure pertinence biologique. C'est pour ces raisons que les auteurs préconisent l'utilisation de ce type de méthodes en métabolomique.

I.1.3.1. Apprentissage par renforcement

Dans le cadre de l'apprentissage par renforcement [18], le système peut interagir avec son environnement afin d'accomplir des actions. En retour, il obtient une « récompense », qui peut être positive si l'action était un bon choix, ou négative dans le cas contraire. La « récompense » peut parfois venir après une longue suite d'actions ; c'est le cas par exemple pour un système apprenant à jouer aux échecs. Ainsi l'apprentissage consiste dans ce cas à définir une politique, c'est à dire une stratégie afin d'obtenir systématiquement la meilleure récompense « possible ». Les applications principales de l'apprentissage par renforcement se trouvent dans les jeux d'échecs [19] et la robotique [20]. Récemment une étude publiée par Liu *et al.* discutent du potentiel que pourrait apporter la combinaison d'une méthode d'apprentissage profond avec une méthode d'apprentissage par renforcement afin de permettre un meilleur diagnostic du cancer du poumon [21]. A l'heure actuelle en oncologie, aucune étude utilisant ce type de méthode n'a encore été publiée dans Pubmed.

I.1.3.2. Apprentissage non supervisé

Les analyses non supervisées ont pour but de mettre en évidence, par exemple, des sous-groupes de patients (ou de gènes) aux profils similaires, indépendamment de toute connaissance clinique ou biologique *a priori* [22]. Ce sont des méthodes à visée purement exploratoires qui ne sont basées sur aucun test d'hypothèse. Les analyses non supervisées les plus communément rencontrées sont l'analyse en Composantes Principales (ACP) et les analyses en *clusters* (k-means, classification hiérarchique...).

I.1.3.2.1. Analyse en Composantes Principales

L'ACP [23-25] est la plus connue et la plus ancienne des méthodes d'apprentissage non supervisé. Dans le cas de données omiques, le nombre de données générées peut s'avérer très important, pouvant aller jusqu'à plusieurs milliers. Il est donc primordial de réduire la dimension de ces données qui, de plus, sont souvent corrélées entre elles (multi-colinéarité), pouvant ainsi engendrer une redondance d'information. C'est une méthode factorielle qui permet la réduction de la dimension de l'espace des données. Elle a pour objectif de décrire un ensemble de données par de nouvelles variables en nombre restreint. La réduction de dimensions est réalisée à partir de la construction de nouvelles variables synthétiques obtenues par combinaison linéaire. C'est une méthode simple et intuitive qui permet une représentation graphique de la structure des données en dimensions réduites [26]. L'ACP est souvent utilisée pour l'analyse de données omiques [26-29] ou l'analyse de données cliniques [30]. En 2000, Adam *et al.* ont utilisé cette méthode afin de fournir un nouvel outil pronostic dans le cancer du sein après une mastectomie [28]. L'autre intérêt de l'ACP est qu'elle peut supprimer les effets de lot (ou batch effect en anglais) liés aux conditions expérimentales (temps de passage différents, lots différents, etc….)[31], ce qui est le cas lors des projets omiques où les échantillons peuvent être analysés sur des plateformes et à des moments différents. L'effet batch fait référence à des variations techniques ou des différences non biologiques entre les mesures de plusieurs échantillons. Bien que les effets expérimentaux puissent être réduits par une harmonisation des techniques, il est difficile de les éliminer complètement. Si ce biais systématique n'est pas supprimé, son effet peut masquer des différences biologiques importantes, ce qui, dans le pire des cas, conduit à des inférences et des conclusions erronées. D'un autre côté, l'inconvénient principal des ACP est qu'elles ne permettent pas d'obtenir directement des *clusters*. Pour cela, il est nécessaire d'appliquer un algorithme de clustering (par exemple k-means) sur le sous-ensemble des composantes principales de l'ACP obtenu [32, 33].

I.1.3.2.2. Les méthodes de clustering

Comme nous avons pu le voir, l'ACP permet d'obtenir une réduction de dimension. Mais les méthodes de regroupement (ou de clustering) permettent d'aller plus loin. Différentes approches de regroupement ont été proposées. Fraley et Raftery [34] ont suggéré de diviser les approches de regroupement en deux familles distinctes: les méthodes hiérarchiques et les méthodes de partitionnement. Dans cette partie, nous dressons un panorama des différentes méthodes de clustering que nous avons appliquées durant cette thèse.

– Les méthodes hiérarchiques

Les méthodes de classifications hiérarchiques (Hierarchical Clustering Analysis) [35] sont certainement les plus connues et les plus utilisées pour l'analyse des données génétique car elles sont simples à mettre en œuvre et les résultats peuvent être visualisés graphiquement. L'analyse par regroupement hiérarchique est une méthode qui permet le groupement des variables selon leur similarité. Cette approche n'a besoin d'aucune spécification et ne nécessite pas de fixer *a priori* le nombre de clusters, elle ne se base que sur une mesure de similarité (distance par exemple). La hiérarchie est représentée par un dendogramme (arbre de clusters) dans lequel chaque cluster est emboîté dans un autre. Plusieurs variantes existent [36, 37]: la classification ascendante hiérarchique (CAD) et la classification descendante hiérarchique (CDH). La CAD [38, 39] débute en considérant chaque objet comme un cluster et réduit de façon itérative le nombre de clusters en fusionnant les objets les plus proches. La CDH [40] débute avec un seul cluster regroupant tous les objets et divise les clusters afin que l'hétérogénéité soit la plus réduite possible. Les partitions évoluent au cours des itérations via un critère de qualité. Il existe de multiples critères tel que le critère de Ward [41] qui allie à la fois la dispersion à l'intérieur d'une classe et la dispersion entre les classes. Ces processus se terminent par un critère d'arrêt que l'opérateur définira lui-même. Ces méthodes sont performantes et s'adaptent à tout type de données et pour toute mesure de similarité. En revanche, le manque d'évaluation de la performance au sein des clusters reste le point faible de ces algorithmes.

En génomique, il est courant de représenter sur un même graphique les résultats de classification des tumeurs et des gènes. Les valeurs d'expression sont présentées sous forme matricielle : en ligne les gènes (ordonnés par le « dendrogramme des gènes ») et en colonne les tumeurs (ordonnées par le « dendrogramme des tumeurs »). La cellule correspondant à l'intersection d'une ligne et d'une colonne est colorée selon le niveau d'expression du gène dans la tumeur considérée. Cette représentation, appelée *heatmap* [42], permet non seulement de visualiser les similitudes d'expression entre gènes et entre tumeurs mais également de déterminer les groupes de gènes qui ont influencé la classification des tumeurs. Nous pouvons citer à titre d'exemple la classification moléculaire hiérarchique désormais classique des profils d'expression de tumeurs du sein (figure 3) qui a permis d'identifier cinq-sous-

types de tumeurs aux caractéristiques histologiques distinctes et dont les pronostics se sont révélés différents [43-45].

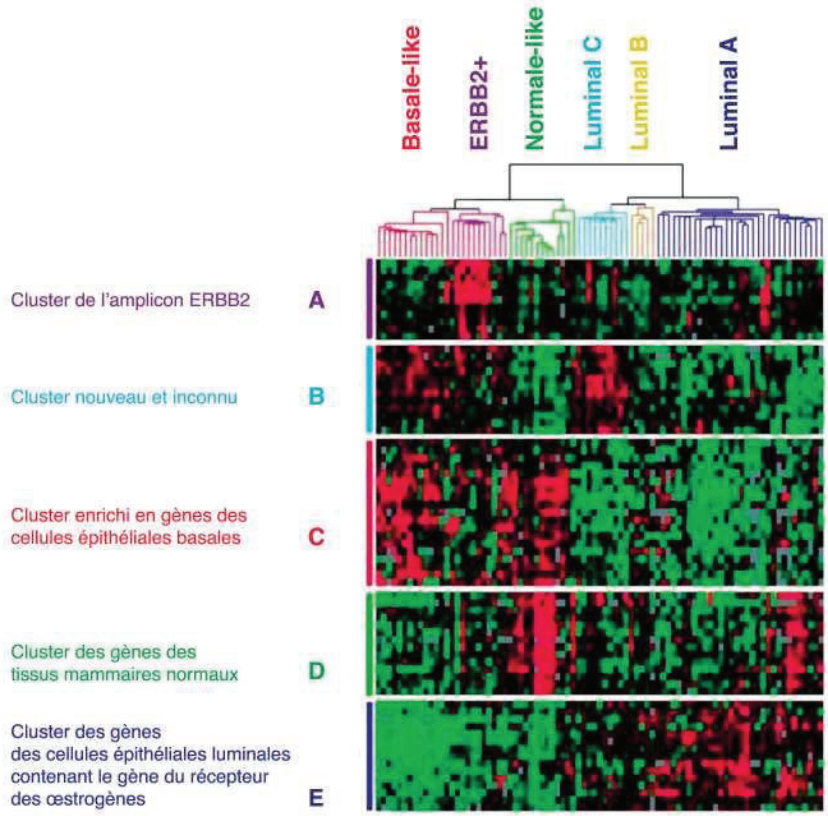


Figure 2 : Classification hiérarchique de 85 tumeurs du sein sur le profil d'expression de 427 gènes [43-45]

– Les méthodes de partitionnement

Comparativement aux méthodes hiérarchiques, cette famille de méthodes propose une partition des données plutôt qu'une structure du type "dendrogramme". Le principe de ces méthodes est alors de comparer plusieurs partitionnements afin de ne sélectionner que le meilleur. Nous présenterons dans cette partie, différentes méthodes de partitionnement dont la plus connue, la méthode des k-means.

– Méthode des K-means

La méthode des *k-means* [46, 47] est une méthode de classification largement utilisée en biologie qui vise à minimiser la somme des distances au carré entre les points d'un même groupe. Chaque observation appartient à un unique groupe. L'algorithme des k-means procède de la façon suivant : 1-Initialisation: l'algorithme choisit aléatoirement k centres initiaux (C_1, …, C_k) parmi l'ensemble des données (que l'on notera X); 2-Itérations: l'algorithme calcule les moyennes empiriques des k centres avec la partition courante et affecte chaque observation au centre dont il est le plus proche de la moyenne; 3-Critère d'arrêt : L'algorithme s'arrête dès qu'il n'y a plus de changement d'affectation.

L'algorithme du k-means est un algorithme dont les performances peuvent varier selon l'étape d'initialisation. Par conséquent, sa convergence vers un minimum global n'est pas assurée.

– Méthode des K-means++

Arthur *et al.*[48] ont proposé une méthode alternative afin de choisir les centres initiaux pour l'algorithme *k-means*. L'étape initiale de l'algorithme est remplacée par 3 étapes : 1- L'algorithme choisit aléatoirement un premier centre C1 parmi l'ensemble des données. 2-L'algorithme identifie un $2^{ème}$ centre C2 associé à l'élément $x \in X$ avec une forte probabilité d'être éloigné de C1. L'algorithme répète l'étape précédente afin d'identifier tous les K centres. Une fois les k centres identifiés, il suffit d'appliquer les étapes 2 à 3 de l'algorithme k-means. Cette initialisation éloigne le prochain centre le plus possible des centres déjà choisis. L'idée essentielle de l'algorithme k-means++ est de choisir les centres un par un de manière contrôlée.

– Le clustering spectral

La classification spectrale (ou clustering spectral) est l'une des méthodes de classification les plus populaires. C'est une méthode issue de la théorie des graphes et de l'analyse numérique. Elle est utilisée pour son efficacité et sa simplicité d'implémentation qui se résume en l'extraction des valeurs et vecteurs propres (spectre) d'une matrice de similarités créée à partir d'un ensemble de données. Contrairement aux algorithmes de classification traditionnelle non supervisée comme celui des K-means, les méthodes de classification spectrale offrent l'avantage de traiter des ensembles de données de structures complexes et non linéairement séparables, comme illustré sur la figure 3. Différents algorithmes de classification spectrale existent [49], ils se décomposent en trois grandes étapes : 1-Construction entre les objets d'un graphe de similarité des données; 2-Représentation spectrale à l'aide d'une projection sur un espace spectral pour que les clusters soient plus facilement identifiables; 3-Application d'un algorithme de clustering de partitionnement (exemple : K-means). De nombreuses études ont prouvé leur efficacité dans le cas de données omiques [50, 51]. Citons par exemple l'étude publiée [52] par Jiang *et al.* où les auteurs ont démontré qu'il était possible d'identifier avec précision des sous-types de cancers en combinant une classification spectrale et des données de génomique.

– Les méthodes à noyaux

De nombreuses applications requièrent des modèles non linéaires pour rendre compte des dépendances et des régularités sous-jacentes dans les données. Les méthodes à noyaux [53, 54] permettent de trouver des fonctions de décision non linéaires, tout en s'appuyant sur des méthodes linéaires (figure 3). Ce concept a suscité un grand intérêt de la part des communautés d'analyse mathématique, de statistique et de Machine Learning [début 20ième siècle Mercer, 1909]. Une fonction noyau correspond à un produit scalaire dans un espace de re-description des données et est

souvent de grande dimension. L'intérêt pour les méthodes à noyaux s'est développé à l'issue de l'introduction des Support Vector Machine (SVM) par Cortes and Vapnik en 1995 [55]. L'algorithme SVM via le concept de *kernel trick*, permet de construire des modèles de classification dont les frontières de séparation complexes permettent de traiter des jeux de données qui ne sont pas linéairement séparables. Ces méthodes simples et rigoureuses sont devenues très populaires pour analyser les données de biologie computationnelle [56] car il est possible de les utiliser pour traiter des problèmes non linéaires.

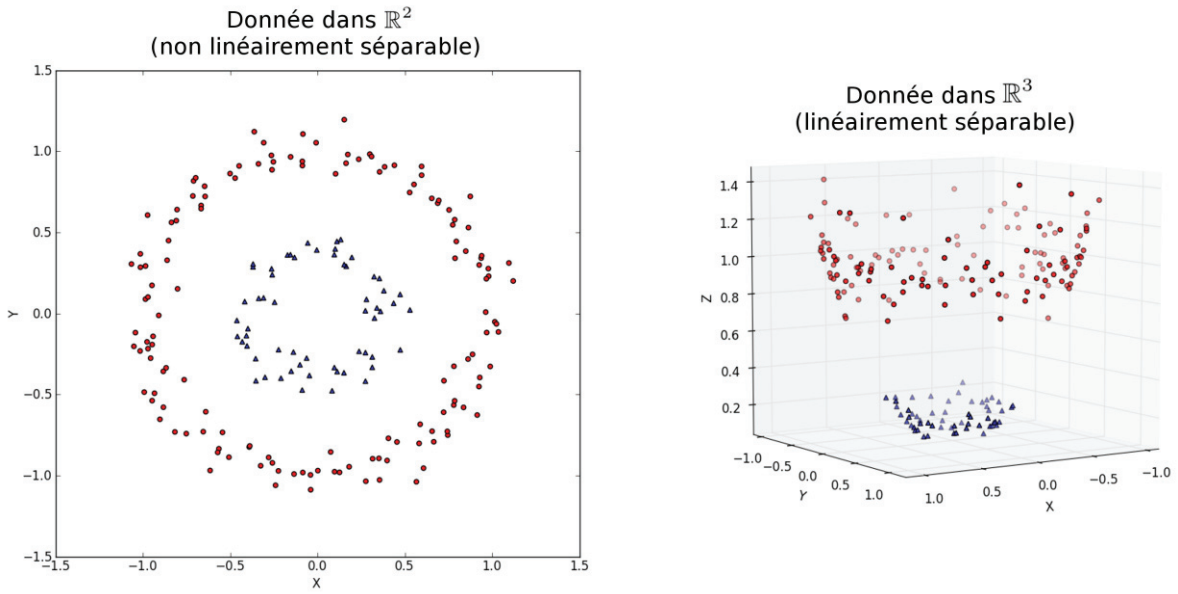


Figure 3 : Passage où les données sont non linéairement séparables vers un espace de description où les données sont linéairement séparables

En outre, afin de prendre en compte la nature hétérogène de certaines données, il est intéressant de combiner plusieurs noyaux, afin d'obtenir un modèle plus souple [57]. L'approche appelée multi-noyaux (*Multiple Kernel Learning*) a ainsi été introduite [58] pour généraliser l'approche mono-noyau. Récemment, la méthode single-cell interpretation via multikernel learning (SIMLR) [59] basée sur une approche multi-noyaux a été développée et appliquée sur un ensemble de données RNA-seq . Cette méthode combine à la fois une approche multi-noyaux, une réduction de dimension et une représentation graphique des données. Ces trois étapes sont représentées en figure 4.

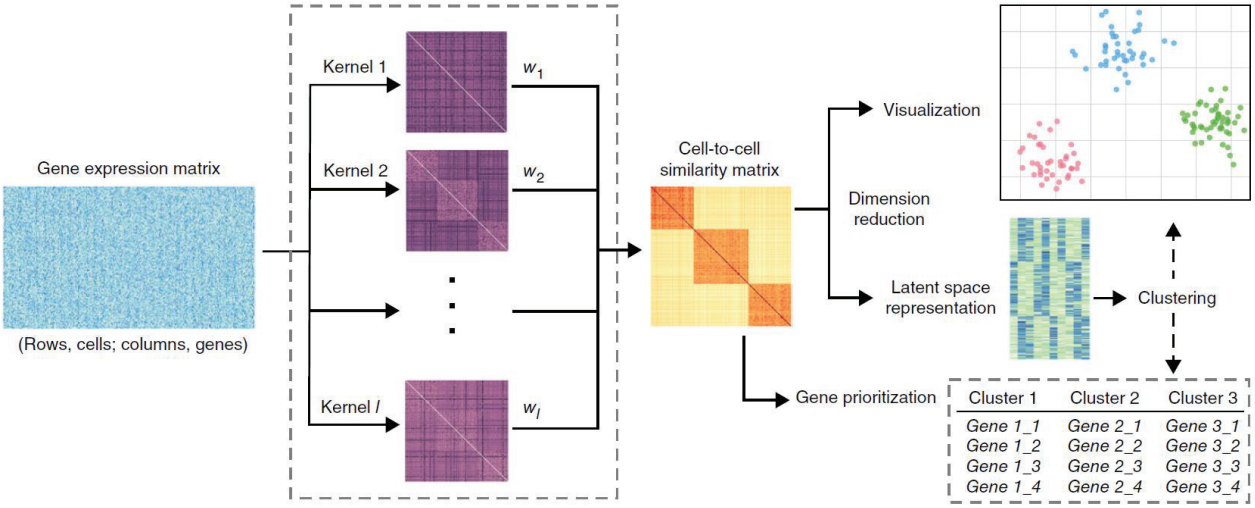


Figure 4 : Visualisation des 3 étapes de la méthode SIMLR appliquée à des données de RNA-seq [59]

– Les méthodes parcimonieuses (sparse)

Les méthodes de classification parcimonieuse, développées plus récemment, permettent de former des groupes de patients en utilisant d'une part les algorithmes de clustering classiques (classification hiérarchique, K-means) et d'autre part, en ajoutant une pénalisation de type Lasso [60] (Least Absolute Shrinkage and Selection Operator) à la fonction objectif pour sélectionner des biomarqueurs. La méthode de clustering sparse k-means [61] est un algorithme amélioré du k-means permettant de partitionner les observations lorsque la base de données contient un grand nombre de données. L'objectif de cette méthode est de ne sélectionner qu'une partie des caractéristiques qui distinguent au mieux les observations en un nombre choisi de k clusters. Cependant, le paramètre utilisé dans la formulation Lasso afin de promouvoir la parcimonie n'est pas simple à estimer. Gilet *et al* en 2017 [62] ont proposé une nouvelle méthode alternative nommée K-sparse. Plutôt qu'une pénalité, les auteurs proposent de définir une contrainte en norme $\ell 1$ et de tirer parti de l'existence d'une projection exacte $\ell 1$ [63]. Cette méthode combine à la fois la méthode de clustering k-means, une réduction de dimension et une sélection de variables. Les auteurs ont comparé leurs résultats avec ceux d'autres méthodes de clustering [59, 61] et ont montré que leur approche améliore significativement les résultats de celles-ci en termes de précision, de silhouette et de temps de calcul.

I.1.3.2.3. Mesure de la performance d'un algorithme de partitionnement

L'évaluation des performances d'un algorithme de clustering n'est pas aussi simple que de compter le nombre d'erreurs ou de mesurer la précision (accuracy) comme il est question avec un algorithme de classification supervisé. Dans ce contexte, l'estimation du nombre optimal de clusters est une question centrale en apprentissage non supervisé. En effet, l'objectif est de partitionner les individus d'un jeu de données, en un nombre restreint de classes les plus homogènes possibles. Les résultats obtenus, par le biais de l'algorithme, dépendent fortement du nombre de classes fixé à l'avance. Il est donc essentiel de choisir un nombre de classes pour que l'attribution des individus au sein d'un cluster soit le meilleur possible. C'est une étape cruciale et les indices utilisés pour mesurer la performance d'un partitionnement peuvent être classés en deux catégories, les indices internes et externes. Les indices de validation interne sont basés sur des informations intrinsèques aux données et évaluent la qualité d'une structure de clustering sans informations externes. Un partitionnement est considéré comme optimal lorsqu'il minimise la moyenne des distances entre les observations au sein d'un cluster (homogénéité) et maximise les distances des clusters 2 à 2 (séparabilité). *L'indice de Davies-Bouldin* [64], de *Calinski-Harabasz* [65], *de silhouette* [66] ou la méthode *Gap statistic* [67] sont des indices fréquemment utilisés. Quant aux mesures de qualité externe, elles sont basées sur une connaissance «*a priori*» des caractéristiques d'un bon clustering. Ces mesures consistent à mesurer le degré de correspondance entre la partition générée par l'algorithme de clustering et une partition connue des données. L'indice Adjusted Rand Index (ARI) [68] et l'information mutuelle normalisée (NMI) [69] font partie des mesures les plus fréquemment utilisées.

I.1.3.3. Apprentissage supervisé

En apprentissage supervisé, la procédure d'analyse comporte habituellement deux étapes. Une phase d'apprentissage qui consiste à estimer un modèle à partir de données puis une phase de validation qui a pour objectif d'évaluer les performances du modèle. Même si elles ne sont pas récentes, les méthodes d'apprentissage supervisé sont de plus en plus utilisées en oncologie comme le montre cette étude publiée en 2019 [70] qui compare différentes méthodes d'apprentissage supervisé afin de prédire le diagnostic préopératoire et le pronostic dans le cadre du cancer de l'ovaire. Parmi l'ensemble des méthodes qui existent en apprentissage supervisé (Support Vector machine, Arbres et forêts aléatoires, Réseaux de neurones, Approches bayésiennes, Régression) [71], nous avons décidé de nous intéresser, dans le cadre de cette thèse, plus particulièrement aux méthodes de régression.

Les méthodes de régression sont très souvent utilisées en pratique en recherche clinique. Cependant, lorsque le nombre de variables explicatives est élevé, voire plus grand que le nombre de patients (fléau

de la dimension), ou lorsqu'il existe une multi-colinéarité entre ces variables explicatives, des problèmes d'estimation apparaissent. Un phénomène de multi-colinéarité peut générer alors des problèmes comme une variance des estimateurs élevée, des coefficients non significatifs dans le modèle même si la qualité globale de celui-ci est bon. Pour s'affranchir de ces deux problèmes, une régularisation du problème est nécessaire et les méthodes de régressions pénalisées constituent une bonne alternative à celles des régressions classiques. L'idée consiste à forcer les solutions à être dans un espace plus petit afin de diminuer la variance des estimateurs. Cet espace plus petit, dit aussi contraint, est obtenu par minimisation du problème initial sous contrainte de norme. La contrainte d'appartenance à l'espace est donnée par une fonction de régularisation pénalisant les solutions ayant de grandes normes.

I.1.3.3.1. Méthodes de régularisation

Dans un contexte de données à haute dimension, c'est-à-dire où le nombre de variables *p* est largement supérieur au nombre de patients *n*, les méthodes de régression classiques ne sont pas adaptées car elles peuvent induire un risque majeur de sur-apprentissage[1] [72]. De plus, le grand nombre de variables et la multi-colinéarité des données ne permettent pas la convergence des modèles. Mais attention, la multi-colinéarité n'est pas directement liée à la dimension des données même si ces deux phénomènes ont plus de chance d'apparaître lorsque l'on augmente le nombre de prédicteurs. Deux solutions existent pour pallier ces problèmes: la première consiste à réaliser une réduction de la dimension de l'espace des données comprenant un nombre de variables indépendantes construites comme des combinaisons linéaires (Partial Least Square, etc…) [73] et la deuxième à réaliser une sélection de variables.

– Méthodes de pénalisation

Les méthodes dites de pénalisation consistent à régulariser un problème de régression en introduisant une pénalité sur le vecteur des coefficients. L'idée principale est de forcer un certain nombre de coefficients à être nuls ou à tendre vers zéro dans le modèle. Ces méthodes permettent de réaliser simultanément une sélection de variables et une estimation des coefficients. Ceux-ci sont estimés en maximisant la vraisemblance partielle pénalisée du modèle de régression utilisé (logistique si critère binaire, etc….) $l(\beta) - \lambda \sum_{j=1}^{p} pen(\beta_j)$. Plusieurs termes de pénalisation ont été proposés, les plus communs étant le Lasso [60] qui pénalise selon la norme $\ell 1$ des coefficients ($pen(\beta_j) = |\beta_j|$), le Ridge [74] qui pénalise selon la norme $\ell 2$ et l'Elastic net [75] qui combine les normes $\ell 1$ et $\ell 2$ ($pen(\beta_j) = \alpha|\beta_j| + (1-\alpha)\beta_j^2$) [75]. La pénalisation Elastic net permet de sélectionner un nombre raisonnable de variables (prédicteurs) tout en offrant la possibilité de conserver des prédicteurs corrélés. Cette méthode de régression semble être plus appropriée pour l'analyse de données omiques. Pour les

[1] Phénomène qui intervient lorsque le modèle s'ajuste trop bien à un jeu de données, diminuant par conséquence sa capacité de généralisation.

méthodes de pénalisation, le choix du paramètre λ est généralement optimisé par validation croisée. De nombreuses études ont utilisé des méthodes parcimonieuses en oncologie afin d'identifier des biomarqueurs. On peut citer par exemple l'outil ENCAPP développé par Das *et al.* en 2015, basé sur une méthode Elactic net, dont l'objectif était d'identifier des biomarqueurs afin de prédire le pronostic dans différents types de cancers [76].

I.1.3.3.2. Création de groupes à risque

Après identification des variables pertinentes et construction du modèle final, un score de risque est calculé sur l'échantillon d'apprentissage (i.e. prédicteur linéaire du modèle). Afin que le clinicien puisse prendre une décision en routine clinique, il est nécessaire qu'il ait à sa disposition un seuil lui permettant de déterminer différents sous-groupes à risque (par exemple, haut risque de toxicité ou faible risque de toxicité). Il existe des techniques comme la méthode des quantiles [77], l'indice de Youden [78] ou maximisation du test du χ^2). Dans le cadre de ce travail, nous nous sommes intéressés à une méthode de discrimination, basé sur l'algorithme de Nelder-Mead [79], dont l'objectif consiste à maximiser l'aire sous la courbe de ROC à partir d'un nombre k de classes fixé à l'avance. Le meilleur découpage constituera celui qui maximise l'aire sous la courbe ROC.

I.1.3.3.3. Validation d'un modèle

– Validation externe

Idéalement pour qu'un modèle soit validé, il faudrait que le modèle construit sur le jeu de données d'apprentissage, le soit sur un jeu de données indépendant. Les capacités discriminantes de la règle de décision sont alors estimées sur ce jeu de données en calculant des indices de performance comme l'AUC, le taux de mauvaises classifications pour un critère binaire, la sensibilité et la spécificité ainsi que les intervalles de confiance associés. Dans le cas où le nombre de patients est limité, l'ensemble des patients (N) de l'étude participent alors à l'apprentissage. Il n'est donc pas possible de disposer d'un jeu de données de validation.

– Validation interne

Dans le cas de figure, où une validation externe est impossible, une solution très couramment utilisée consiste à construire le modèle par «validation croisée» [80]. Le principe de la validation croisée est de partitionner l'ensemble du jeu disponible composé de N patients en *k* sous-groupes de patients. La règle de décision est ensuite construite sur les patients des *k-1* premiers sous-groupes (qui jouent le rôle du jeu d'apprentissage), et les performances sont alors évaluées sur le dernier sous-groupe *k* (qui joue le rôle de jeu de validation). La procédure est réitérée *k* fois de sorte que chaque sous-groupe

joue le rôle de jeu de validation. Les performances sont finalement estimées en moyennant les *k* taux de mauvaises classifications obtenus. Cette méthode de ré-échantillonnage porte le nom de *k-folds cross-validation* [81] et dans le cas *où k=N,* elle est nommée *leave-one-out* [82].

– Critère de performance d'un modèle

L'évaluation de la performance d'un modèle est une étape importante lors de la création d'un modèle de classification, car il est nécessaire de prouver la précision de celui-ci et ainsi d'apporter une aide à la décision en pratique clinique. La forme la plus simple d'évaluation des performances est la précision prédictive, qui donne le pourcentage de patients correctement étiquetés par le modèle. Cela peut être une mesure biaisée, en particulier en oncologie, car il arrive souvent qu'il y ait un déséquilibre important entre les patients présentant l'événement (toxicité, progression) et ceux qui ne le présentent pas. Différentes techniques peuvent être utilisées comme la matrice de confusion ou la courbe ROC [83]. Le choix de la méthode dépend évidemment de la question clinique sous-jacente et nécessite une discussion en amont avec les utilisateurs.

Afin d'évaluer les performances d'un modèle de classification, nous venons de voir qu'une des techniques qui peut être utilisée est la courbe ROC. Son objectif est d'analyser les variations de la sensibilité et de la spécificité, afin de visualiser la performance globale du modèle. L'utilisation des courbes ROC est apparue en médecine dans les années 1970 afin d'améliorer la prise de décision en imagerie médicale [84]. Elles sont désormais présentes dans de nombreux domaines, notamment en recherche clinique où elles caractérisent les capacités d'un test à prédire, par exemple, l'efficacité d'un traitement chez un patient. Une courbe ROC est obtenue à l'aide d'un graphique où l'axe des ordonnées correspond à la « sensibilité » et l'axe des abscisses à « 1 – spécificité ». La performance du modèle dépend donc du meilleur compromis entre ceux deux critères. À partir de cette courbe, l'indice permettant d'évaluer numériquement la performance du modèle correspond à l'aire sous la courbe ROC (AUC, *Area Under the Curve*). Cet indice peut être analysé comme la probabilité du modèle à prédire correctement la réponse ou la toxicité chez un patient. Cet indice varie entre 0.5 (la performance du modèle est liée au hasard) et 1 (la performance du modèle est parfaite).

I.2. Contexte biologique : Les approches omiques

Les approches omiques comprennent principalement (figure 5): la génomique (étude des variations génétiques), l'épigénétique (étude des modifications de l'ADN) la transcriptomique (étude de l'ARN), la protéomique (étude des protéines) et la métabolomique (étude des métabolites produits).

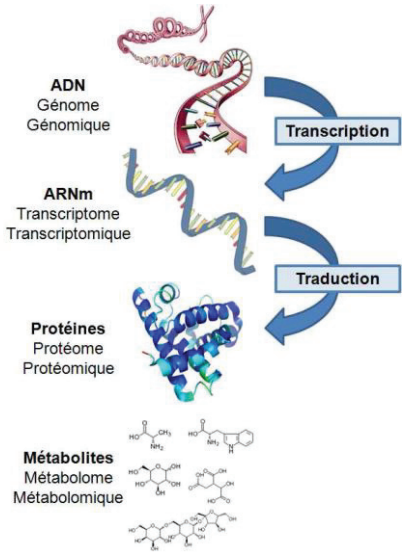


Figure 5 : Représentation des technologies omiques

Le développement des techniques d'analyses génomiques à haut débit permettant ce type de profilage, l'existence de méthodes statistiques historiques telles que l'ACP ou la méthode des K-means [46, 47] et la disponibilité de l'ensemble de ces bases de données, font qu'actuellement de plus en plus d'équipes se sont engagées dans cette voie potentiellement prometteuse. La conséquence est que, depuis quelques années s'est développé le terme de « Big data » qui est un concept reposant sur l'accumulation et l'analyse d'un nombre de données jamais atteint auparavant [85]. L'objectif du « big data » est de permettre d'exploiter et d'analyser simultanément les données issues de ces différentes approches, comme autant d'angles de prise de vue complémentaires décrivant les systèmes biologiques étudiés. La principale difficulté de cette approche réside dans l'interprétation de ces données qui nécessite non seulement des moyens bioinformatiques mais aussi des méthodes statistiques innovantes.

Dans le cadre de ce travail, nous aborderons deux applications avec d'une part une étude génomique et d'autre part une exploration métabolomique.

I.2.1. La génomique

La caractérisation et la compréhension des variations génétiques constituent de véritables challenges en génétique humaine tant pour les personnes indemnes de toute pathologie que pour les personnes malades. Des progrès perceptibles en recherche et thérapeutiques du cancer n'ont été possibles qu'à partir du séquençage de l'ensemble du génome et des avancées technologiques nécessaires à son analyse.

I.2.1.1. Polymorphismes génétiques

Les polymorphismes génétiques sont des différences génétiques entre individus qui sont transmissibles d'une génération à l'autre. Les différents variants observables sont appelés allèles. Les polymorphismes ont des mécanismes de genèse différents mais ont souvent pour origine des erreurs lors de la réplication du génome [86].

I.2.1.1.1. Insertion-délétion

Les insertions-délétions sont des fragments nucléotidiques rajoutés ou retirés par rapport au génome de référence. En général, lorsqu'ils sont présents dans la séquence codante, ils entraînent un décalage de lecture entraînant une traduction différente de l'original. Chez un être humain, on compte environ 200 décalages du cadre de lecture dans son génome [87].

I.2.1.1.2. Single Nucléotide Polymorphismes (SNP)

Les SNPs sont des variations ponctuelles d'un seul nucléotide (figure 6). Leur répartition uniforme sur tout le génome et la simplicité pour les caractériser en font le marqueur d'intérêt chez les chercheurs afin d'établir une cartographie du génome (e.g. dbSNP, HapMap [88-90], le projet 1000 génomes [87]).

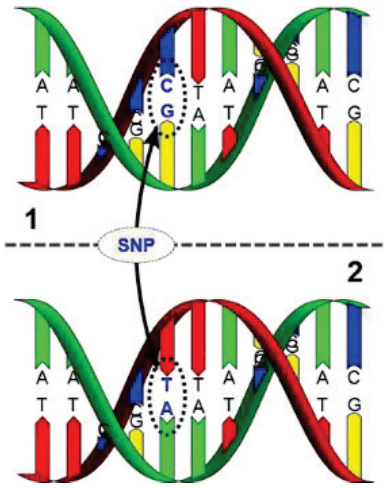


Figure 6 : Représentation d'un polymorphisme entre 2 individus ou un seul nucléotide les diffère.

Le nombre de SNPs répertoriés aujourd'hui est d'environ 40 millions (source dbSNP) et ils représentent plus de 90% de la diversité génétique humaine connue. Un SNP se caractérise par sa position chromosomique, ses allèles et sa fréquence allélique mineure appelée (Minor Allele Frequency ou MAF). Il faut souligner que seule une minorité de SNPs a un impact fonctionnel.

I.2.1.1.3. Modèle d'équilibre d'Hardy-Weinberg

Le modèle d'équilibre d'Hardy [91]-Weinberg [92] est l'un des principes fondamentaux de la génétique des populations. Il modélise le comportement des fréquences alléliques et génotypiques pour un polymorphisme, plus particulièrement les SNPs, au sein d'une population au fil des générations sous différentes conditions. Il stipule, sous certaines hypothèses, que les fréquences alléliques et génotypiques d'un polymorphisme sont stables au sein de la population au fil des générations. Lorsque les hypothèses sont respectées, on dit alors que le modèle est à l'équilibre.

L'équation $p^2 + 2pq + q^2 = 1$ caractérise ce principe où p est la fréquence d'un allèle A, q celle d'un allèle B, p^2 correspond à la fréquence du génotype « AA », 2pq à celle du génotype « AB » et q^2 à celle du génotype « BB ». Les fréquences génotypiques observées permettent grâce à l'équation d'Hardy-Weinberg de calculer les fréquences alléliques attendues puis d'en déduire les fréquences génotypiques attendues qui sont finalement comparées aux fréquences génotypiques observées.

I.2.1.1.4. Le déséquilibre de liaison

En génétique des populations, le déséquilibre de liaison (Linkage disequilibrium ou LD en anglais) correspond à une association non aléatoire d'allèles et à différents locus (position fixe sur un chromosome) d'une population donnée. Les *loci* sont en déséquilibre de liaison lorsque la fréquence d'association de leurs différents allèles est supérieure ou inférieure à celle attendue si les *loci* étaient indépendants et associés de manière aléatoire [93]. Le déséquilibre de liaison est influencé par des facteurs, comme la sélection, le taux de recombinaison génétique et le taux de mutation. La mesure de déséquilibre de liaison la plus utilisée est celle du R^2 [94]. Deux SNPs sont en déséquilibre de liaison lorsque leur LD > 0,80. Lorsque celui-ci est égal à 1, on peut en déduire :

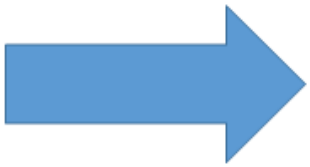
- que les deux SNPs ont les mêmes fréquences alléliques ;
- qu'il n'y a plus que deux conformations haplotypiques observées

I.2.1.1.5. Codages des SNPs pour les analyses statistiques post-traitement

Les SNPs sont en général bialléliques avec deux des quatre bases A, C, T, G. D'un point de vue statistique, un SNP peut être considéré comme une variable catégorielle à trois modalités : si les allèles des SNPs sont, par exemple, A et C, les génotypes possibles sont « AA », « AC » et « CC ». Admettons que C soit l'allèle majeur (c'est à dire l'allèle le plus fréquent) et A l'allèle mineur, alors « CC » est codé 0 (homozygote fréquent), « AC » est codé 1 (hétérozygote) et enfin « AA » est codé 2 (homozygote rare). Le codage 0, 1, 2, est souvent rencontré mais peut-être contesté d'un point de vue statistique et biologique lorsque ces variables sont traitées comme des variables quantitatives, c'est-à-dire numériques, dont les valeurs suivent une échelle discrète ou ordinale [95]. En effet, il n'y a pas de notion d'ordre et/ou de proportionnalité entre les différentes modalités, c'est pourquoi nous avons envisagé un autre codage pour l'ensemble des analyses.

Le principe est de créer une nouvelle variable qualitative à *k* modalités (ici deux ou trois modalités pour un SNP) dans un modèle de régression par k variables binaires Z_i (dummy variable). La variable Z_i vaut 1 si elle est au niveau i, 0 sinon. Nous allons donc considérer un SNP comme étant l'ensemble des deux ou trois variables binaires (figure 7).

PATIENT	SNP rs48256
1	AA
2	AC
3	CC
4	AC
5	AA
6	AA
7	CC
9	AC
10	AA



AA	AC	CC
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0
1	0	0
0	0	1
0	1	0
1	0	0

Figure 7 : Recodage des variables SNPs en variables binaires.

Dans le cadre de cette thèse, nous appliquerons une méthode d'apprentissage supervisé sur des données de SNPs issues d'une plateforme de séquençage haut débit (MassArray; Equipement Agena)

I.2.2. La métabolomique

La métabolomique est une approche analytique biologique émergente. C'est une science multifactorielle qui étudie l'ensemble des métabolites, ou métabolome, présents dans un système biologique donné, par exemple une cellule, un tissu ou un fluide biologique. Les métabolites sont des molécules de taille inférieure à 1500 Dalton. Ils peuvent être classés en deux catégories en fonction de leur origine : les métabolites endogènes qui sont générés par l'organisme et les métabolites exogènes qui proviennent de l'environnement extérieur (exemple: les médicaments ou les polluants environnementaux). La métabolomique donne une vision globale sur les évènements biochimiques présents. Les métabolites, produits par les protéines, prennent part aux réactions biochimiques de la cellule. Ils sont regroupés en deux classes: 1-Les métabolites primaires ; 2-Les métabolites secondaires. Les métabolites primaires sont directement impliqués dans plusieurs processus tels que la reproduction, la croissance ou le développement d'une cellule. Ils sont communs à de nombreux organismes, comme par exemple, les acides aminés. Quant aux métabolites secondaires (xénobiotiques, les toxines ou les médicaments), ils ne sont pas impliqués directement dans ces processus. Les métabolites sont les conséquences finales des processus de régulation cellulaire et leur variation peut être considérée comme la réponse du système biologique envers les changements génétiques, pathologiques, environnementaux et toxicologiques [96]. La métabolomique est probablement l'approche qui est la plus proche du phénotype d'un système biologique, présentant ainsi un intérêt particulier dans divers domaines de recherche (Figure 8) [97].

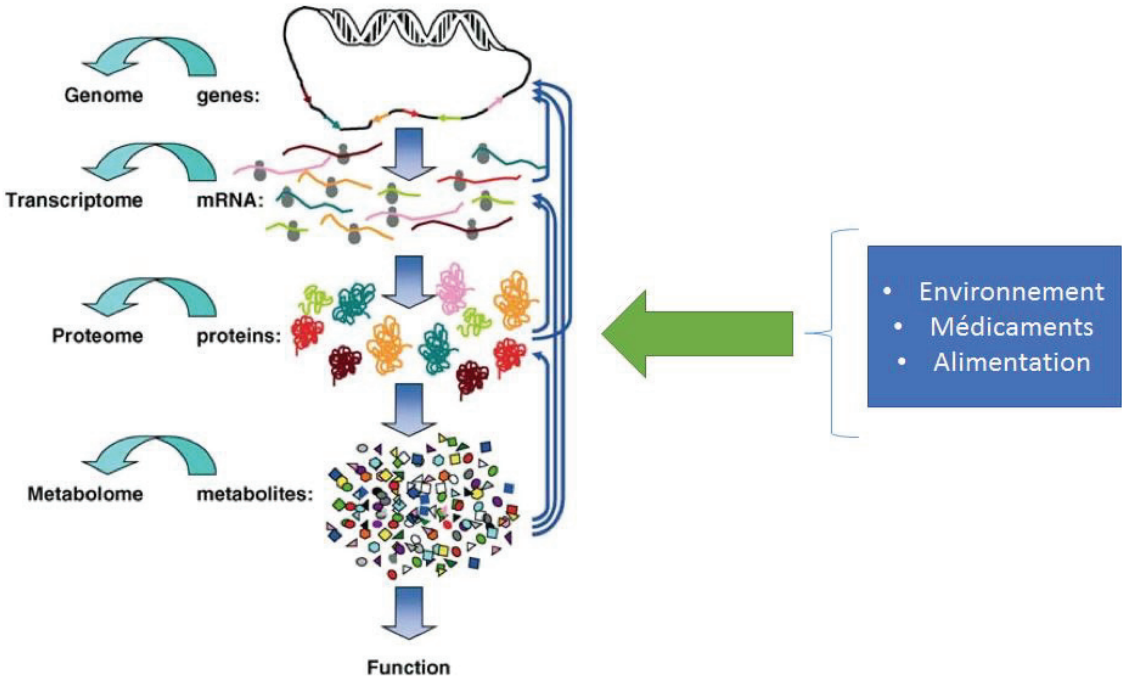


Figure 8 : Interactions entre les différents niveaux du système biologique [97].

I.2.2.1. Les différentes approches métabolomiques

On peut distinguer deux grands types d'approches en métabolomique: 1-L'approche ciblée dont l'objectif est de mesurer spécifiquement un nombre réduit de métabolites comme les substrats ou les produits d'une réaction enzymatique par exemple. Elle peut aussi se focaliser sur des composés appartenant à une voie métabolique ou à une famille chimique donnée. Ce type d'approche a pour but une meilleure compréhension de la fonction ou de la régulation d'une voie métabolique et/ou de ses liens avec les autres voies métaboliques. Elle fait le plus souvent appel à des méthodes d'extraction sélectives de métabolites. Les analyses effectuées dans ce cadre sont souvent quantitatives, précises et sensibles et s'apparentent à des dosages classiques en milieu biologique; 2-L'approche non-ciblée qui mesure un maximum de métabolites présents dans un échantillon sans *a priori*, et qui fournit par conséquent une vue d'ensemble sur l'état métabolique de l'échantillon permettant ainsi la découverte de nouvelles perturbations métaboliques associées à une maladie, un médicament ou un changement environnemental. L'approche non-ciblée présente un intérêt particulier dans la recherche de nouveaux mécanismes ou de nouveaux biomarqueurs. Par contre, la quantification est relative, la sensibilité est plus faible par rapport à une approche ciblée et l'identification des métabolites nécessite une validation supplémentaire. L'information générée est complexe et nécessite un traitement de données approprié.

I.2.2.2. Plateformes analytiques en métabolomique

Les deux plateformes analytiques les plus couramment utilisées pour l'analyse métabolomique sont: 1- La spectroscopie en résonance magnétique nucléaire (RMN) qui est une technique analytique qui exploite les propriétés magnétiques de certains noyaux atomiques, et donne par conséquence des informations structurales sur les métabolites. La RMN est une technique non-destructive, reproductible et quantitative. La préparation des échantillons est simple et une large gamme de métabolites peut être analysée. Le principal inconvénient de la RMN est sa faible sensibilité, qui restreint son application à la mesure des métabolites les plus abondants dans un échantillon [98, 99]. 2-La spectrométrie de masse (SM) est une technique analytique qui sépare les molécules chargées (ions) en fonction de leur rapport masse sur charge (m/z) permettant ainsi la détection et l'identification des métabolites. L'échantillon est injecté soit directement dans le spectromètre de masse, soit dans un système de séparation (par exemple une chromatographie en phase liquide (CL) ou gazeuse (CG)), qui sépare préalablement les métabolites. Les molécules à analyser contenues dans l'échantillon sont ionisées. Il existe plusieurs types de sources d'ionisation dont le choix dépend du type de molécules à analyser. Les molécules chargées sont par la suite séparées par un analyseur en fonction de leur rapport m/z et pourront être fragmentées dans une cellule de collision permettant

l'identification de leur structure chimique [100]. Par rapport à la RMN, la SM est une technique d'une grande sensibilité permettant la détection de variations métaboliques moins abondantes. En revanche, la SM est parfois confrontée à des problèmes de reproductibilité et la quantification reste relative dans le cas de l'approche non-ciblée [101, 102]. La spectroscopie en RMN, la CG-SM et la CL-SM sont les trois techniques les plus utilisées aujourd'hui en métabolomique. Elles présentent chacune des avantages et des inconvénients [103, 104].

I.2.2.3. Prétraitement et normalisation des données

Toutes les approches de génomique fonctionnelle reposent sur l'informatique aussi bien au niveau de l'acquisition et du stockage, que de l'analyse statistique. Si les techniques comme la génomique, la transcriptomique et la protéomique disposent aujourd'hui d'outils de traitement des données performants, de nombreuses limitations subsistent en métabolomique. L'étape de prétraitement des données est une étape cruciale et nécessaire afin de présenter les empreintes métaboliques sous une forme compatible avec la réalisation d'analyses statistiques ultérieures. Elle a pour objectif principal d'extraire les signaux obtenus grâce aux différentes techniques analytiques, d'aligner les empreintes des différents échantillons, d'éliminer le bruit de fond, et de présenter les données extraites sous un format compatible avec leur injection dans les logiciels d'analyses statistiques. De nombreux logiciels existent pour le prétraitement des données métabolomiques [105]. Jusqu'à aujourd'hui, les méthodes de prétraitement des données étaient rarement décrites dans la littérature, mais l'apparition de «normes minimales de restitution des analyses métabolomiques» [106-110] et les nombreux travaux menés sur ce thème ont contribué à améliorer la situation. Suite au prétraitement des données métabolomiques, une matrice de données contenant les métabolites en colonne et les échantillons en ligne est générée. L'étape de normalisation a pour but de supprimer les biais systématiques dans l'intensité des métabolites entre les mesures tout en gardant les variations biologiques intéressantes. La normalisation peut être effectuée entre les différents échantillons ou entre les différents métabolites [111]. Il n'existe pas de méthode standardisée ou unifiée de normalisation en métabolomique non-ciblée. Le choix de la méthode de normalisation doit être adapté à la nature de l'échantillon (urine, sang, tissu, culture cellulaire, etc.) et doit être considéré dès l'étape de la collection et de l'extraction des échantillons [112].

I.2.2.4. Reconstruction des voies métaboliques

Une fois que les métabolites discriminants ont été identifiés, via des méthodes d'apprentissage automatique, ils peuvent être interprétés en termes de voies métaboliques pour comprendre la régulation biologique sous-jacente. L'analyse des voies métaboliques est le plus souvent basée sur une analyse d'enrichissement. La distribution des métabolites appartenant à chaque voie est

comparée entre la liste identifiée à partir de l'échantillon et l'ensemble des métabolites dans une banque de donnée (référence), par exemple KEGG (Figure 9). Si les métabolites appartenant à une certaine voie sont observés plus fréquemment dans le set échantillon par rapport au set de référence, la voie métabolique en question est considérée comme enrichie et suggère ainsi une importance biologique dans la condition expérimentale observée [113, 114]. Plusieurs logiciels permettent de réaliser des analyses d'enrichissement en fournissant pour chaque voie du métabolisme un taux d'enrichissement et une valeur-p qui correspond à la probabilité d'un tel enrichissement. Les deux logiciels les plus utilisés sont MetaboAnalyst [115] (https://www.metaboanalyst.ca/) et Ingenuity (https://www.qiagenbioinformatics.com/) [116].

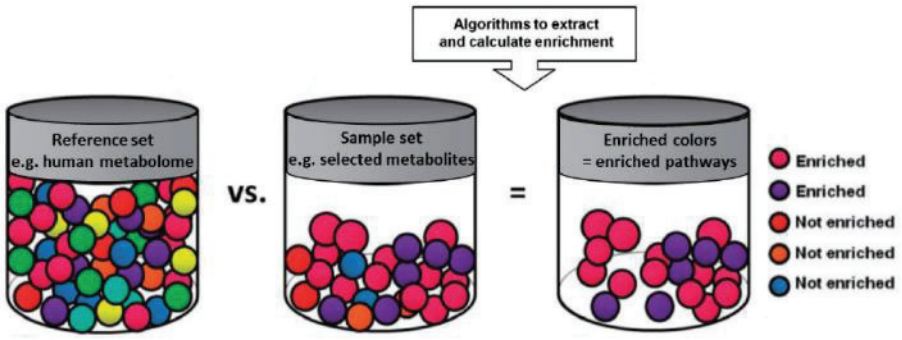


Figure 9 : Principe de l'analyse par enrichissement.([113] modifié)

Dans le cadre de cette thèse, nous appliquerons des méthodes d'apprentissage non supervisé sur des données de métabolomique issues d'un chromatographe en phase liquide couplé à un spectromètre de masse.

I.3. Objectifs des travaux développés

Nous avons présenté, dans les deux premiers chapitres un éventail des méthodes les plus fréquemment utilisées en apprentissage supervisé et non supervisé ainsi que deux approches en sciences omiques qui seront misent en application dans le cadre de cette thèse.

Comme nous avons pu le voir, la haute dimensionnalité des données de type omique pose des problèmes méthodologiques lors de l'analyse statistique. De nombreuses approches ont été développées ces dernières années afin de prendre en compte cette spécificité, mais aucun consensus n'a vraiment été établi concernant le choix de l'une ou l'autre de ces méthodes. Un modèle ne doit pas répondre uniquement aux questions que se pose le biostatisticien puisque sa pertinence statistique ne garantit en rien celle biologique ou clinique. Afin de favoriser la reproductibilité des résultats, il est nécessaire, même s'ils ont des performances statistiques inférieures, que les modèles utilisés aient une pertinence biologique et clinique solide. C'est pour cela que toute étude clinique nécessite une collaboration étroite entre cliniciens, biologistes et biostatisticiens.

La complexité et la quantité d'informations à intégrer lors d'une prise de décision médicale ont largement dépassé les capacités humaines. Ces informations comprennent des variables démographiques, cliniques ou radiologiques mais également des variables biologiques et en particulier omiques (génomique, protéomique, transcriptomique et métabolomique) caractérisées par un grand nombre de variables mesurées relativement au faible nombre de patients. L'analyse de ce type de données représente un véritable challenge dans la mesure où elles peuvent être non informatives (bruitées) et associées à des situations de multi-colinéarité. C'est pour cela qu'il est nécessaire de développer des méthodes capables de prendre en compte ces problèmes. Les méthodes d'apprentissage automatique, branche de l'IA, sont devenues un outil essentiel pour les chercheurs. A partir de jeux de données complexes, ces méthodes permettent d'identifier des modèles pour prédire l'évolution du cancer. L'objectif de ce travail était d'appliquer des méthodes d'apprentissage supervisé et non supervisé à des données biologiques de patients atteints de cancers, dans le but d'améliorer la classification tumorale et d'optimiser la prise en charge thérapeutique. Ce travail a débuté par un premier article, publié dans *Investigational New Drugs*, dont l'objectif était d'évaluer l'apport d'une méthode d'apprentissage supervisé dans l'analyse de données d'immunogénomique germinale (multi-SNPs) pour la prédiction de la réponse et de la tolérance des traitements des cancers par immunothérapie [117]. Dans un deuxième article, soumis pour publication, nous avons cherché à mettre en évidence des sous-groupes de patientes de pronostics différents dans les cancers du sein traitées par chimiothérapie adjuvante, en comparant différentes méthodes d'apprentissage non supervisé appliquées à des données de métabolomique [118]. Enfin, dans un troisième article publié dans *Briefings in Bioinformatics,* nous avons réalisé une revue systématique de la littérature afin de

mieux cerner le sujet complexe des essais thérapeutique simulés (ETS) [85] qui pourraient s'avérer être un outil complémentaire, des méthodes d'IA afin de réduire la taille, la durée et le coût des essais thérapeutiques mais aussi de limiter le nombre de patients inclus à tort. Ces méthodes, fréquemment utilisées en pharmacocinétique (PK) et pharmacodynamique (PD), ont permis par exemple d'optimiser les doses individuelles à administrer tout en réduisant de 27% la durée du syndrome pied-main induite par la Capécitabine et ceci sans compromettre son efficacité anti-tumorale [119]. C'est pour cela que les ETS suscitent depuis plusieurs années un intérêt grandissant car ils ouvrent des perspectives pour une meilleure compréhension de la tolérance et/ou de la réponse à un traitement. L'objectif à terme est de pouvoir réaliser un ETS en oncologie avec les données des patients traités au Centre Antoine Lacassagne. En conclusion, nos travaux permettent de mieux cerner ces outils et d'entrevoir des perspectives d'élargissement de ces applications en IA et biologie clinique du cancer.

Partie II : Travaux réalisés

II. Partie II : Travaux réalisés

II.1. L'immunogénétique germinale, un partenaire potentiel dans l'arsenal des marqueurs prédictifs

II.1.1. Contexte

L'immunogénétique germinale vise à interroger le statut génétique individuel sur une base multifactorielle permettant de prévoir l'efficacité ou la toxicité aux traitements des inhibiteurs de point de contrôle immunitaire (IPCI). Ces traitements apportent des gains notables en termes de survie, mais seulement chez une minorité de patients, et ils peuvent être aussi associés à des événements indésirables graves. Nous avons émis l'hypothèse que la génétique de l'hôte pourrait être utilisée comme biomarqueur prédictif de la réponse aux IPCI et des événements indésirables liés au système immunitaire. Pour cela, nous avons mené une étude basée sur une approche par analyse génétique sur l'ADN germinal ciblant des gènes associés aux réponses immunitaires (163 SNPs). L'objectif de cette application était de prédire, à l'aide d'une méthode d'apprentissage supervisé, l'impact du traitement par IPCI sur le plan de l'efficacité thérapeutique mais aussi sur la survenue d'effets indésirables sur 94 patients traités en monothérapie par IPCI. Les résultats ont mis en évidence l'existence de biomarqueurs germinaux spécifiques au patient, capables de prédire la réponse à aux IPCI et également, de prédire les événements indésirables liés au traitement. Nos résultats montrent que : 1-Le taux de réponse objective (réponse complète ou partielle) était significativement associé aux SNPs liés au microenvironnement immunitaire de la tumeur tels que les gènes CCL2, NOS3, IL1RN, IL12B, CXCR3 et IL6R ; 2-La toxicité était reliée aux SNPs de gènes cibles (directement ou indirectement) du traitement par IPCI et appartenaient tous à la même voie de signalisation, notamment les gènes UNG, IFNW1, CTLA4, PD-L1 et IFNL4. A partir du modèle d'efficacité et de toxicité, nous avons établi deux scores génomiques prédictifs en utilisant les coefficients associés au SNPs identifiés. A partir de ces deux scores, trois niveaux de risques d'échec thérapeutique (faible, modéré, élevé) ainsi que deux niveaux de risques de survenue d'effets indésirables (faible, élevé) ont été créés afin de leur donner une pertinence biologique et clinique. Les patients se trouvant dans le groupe « risque faible» ont 80,5% de chance de répondre au traitement alors que les patients se trouvant dans le groupe « risque élevé » ont 88,5% de chance de ne pas répondre au traitement. De même en termes de tolérance, les patients qui se trouvent dans le groupe « risque faible » ont 4,3% de chance de présenter une toxicité de grade ≥ 3 au traitement alors que les patients qui se trouvent dans le groupe « risque élevé » ont 50% de chance de faire une toxicité de grade ≥ 3. Une analyse univariée de la réponse objective en fonction du score génomique prédictif montre une différence statistiquement significative entre les 3 groupes (OR_{mod}= 4,44 [1,47-13,47], p=0,006; OR_{haut} = 31,62 [IC95%: 7,36-135,91]). Enfin, une analyse

univariée de la toxicité en fonction du score génomique prédictif met en évidence une différence statistiquement significative entre les 2 groupes (OR_{haut} = 22,33 [IC95%: 5,31-93,87], p<0,001). La validation de cet outil prédictif original est en cours sur une série plus importante de patients (essai multicentrique ORL TOPNIVO N°NCT03226756).

II.1.2. Publication: Germinal Immunogenetics predict treatment outcome for PD-1/PD-L1 checkpoint inhibitors

Investigational New Drugs
https://doi.org/10.1007/s10637-019-00845-w

PHASE II STUDIES

Germinal Immunogenetics predict treatment outcome for PD-1/PD-L1 checkpoint inhibitors

Sadal Refae[1] · Jocelyn Gal[2] · Nathalie Ebran[1] · Josiane Otto[3] · Delphine Borchiellini[3] · Frederic Peyrade[3] · Emmanuel Chamorey[2] · Patrick Brest[4] · Gérard Milano[1] · Esma Saada-Bouzid[3]

Received: 10 July 2019 / Accepted: 1 August 2019

Summary
Background Checkpoint inhibitors bring marked benefits but only in a minority of patients and may also be associated with severe adverse events. Treatment outcome still cannot be faithfully predicted. The following study hypothesized that host genetics could be applied as predictive biomarkers for checkpoint inhibitor response and immune-related adverse events. We conducted a study based on germinal polymorphisms from genes coding for proteins involved in immune regulation. *Methods* Germinal DNA was obtained from advanced cancer patients treated with anti-PD-1/PD-L1 checkpoint inhibitors. DNA was genotyped using a custom panel of 166 single nucleotide polymorphisms covering 86 preselected immunogenetic-related genes. Computational analysis using a GTEX portal was made to determine potential expression Quantitative Trait Loci in tissues. *Results* Ninety-four consecutive patients were included. Objective response rate (complete or partial response) was significantly correlated to tumor microenvironment-related SNPs concerning *CCL2*, *NOS3*, *IL1RN*, *IL12B*, *CXCR3* and *IL6R* genes. Toxicity were linked to target-related gene SNPs including *UNG*, *IFNW1*, *CTLA4*, *PD-L1* and *IFNL4* genes. The Area Under the ROC curve (AUC) was 0.81 (95% CI: 0.72–0.9) for response and 0.89 (95% CI: 0.76–1.00) for toxicity. In silico functionality exploring pointed rs4845618 (*IL6R*), rs10964859 (*IFNW1*) and rs3087243 (*CTLA4*) as potentially impacting gene expression. *Conclusion* These results strongly support a role for distinct immunogenetic-related gene SNPs able to predict efficacy and safety of anti-PD1/PD-L1 therapies. The results highlight the existence of patient-specific, germinal biomarkers able predict response to checkpoint inhibitor efficacy and, possibly, to predict treatment-related adverse events.

Keywords Immunotherapy · Predictive test · Precision medicine · Germinal polymorphisms

Introduction

Immunotherapy by so-called checkpoint inhibitors (CPI) has now reached a high level of clinical demonstration in terms of durable antitumor activity and acceptable safety across a spectrum of solid and hematologic malignancies [1, 2]. The fact that clinical response to CPI-based therapies can vary across tumor types and between patients has motivated significant pre-clinical and clinical search to identify biomarkers capable of predicting accurately response and resistance to immunotherapeutic treatment in different tumor types [3]. These biomarkers cover both the tumor itself and the tumor microenvironment. Expression of PD-L1 has been reported to be predictive of response to CPIs targeting PD-1 or PD-L1 in several tumor types [4–6]. As T cells recognize immunogenic antigens, it has been shown that tumor antigenicity, such as tumor mutational burden (TMB) or neoantigen load, could be associated with response to CPIs [7, 8]. Tumor microsatellite status, whether linked or not to TMB, has been identified as a predictor of CPI antitumor efficacy with high microsatellite instability and irrespective of the tumor type [9, 10]. Amultifactorial approach combining these biomarkers was

✉ Gérard Milano
gerard.milano@nice.unicancer.fr

✉ Esma Saada-Bouzid
esma.saada-bouzid@nice.unicancer.fr

[1] Centre Antoine Lacassagne, University Côte d'Azur, Oncopharmacology Unit, F-06189 Nice, France

[2] Centre Antoine Lacassagne, Epidemiology and Biostatistics Department, University Côte d'Azur, F-06189 Nice, France

[3] Centre Antoine Lacassagne, Medical Oncology Department, University Côte d'Azur, F-06189 Nice, France

[4] Centre Antoine Lacassagne, CNRS, Inserm, Ircan, FHU-Oncoage, University Côte d'Azur, F-06189 Nice, France

Published online: 11 August 2019

recently reported [3]. The predictor profile covered both the tumor microenvironment and the tumor characteristics, including a T cell-inflamed gene expression profile (GEP), PD-L1 expression and TMB [3]. Nevertheless, predictive markers are very scarce and even no-existent regarding the risk of side-effects associated with CPI treatment. Although toxicity related to CPI use is relatively infrequent, its severity remains significant with an approximate 1% of treatment-related deaths [11]. Reliable predictors are needed therefore to identify patients at risk.

There is cumulative evidence that pharmacodynamics variability (both response and toxicity) related to conventional anti-cancer therapy including chemotherapy and targeted therapy may be linked to patient characteristics under the general denomination of pharmacogenetics [12, 13]. To date, while most research aimed to predict the clinical efficacy of CPI treatment has focused on tumor immune phenotype and somatic genomic features, there are few reports on how host germline genetics may affect response and toxicity. Consequently, it may be a worthwhile strategy to consider the host in order to identify factors able to predict treatment efficacy as well as treatment-related toxicity. The present study aimed to identify such relationships in a group of patients treated by CPI monotherapy. Patient genotyping was based on extensive SNP analysis covering genes associated with immune reaction in general as well as response to immunotherapy, to autoimmune disease development and in general cancer biology.

Material and methods

Study design and patients

The study covered a cohort of 94 consecutive patients all treated by CPI monotherapy (anti-PD-1 or anti-PD-L1) in the Antoine Lacassagne Center (Nice, France) between July 2012 and January 2018. All data were retrieved from the clinical database of the Antoine Lacassagne Center. Patient tumor responses were defined according to RECIST 1.1 criteria: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). Objective response rate (ORR) was defined as the proportion of patients whose best overall response over the entire treatment period was CR or PR. Immune-related adverse events (irAEs) were evaluated according to National Cancer Institute Common Terminology Criteria for Adverse Events (NCI-CTCAE V5) and were defined as grade ≥ 3. Informed written consent was obtained from each patient. The Institutional Ethics Committee approved the study.

SNP selection and genotyping

An extensive literature search for genes implicated in immune reaction, immunotherapy response, autoimmune diseases and cancer biology was performed through Pubmed (www.pubmed.com) using the keywords "Polymorphism" and "Immunotherapy" and "Cancer" from "January 2000 "to "May 2017". This enabled us to identify and select 166 SNPs of 86 genes. Genotyping was performed using custom-designed sequenom MassArray iPLEX assay (Agena Bioscience) [14]. All SNPs with minor allelic frequency < 5% were excluded from analyses. Furthermore, genes were grouped into different families according to their functionalities, as shown in Table 1.

In silico analysis

The potential functional impact of SNPs was examined using several in silico tools e.g. HaploReg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php). The aim was to collect all potential functional SNPs in the regulatory regions (Linkage disequilibrium and the haplotype blocks within the genes were examined). Regulome DB (http://www.regulomedb.org/) was consulted in order to explore chromatin status, conservation, and regulatory motif alterations within sets of genetically-linked variants. Gtex Portal (https://gtexportal.org/home/) was consulted in order to identify all cis-eQTL SNPs affecting the expression of genes of interest and microRNA binding-site prediction tools. SNPMIR (http://www.genomique.info/joomla_2.5.9/) was taken into account in order to investigate the 3'-UTR and to predict whether a SNP within the target site would disrupt/eliminate or enhance/create a microRNA binding site.

Statistical analysis

An evaluation of missing data rate was performed on 166 SNPs. All SNPs with a missing data rate > 10% were excluded from the analysis. One hundred and sixty-three SNPs were included. Missing genotypes (28% of SNPs) were imputed using multiple imputations by chained equations (MICE) [15]. Dominant and recessive models were investigated to test possible associations between SNPs and treatment outcome and toxicity. For each SNP, risk alleles were coded as 1 and non-risk alleles as 0. For each SNP, odds ratio (OR) and 95% confidence intervals were calculated in a first step by logistic regression for associations between genotypes and treatment outcome and, in a second step, between genotypes and toxicity. Pairwise linkage disequilibrium (LD) analyses were performed [16]. A penalized regression method was used to select biomarkers in high-dimensional data [17, 18]. Toxicity and treatment outcome prediction were thus investigated by

Table 1 Incorporating host immunogenetics for an optimal algorithm in treatment outcome prediction

Immunocheckpoint stimulators	Immunocheckpoint inhibitors	Immunocheckpoint receptor	DNA repair	RNA genes	Chemokines
ICOS	PD-1	LAG3	NIEL2	MIR146a	CXCR3
OX40	PD-L1			MIR-499	CCR2
TIM3	CTLA4			MIR-125A	CCL2
CD137					CCR5
CD28					CCL5
					CCRL2
					CXCR4
					CXCL12
Cytokines	Cell adhesion	Enzyme coding or signal transduction	Angiogenesis	Cell death and apoptosis	Interleukins
IFNL4, 28B	CDH1	IDO1	VEGF-A	TRAIL-R1	IL1B
IFNG		NOS3	VEGFR1	FAS	IL1RN
IRF5		NT5E	VEGFR2	FASL	IL-2
IFNW1		UNG	VEGFR3		IL6R
TNFA		GzmB	PDGFR-α		IL2RA
TNFB		CD3G	HIF1A		IL23R
LTA		JAK1			IL4
		JAK2			IL6
		CD247			IL10
		STAT6			IL10RA
					IL12B
					IL17A
					IL18
Tumor Suppressor gene	Antigen presenting cell	Costimulation T-Helper cell	Other		
TP53	HLA	B7-H4	EGF		
	ITGAM	CD80	CDH1		
	TLR1		TNXB		
	TLR3		AADAT		
	TLR4		KMO		
	P2RX7		ERCC1		
	MYG1		GITR		
	ICAM1		FOXOP3		
	TGFBR1				
	TRAF-3				
	CD40				
	CD103				
	FCGR2A				
	FcGRIIIA				

elastic-net penalized logistic regression [19] with optimal value tuning parameters obtained from 5-fold cross-validation [20]. We stratified by gender when a SNP was located on an X chromosome. A predictive risk score for each patient was constructed. An analysis of the importance of the variables (weight) in the prediction models was performed. The quality of the predictive classification ability model using Area Under the ROC curve (AUC) was established. The AUC varies between 0.5 and 1, values equal to 1 indicate a perfect discrimination power. The optimal number of risk groups for predictive models was obtained using the Nelder-Mead algorithm [21]. Thus, two risk groups of favorable alleles (Low, High) for toxicity and three categories for treatment outcome were defined. Statistical analyses were performed with R.3.5.2 software on Windows® and glmnet [22], pRoc [23], Epi [24] and LDcorSV packages [25].

Results

Clinical and tumor characteristics

Patients, disease and treatment characteristics are summarized in Table 2. Median follow-up was 16.3 months (95% CI: 12.5–18.3). Median age was 68 years (range 32–80); 63 patients were male (67%) and 31 female (33%). A majority of patients were smokers ($n = 66$; 82.5%). Forty-eight had non-small cell lung carcinoma (NSCLC) (51%), 14 renal cell carcinoma (15%), 13 head and neck squamous cell carcinoma (HNSCC) (14%), 12 melanoma (13%) and 7 another cancer type (7%). Among the latter, 2 had bladder cancer, 2 ovarian cancer, 2 hematological cancer and one gastrointestinal cancer. Eighty-two and twenty-two patients, respectively, were treated with anti-PD1 (87%) or anti-PD1 (33%). Sixty-three patients had received previous radiotherapy (68.5%). Fifteen patients experienced grade 3–4 IrAEs (16%), 64 grade 1–2 (68%) and 15 patients showed no IrAE (16%). Among study patients, we identified 89 IrAEs including 49 grade 1 IrAEs, 24 grade 2 IrAEs, 15 grade 3 IrAEs and 1 grade 4 IrAE (renal failure). We recorded 9 (9.5%) patients with CR, 40 (42.5%) with PR, 26 (28%) with SD and 19 (20%) with PD. ORR was observed in 49/94 (52%) of patients.

Associations between treatment outcome and SNPs

A predictive model for treatment response was also elaborated. To that end, the relationship between treatment response and the 163 SNPs and different patient characteristics was first assessed by univariate analysis (Table 3). No clinical characteristics were shown to be correlated to treatment efficacy by univariate analysis. Among the 163 SNPs, 8 SNPs were associated to efficacy in univariate analysis: rs1799983 (*NOS3*); rs4845618 (*IL6R*); rs3212227 (*IL12B*); rs419598 (*IL1RN*); rs2280964 (*CXCR3*); rs13900 (*CCL2*); rs4586 (*CCL2*); rs1024611 (*CCL2*). SNP rs1024611 (*CCL2*) was removed from the multivariate predictive model as it was correlated with rs13900 (*CCL2*; LD = 1). rs13900 (*CCL2*) and rs4586 (*CCL2*) were kept in the model because their LD was equal to 0.52, despite they were on the same gene. Finally, 7 SNPs were kept in the multivariate predictive model. As *rs2280964* (*CXCR3)* is located on the X chromosome, a possible link between gender and *CXCR3* was tested. No interaction was found (not significant). Figure 1a illustrates the performance of the predictive model. The AUC was equal to 0.81 (95% CI: 0.72–0.9) associated with a sensitivity and a specificity of 0.73 and 0.69, respectively. Table 5 shows the respective weight of each SNP in the multivariate predictive model. SNPs *rs13900* C/T (*CCL2*) was the most influential SNP in the model. rs1799983 G/G (*NOS3*), rs321227 T/T (*IL12B*), rs419598 C/T or C/C (*IL1RN*), rs2280964 C/C (*CXCR3*) and *rs4845618* T/T (*IL6R*) had similar weights. *rs4586* (*CCL2*) was the least influential SNP in the model (weight 32).

Table 2 Patients and tumor characteristics

Characteristics	No of patients ($N = 94$)	%
Median Age $_{(min\text{-}max)}$	$68_{32–84}$	
Gender		
Female	31	33
Male	63	67
Histology		
NSCLC*	48	51
Renal cell carcinoma	14	15
HNSCC**	13	14
Melanoma	12	13
Other	7	7
Smoker		
No	14	17.5
Yes	66	82.5
Unknown	14	
Previous irradiation		
No	29	31.5
Yes	63	68.5
Unknown	2	
Number of lines before recurrence		
0	14	15
1	49	52
2	21	22.5
≥3	10	10.5
Anti-PD-1/PD-L1		
Anti-PD1	82	87
Nivolumab	78	83
Prembroluzimab	4	4.25
Anti-PD-L1	12	13
Avelumab	5	5.25
Durvalumab	4	4.25
Atezolizumab	3	3.25
Progress of treatment		
In progress	46	49
Completed	48	51
Reason for stopping treatment		
Progression	32	66.75
Toxicity	9	18.75
Durable response	5	10.5
Patient	2	4
Response		
Complete response	9	9.5
Partial response	40	42.5
Stable disease	26	28.0
Progressive disease	19	20.0
IrAE***		
<3	79	84
≥3	15	16
Type IrAE		
Hematologic	17	19.25
Dermatologic	17	19.25
Thyroid	13	14.5
Digestif	7	8
Metabolic	5	5.5
Articulaire	12	13.5
Rhinitis	5	5.5
Others	13	14.5

* *NSCLC* Non-small cell lung carcinoma, ** *HNSCC* Head and neck squamous cell carcinoma, *** *IrAE* Immune-related adverse event

Three levels of risk of failure (high, intermediate and low) were then created based on the Immunocarta efficacy predictive model (Fig. 2a). Almost half the patients (43.5%, $N = 53$) were in the low-risk group. Patients in the low-risk group had 80.5% chance of response to treatment. Patients in the intermediate risk group had 48% chance of response, while patients in the high-risk group had 11.5% chance of response (Table 6). As compared with patients in the low-risk group, patients at intermediate risk and those at high risk had an OR of 4.42 (95% CI: 1.48–13.37, $p = 0.006$) and 31.62 (95% CI: 7.36–135.91, $p < 0.001$), respectively (Table 6).

Associations between toxicity and SNPs

All tested SNPs were in Hardy-Weinberg equilibrium using χ^2 test. First, a predictive model for toxicity was built. The association between toxicity risk and the 163 SNPs and different patient characteristics was assessed by univariate analysis (Table 4). Toxicity was not correlated with clinical characteristics in the univariate analysis. Seven SNPs out of the 163 analyzed were significantly associated with toxicity: rs246079 (*UNG*), rs10964859 (*IFNW1*), rs4143815 (*PDL-1*), rs12979860 (*IFNL4*), rs3087243 (*CTLA4*), rs11571302 (*CTLA4*) and rs7565213 (*CTLA4*). SNPs rs11571302 (*CTLA4*) and rs7565213 (*CTLA4*) were then removed from the multivariate predictive model as they were strongly correlated (LD = 0.98). LD between other SNPs was inferior to 0.1. As shown in Fig. 1b illustrating the performance of the Immunocarta predictive model for toxicity, AUC was 0.89 (95%CI: 0.76–1) associated with a sensitivity and specificity of 0.80 and 0.85, respectively. The respective weight of each SNP in the multivariate predictive model is presented in Table 5. SNPs rs10964859 G/G (*IFNW1*) and rs246079 A/A (*UNG*) were found to be the two most influential SNPs in the model. rs3087243 G/G (*CTL4*), rs4143815 C/C (*PDL-1*) and rs12979860 T/T (*IFNL4*) had similar weights.

Two risk groups (high and low) were then defined based on a predictive model for toxicity (Fig. 2b). According to this model, patients classified in the low-risk group (74.5%, $n = 70$) had a 4.3% risk of irAE. Conversely, patients classified in the high-risk group (25.5%, $n = 24$) had a 50% risk of irAE (OR: 22.33, CI 95% [5.31–93.87], $p < 0.001$, Table 6).

Discussion

General considerations

The individual immunological profile translated by the platelet-to-lymphocyte ratio has recently been shown to have a significant heritable SNP component [26]. By introducing a potential role played by the host, the present study points that germinal immunogenetics may provide additional information regarding predictive biomarkers for immunotherapy by CPI. Generally, these latter biomarkers are centered on the tumor itself or on its environment. It is clear that germline variants may offer efficient and easily-assessable indicators for enlarging the spectrum of immunotherapy predictive markers. Moreover, targeted SNP identification by genotyping is a particularly cost-effective genomic analysis [27]. It is however important to bear in mind the potential limitations of germinal immunogenetics for accurate predictability in patients receiving CPI-based therapy. These limitations have already been identified through the application of germinal polymorphisms in anticancer treatment in general and in the area of CPI-based therapy in particular. The pharmacogenetics of anticancer agents has largely proven its clinical utility (DPD and fluoropyrimidines, UGT1A1 for irinotecan) [28, 29]. However, this predictive tool suffers from inherent shortcomings, e.g. the usually limited number of cases for which links between pharmacogenetics and pharmacodynamics have been demonstrated, and a lack of independent validations on large cohorts. Prospective controlled trials in which the clinical usefulness of gene polymorphisms is more firmly established are generally rare and have recently been reported for only *DPD* [28] and *UGT1A1* [29]. Penalized regression methods as presently used to select biomarkers in high-dimensional data are well adapted to the general objectives of the study, but these methods do not take into account knowledge of biomarker biological pathways [30]. Thus, another important issue of the present study is the true functional significance of the reported SNPs linked to treatment outcome. This insuffisant information may be explained by the complexity of the exploratory studies needed. Usually, in silico simulations using dedicated software are made and can shed some light on this important aspect of the functional impact of reported predictive SNPs. These important general considerations apply to the present study devoted to potential links between germinal polymorphism and treatment outcome under CPI.

SNPs and CPI efficacy

In the present study, it is interesting to note that the SNPs identified as being related to CPI treatment efficacy were located in genes mostly linked to the tumor microenvironment (*CCL2*, *NOS3*, *IL12B*, *IL1RN*, *CXCR3*, *IL6R*). *NOS3* regulates NO synthesis and thus may play a role in tumor microvascularization [31]. The rs1799983 is located in the coding region of *NOS3* inducing a missense mutation (ASP>GLU) and is associated with a synonymous SNP (rs1549758). As the impact of these two SNPs was not obvious at the protein level missense, we further examined a possible alteration of miRNA binding. Interestingly, a possible gain in miR-24-3p binding capacity was found at the minor allele with a subsequently reduced level of protein translation (SNPMIR website).

Table 4 Univariate analysis for the 163 SNPs according to toxicity

Variable			IrAe <3	IrAe ≥3	OR	CI95%	*p* value
SNPs	Model	Genotype					
rs246079 (UNG)							
	Dominant	A/A	21 (26.6)	12 (80)	1	Referent	
		A/G or G/G	58 (73.4)	3 (20)	0.09	[0.02–0.36]	< 0.001
rs10964859 (IFNW1)							
	Recessive	C/C or C/G	73 (92.4)	10 (66.7)	1	Referent	
		G/G	6 (7.6)	5 (33.3)	6.08	[1.52–24.33]	0.014
rs4143815 (PD-L1)							
	Recessive	G/G or G/C	73 (92.4)	9 (60)	1	Referent	
		C/C	6 (7.6)	6 (40)	8.11	[2.09–31.40]	0.003
rs12979860 (IFNL4)							
	Recessive	C/C or C/T	67 (84.8)	9 (60)	1	Referent	
		T/T	12 (15.2)	6 (40)	3.72	[1.09–12.68]	0.036
rs3087243 (CTLA4)							
	Recessive	A/A or A/G	62 (78.5)	8 (53.3)	1	Referent	
		G/G	17 (21.5)	7 (46.7)	3.19	[1.01–10.27]	0.048
rs11571302 (CTLA4)							
	Recessive	T/T or G/T	65 (82.3)	8 (53.3)	1	Referent	
		G/G	14 (17.7)	7 (46.7)	4.06	[1.23–13.38]	0.018
rs7565213 (CTLA4)							
	Recessive	A/A or A/G	64 (81)	8 (53.3)	1	Referent	
		G/G	15 (19)	7 (46.7)	3.73	[1.14–12.19]	0.026
Patients and tumor characteristics							
Age			65.5 (10)	70.5 (8.4)	1.062	[0.99–1.13]	0.075
Gender							
		Female	25 (31.6)	6 (40)	1	Referent	
		Male	54 (68.4)	9 (60)	0.69	[0.21–2.21]	0.53
Tobacco							
		No	20 (25.3)	4 (26.7)	1	Referent	
		Yes	59 (74.7)	11 (73.3)	0.93	[0.26–3.34]	0.91
Histology							
		HNSCC*	11 (13.9)	2 (13.3)	1	Referent	
		NSCLC**	41 (51.9)	7 (46.7)	0.94	[0.16–5.36]	0.94
		Melanoma	11 (13.9)	1 (6.7)	0.50	[0.03–6.69]	0.59
		Renal cell carcinoma	11 (13.9)	3 (20)	1.50	[0.20–11.25]	0.68
		Other	5 (6.3)	2 (13.3)	2.20	[0.22–21.33]	0.48
Radiotherapy							
		No	23 (29.1)	7 (46.7)	1	Referent	
		Yes	56 (70.9)	8 (53.3)	0.47	[0.14–1.48]	0.18
Number of lines before recurrence		0–1	53 (67.1)	10 (66.7)	1	Referent	0.97
		≥2	26 (32.9)	5 (33.3)	1.01	[0.30–3.37]	

* *HNSCC* Head and neck squamous cell carcinoma, ** *NSCLC* Non-small cell lung carcinoma

CCL2 is a member of the chemokine family that displays chemoattractant activity for immunity cells, and particularly T cells [32]. rs13900 C/T (*CCL2*) was found to be the one with the largest weight in the present multi-SNP model (Table 5). rs13900 is located on the 3'UTR of *CCL2* gene and is in allelic association with more than 20 other SNPs ($r^2 > 0.8$), all located in the *CCL2*, *CCL7*, *CCL11* gene cluster. Furthermore, analysis of this variant effect on possible miRNA binding alteration showed that CCL2 regulation by the hsa-miR-18 family is likely an affinity increase, thus leading to decreased protein expression. This finding concords with previous studies showing that the loss of *CCL2* improved

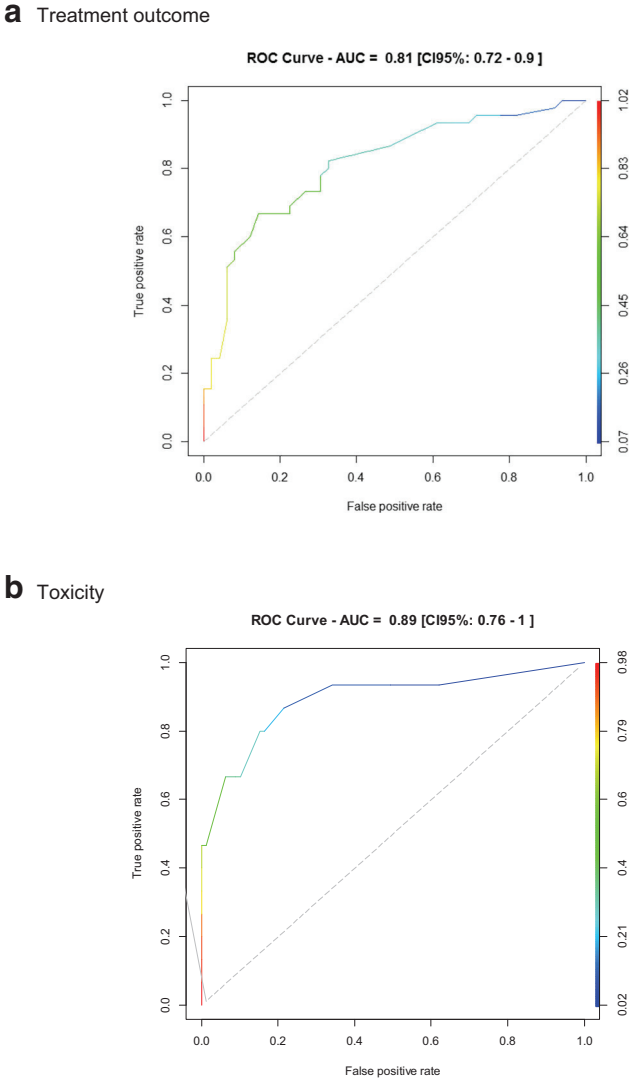


Fig. 1 ROC curves representation and AUC estimation for predictive model

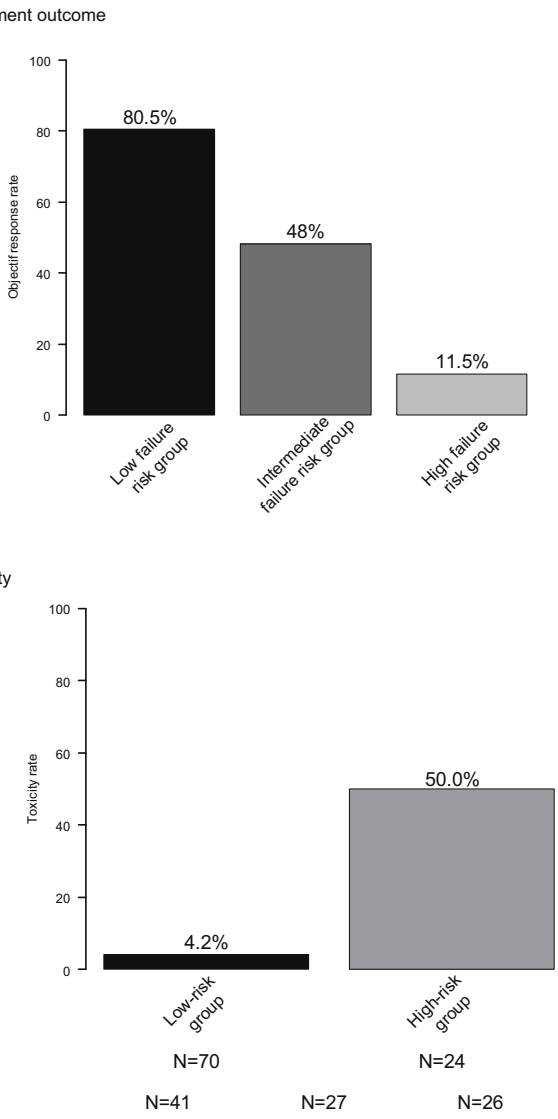


Fig. 2 Patients distribution in the different risk groups according on the predictive model

immunotherapy efficiency [33]. rs4586 is a synonymous SNP located on the coding DNA sequence of *CCL2* gene and is in allelic association ($r^2 > 0.9$) with more than 8 other SNPs ($r^2 > 0.8$), all located in the *CCL2* and *CCL7* gene cluster. Further analysis of this variant effect on possible miRNA alteration of binding showed that *CCL2* regulation by the hsa-miR-15/16 family is likely a capacity fixation increase, thus inducing decreased protein expression.

IL12B, also known as natural killer-cell stimulating factor 2, is a common subunit of interleukin 12 and interleukin 23. This cytokine is expressed by activated macrophages that serve as an essential inducer of Th1 cell biological activity [34]. A *IL12B* gene polymorphism (different, however, from the one reported here) was recently reported to be associated with auto-immune disease [35]. The underlying mechanism leading to autoimmune disease development and those linked with a higher sensitivity to CPI may thus have common biological sources.

IL1RN is a member of the interleukin 1 cytokine family. This protein modulates a variety of *IL1*-related immune inflammatory functions, and *IL1* gene polymorphism has been shown to be related to inflammatory disorders such as bronchopulmonary dysplasia [36]. rs419598 is located on the coding region of IL1RN inducing a synonymous SNP. rs419598 is in allelic association with more than 45 SNPs ($r^2 > 0.9$), thus making it difficult to discuss the possible impact of each individual SNP. This latter cluster of SNPs is associated with eQTL effect in numerous genes (*IL36RN*, *IL1RN*, *IL36A*, *PAX8-AS1*, *CDK8P2*, *TTL*, *IL36B*, and *RP11-65I12.1*) in different tissues. The corresponding effect

Table 3 Univariate analysis for the 163 SNPs according to treatment outcome

Variable			RC or RP	SD or PR	OR	CI95%	*P* value
SNPs	Model	Genotype					
rs1799983 (NOS3)							
	Dominant	G/G	13 (26.5)	22 (48.9)	1	Referent	
		G/T or T/T	36 (73.5)	23 (51.1)	0.37	[0.15–0.91]	**0.026**
rs4845618 (IL6R)							
	Dominant	T/T	14 (28.6)	23 (51.1)	1	Referent	
		G/T or G/G	35 (71.4)	22 (48.9)	0.38	[0.16–0.91]	**0.027**
rs3212227 (IL12B)							
	Dominant	T/T	23 (46.9)	31 (68.9)	1	Referent	
		G/T or G/G	26 (53.1)	14 (31.1)	0.39	[0.16–0.94]	**0.033**
rs419598 (IL1RN)							
	Dominant	T/T	33 (67.3)	18 (40)	1	Referent	
		C/T or C/C	16 (32.7)	27 (60)	3.09	[1.30–7.32]	**0.008**
rs2280964 (CXCR3)							
	Dominant	C/C	33 (67.3)	39 (86.7)	1	Referent	
		C/T or T/T	16 (32.7)	6 (13.3)	0.31	[0.10–0.92]	**0.031**
rs13900 (CCL2)							
	Recessive	C/C or T/T	49 (100)	41 (91.1)	1	Referent	
		C/T	0 (0)	4 (8.9)	–	[− −]	**0.048***
rs4586 (CCL2)							
	Recessive	T/T or C/T	47 (95.9)	37 (82.2)	1	Referent	
		C/C	2 (4.1)	8 (17.8)	5.08	[0.98–26.25]	**0.047**
rs1024611 (CCL2)							
	Recessive	A/A or A/G	49 (100)	41 (91.1)	1	Referent	
		G/G	0 (0)	4 (8.9)	–	[− −]	**0.048***
Patients and tumor characteristics							
Age			65.3 (10.3)	67.4 (9.5)	1.02	[0.98–1.066]	0.31
Gender							
		Female	18 (42.9)	22 (48.9)	1	Referent	
		Male	28 (57.1)	23 (51.1)	1.43	[0.589–3.466]	0.419
Tabac							
		No	12 (24.5)	12 (26.7)	1	Referent	
		Yes	37 (75.5)	33 (73.3)	0.89	[0.346–2.298]	0.809
Histology							
		HNSCC	6 (12.2)	7 (15.6)	1	Referent	
		NSCLC	24 (49)	24 (53.3)	0.85	[0.24–3.00]	0.806
		Melanoma	9 (18.4)	3 (6.7)	0.28	[0.05–1.62]	0.149
		Renal cell carcinoma	6 (12.2)	8 (17.8)	1.14	[0.24–5.38]	0.863
		Other	4 (8.2)	3 (6.7)	0.64	[0.09–4.25]	0.640
Radiotherapy							
		No	18 (36.7)	12 (26.7)	1	Referent	
		Yes	31 (53.1)	33 (73.3	1.59	[0.65–3.92]	0.297
Number of lines before recurrence		0–1	31 (63.3)	32 (71.1)	1	Referent	0.419
		≥2	18 (36.7)	13 (28.9)	0.7	[0.28 - 1.69]	

* OR not calculable replaced by Fisher's test

Table 5 Weight of each variable for multivariate predictive model

rs ID (Gene)	Chr	Ancestral allele	Minor Allele Frequency	Modality	Weight
Treatment outcome					
rs13900 (CCL2)	17	C	0.5	C/T	347
rs1799983 (NOS3)	7	G	0.4	G/G	133
rs3212227 (IL12B)	5	T	0.5	T/T	133
rs419598 (IL1RN)	2	T	0.42	C/T or C/C	133
rs2280964 (CXCR3)	X	C	0.49	C/C	119
rs4845618 (IL6R)	1	G	0.49	T/T	102
rs4586 (CCL2)	17	C	0.46	C/C	32
Toxicity					
rs10964859 (IFNW1)	9	C	0.37	G/G	243
rs246079 (UNG)	12	G	0.41	A/A	243
rs3087243 (CTLA4)	2	G	0.45	G/G	187
rs4143815 (PDL-1)	9	G	0.33	C/C	175
rs12979860 (IFNL4)	19	T	0.32	T/T	152

size was often associated with decreased expression of the corresponding gene in both heterogeneous and homozygous mutant populations.

CXCR3 is a G protein-coupled receptor in the CXC chemokine receptor family expressed primarily on activated T lymphocytes and NK cells. *CXCR3* is able to regulate leukocytes trafficking. *CXCR3* gene polymorphisms have been implicated in various auto-inflammatory diseases [37].

IL6R gene encodes a TL6 receptor complex and plays a significant role in growth and differentiation of immune cells. Dysregulated production of *IL6* and its receptor is implicated in the pathogenesis of many diseases, and with autoimmune diseases in particular [38]. rs4845618 is an intronic variant located on the *IL6R* gene. This SNP is in allelic association with more than 20 other SNPs ($r^2 > 0.8$), all located along *IL6R* and *TDRD10* genes. It is associated with eQTL effect on these two genes and is dependent upon tissue location. Since this SNP is associated with other different SNPs, it is difficult to obtain a clear-cut view of its effect.

Chowell and coworkers recently reported on HLA-1 variations at each of the genes *HLA-A*, *-B* and *-C* as being possibly associated with treatment outcome under CPIs [39]. They examined 1535 patients and found that homozygosity at one HLA-I locus ("A" "B" or "C") was associated with a significant loss in overall survival. HLA polymorphisms were included in the explored cohort of the present study but the investigated polymorphisms did not demonstrate predictive power. The broader coverage of HLA polymorphisms and the higher number of cases in the Chowell study as compared to our own may explain the differing findings.

An approach with a multi-SNP score was adopted by Lima and coworkers [40] in order to establish a predictive profile for BCG immunotherapy in bladder cancer. They noted that polymorphisms impacting interleukins and their receptors could play an independent role within the predictive panel. The present study focused on possible relationships between SNPs and CPI-treatment outcome by analyzing CPI monotherapy treatment. This approach limits the sources of variability which can occur in association studies. Links between SNPs and survival were not considered for one main reason since the profile of the patient population included different tumor locations, each exhibiting its own specific evolution profile.

Table 6 Classification of patients based on risk group and risk evaluation of each group

		Toxicity			
Risk group	Total n (%)	No IrAE	IrAE	Odds Ratio (CI 95%)	p
Low risk	70 (74.5%)	67 (95.7%)	3 (4.3%)	1.0 referent	
High risk	24 (25.5%)	12 (50%)	12 (50%)	22.33 [5.31–93.87]	<0.001
Treatment Outcome					
		RC or PR	SD or PD		
Low risk	41 (43.5%)	33 (80.5%)	8 (19.5%)	1.0 referent	
Intermediate risk	27 (29%)	13 (48%)	14 (52%)	4.44 [1.47–13.37]	0.006
High risk	26 (27.5%)	3 (11.5%)	23 (88.5%)	31.62 [7.36–135.91]	<0.001

SNPs and CPI toxicity

In contrats with SNPs related to treatment efficacy, those linked to treatment toxicity were associated with target cell-related genes, with a final panel including *PD-L1*, *CTLA-4*, *UNG*, *ILNL4* and *IFNW1*. This finding deserves to be emphasized since toxicity can be understood as treatment effect outside the pre-supposed specific tumor target. This observation suggests that there are probably no SNPs common to both responders and toxic patients. Thus, it is unlikely that common SNPs could explain the reported finding that the same patients treated by single agent CPI may show both significant adverse events and greater efficacy [41, 42]. Among the final panel for SNPs linked to CPI adverse events, two SNPs, rs10964859 (*IFNW1)* and rs246079 (*UNG)*, were found to have the highest weight in the model (Table 5). The protein encoded by the *INFW1* gene is interferon Omega1. The protein binds to the interferon alpha/beta receptor but not to the interferon gamma receptor and plays a role in adaptive immune response and T cell activation as well as in NK-cell activation and tumor immune response in general [43]. rs10964859 is located on the 3'UTR of *IFNW1* gene and is not genetically associated with other SNPs. Furthermore, analysis of this variant effect on possible miRNA loss of binding revealed that *IFNW1* expression regulation by hsa-miR-7-1-3p is likely in the presence of this polymorphism. Interestingly, the increase in intrinsic *PD-L1* expression conferred by *INF* pathway activation [44] may explain the increased treatment sensitivity of patients under CPI.

UNG gene product should not be stringently considered as being related to CPI treatment targets [45]. The encoded protein eliminates uracil from DNA molecules. However, *UNG* has been shown to play a significant role in adaptive immunity [46]. It has been demonstrated that *UNG* play a role in immune tolerance mechanisms [47]. rs246079 is present in the last intron of the *UNG* gene and is in allelic association ($r^2 > 0.9$) with more than 10 other SNPs also involving *ACAB* (Acetyl CoA carboxylase gene) and *RP11–98801.5* (human genomic BAC library clone name). Our analysis showed a negative effect size of rs246079 on *ACAB* expression only in blood cells, thus making it difficult to interpret the real impact of the variant on tumor biology.

We also found rs4143815 (*PD-L1*) to be related to adverse events, although with a lesser intrinsic impact in the predictive model. This SNP was recently reported by Nomizo and coworkers to be a possible biomarker for the efficacy of nivolumab [48]. *PD-L1* is expressed not only on the tumor cell surface but also on host immunity cells [49]. Consequently, the dual impact of rs4143815 is not surprising.

rs4143815 is located in the 3'UTR of *CD274* gene and is not in strong genetic association with other SNPs. Interestingly, we found no alteration in miRNA binding generated by this SNP. Rather, it was found to be associated with positive eQTL of the downstream *PDCD1LG2* gene, coding for *PD-L2*, in several tissues (GTEX portal: https://gtexportal.org/home/). This suggests a possible effect of the *PDCD1LG2* gene, an important paralog of this gene being *CD274*, which is involved in the costimulatory signal essential for T cell proliferation and *IFNG* production independently of *PDCD1*. This increase in intrinsic *PD-L2* expression may contribute to excessive sensitivity to CPI action.

The other *rs* identified in the final model for toxicity prediction were related to *CTLA4* and *IL12B*. rs3087243 is located on 3′ of the *CTLA4* gene region. rs3087243 is in allelic association ($r^2 > 0.9$) with rs11571316 and rs11571302 and more than 20 other SNPs ($r^2 > 0.8$), all located in the 5′ or 3′ gene region of *CTLA4*. Altogether, these SNPs have a positive eQTL effect on *CTLA4* expression in testis and lung tissue (GTEX portal), possibly due to improved enhancer activity (http://www.regulomedb.org/). However, it remains to be determined how the increase in *CTLA4* expression in these tissues can account for toxic effects. rs12979860 is located on 5′ gene region of the *IFNL4* gene and is in allelic association with 11 SNPs ($r^2 > 0.8$), all located in the 5′ or 3′ gene region of the *IFNL4* gene. rs12979860 is predicted to alter numerous transcription factors (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) and is associated with increased function of bivalent enhancer and repressed polycomb (http://www.regulomedb.org/). However, no direct eQTL was observed in the GTEX portal.

In addition to the present study, a recently reported investigation by Bins and coworkers assessed the association with CPI toxicity of seven SNPs in four genes [50]. A multivariate analysis in an exploration cohort revealed that homozygous variant patients for *PDCD1* B04C > T had decreased odds for toxicity. However, in a prospective validation group, this link was no longer observed [50].

Conclusion and perspectives

This study identified germinal genetic markers potentially predicting immunotherapy outcome. The data support the role of complementary distinct SNPs in immunogenetics-related genes in predicting efficacy on one hand and safety on the other. These data support the idea that patient-specific, germinal biomarkers may predict response and toxicity to CPI. New molecular technologies and novel analytical methods should provide opportunities to bridge the knowledge gap between SNPs and CPI treatment associations and the functional impact of these SNPs. However, due to the retrospective design of the study and the small number of patients, these results should be validated in a larger cohort and in the context of a prospective clinical trial. Thus, our germinal immunogenetics approach is prosptectively evaluated in TOPNIVO a French multicentric immunotherapy clinical trial (NCT03226756).

Overall, the design of novel computational methods incorporating machine learning and bioinformatic techniques should provide particularly suitable tools for predicting immunosensitivity at individual level and for identifying SNP-related biological mechanisms. These new bases will certainly improve the performance of the germinal immunogenetics approach proposed in the present study.

Acknowledgements The authors acknowledge support from University Côte d'Azur, Centre Antoine Lacassagne, Oncopharmacology Unit, France.

Author contributions All authors have been participated in the writing and involved in critical revision of this manuscript for important intellectual content. All authors approved this manuscript.

Compliance with ethical standards

Conflict of interest Gérard Milano is a member of an advisory board at BMS, MSD and Merck. Fréderic Peyrade is a member of an advisory board at MSD and Merck. Delphine Borchiellini is a member of an advisory board at MSD, Pfizer, Astra-Zeneca, Roche, BMS. The remaining authors declare no competing interests.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent All patients provided written informed consent before enrollment.

References

1. Ribas A, Wolchok JD (2018) Cancer immunotherapy using checkpoint blockade. Science 359(6382):1350–1355
2. Sharma P, Allison JP (2015) The future of immune checkpoint therapy. Science 348(6230):56–61
3. Havel JJ, Chowell D, Chan TA (2019) The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer 19(3):133–150
4. Ansell SM, Lesokhin AM, Borrello I, Halwani A, Scott EC, Gutierrez M, Schuster SJ, Millenson MM, Cattry D, Freeman GJ, Rodig SJ, Chapuy B, Ligon AH, Zhu L, Grosso JF, Kim SY, Timmerman JM, Shipp MA, Armand P (2015) PD-1 blockade with nivolumab in relapsed or refractory Hodgkin's lymphoma. N Engl J Med 372(4):311–319
5. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, Patnaik A, Aggarwal C, Gubens M, Horn L, Carcereny E, Ahn MJ, Felip E, Lee JS, Hellmann MD, Hamid O, Goldman JW, Soria JC, Dolled-Filhart M, Rutledge RZ, Zhang J, Lunceford JK, Rangwala R, Lubiniecki GM, Roach C, Emancipator K, Gandhi L (2015) Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med 372(21):2018–2028
6. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Csoszi T, Fulop A et al (2016) Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung Cancer. N Engl J Med 375(19):1823–1833
7. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J et al (2018) Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science 362(6411)
8. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, Ibrahim F, Bruggeman C, Gasmi B, Zappasodi R, Maeda Y, Sander C, Garon EB, Merghoub T, Wolchok JD, Schumacher TN, Chan TA (2015) Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 348(6230):124–128
9. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD et al (2015) PD-1 blockade in tumors with mismatch-repair deficiency. N Engl J Med 372(26):2509–2520
10. Miao D, Margolis CA, Vokes NI, Liu D, Taylor-Weiner A, Wankowicz SM, Adeegbe D, Keliher D, Schilling B, Tracy A, Manos M, Chau NG, Hanna GJ, Polak P, Rodig SJ, Signoretti S, Sholl LM, Engelman JA, Getz G, Jänne PA, Haddad RI, Choueiri TK, Barbie DA, Haq R, Awad MM, Schadendorf D, Hodi FS, Bellmunt J, Wong KK, Hammerman P, van Allen EM (2018) Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. Nat Genet 50(9):1271–1281
11. Wang DY, Salem JE, Cohen JV, Chandra S, Menzer C, Ye F, Zhao S, Das S, Beckermann KE, Ha L, Rathmell WK, Ancell KK, Balko JM, Bowman C, Davis EJ, Chism DD, Horn L, Long GV, Carlino MS, Lebrun-Vignes B, Eroglu Z, Hassel JC, Menzies AM, Sosman JA, Sullivan RJ, Moslehi JJ, Johnson DB (2018) Fatal toxic effects associated with immune checkpoint inhibitors: a systematic review and meta-analysis. JAMA Oncol 4(12):1721–1728
12. Ciccolini J, Fanciullino R, Serdjebi C, Milano G (2015) Pharmacogenetics and breast cancer management: current status and perspectives. Expert Opin Drug Metab Toxicol 11(5):719–729
13. Hertz DL, McLeod HL (2013) Use of pharmacogenetics for predicting cancer prognosis and treatment exposure, response and toxicity. J Hum Genet 58(6):346–352
14. Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet. 2009;Chapter 2:Unit 2 12
15. White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. Stat Med 30(4): 377–399
16. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity (Edinb) 108(3):285–291
17. Lu M, Zhou J, Naylor C, Kirkpatrick BD, Haque R, Petri WA Jr et al (2017) Application of penalized linear regression methods to the selection of environmental enteropathy biomarkers. Biomark Res 5:9
18. Kim S, Baladandayuthapani V, Lee JJ (2017) Prediction-oriented marker selection (PROMISE): with application to high-dimensional regression. Stat Biosci 9(1):217–245
19. Friedman JH, Hastie T, Tibshirani R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33(1):22
20. Stone M (1974) Cross-Validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Methodol 36(2):111–147
21. Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7(4):308–313
22. Friedman J, Hastie T, Simon N, Tibshirani R. Lasso and Elastic-Net Regularized Generalized Linear Models. R-package version 2.0–5. 2016. 2016
23. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77
24. Carstensen B, Plummer M, Laara E, Hills M. Epi: a package for statistical analysis in epidemiology. R package version 1.1. 71. 2015

25. Desrousseaux D, Sandron F, Siberchicot A, Cierco-Ayrolles C, Mangin B. LDcorSV: Linkage disequilibrium corrected by the structure and the relatedness. R package version 1.3. 1. 2016
26. Lin BD, Carnero-Montoro E, Bell JT, Boomsma DI, de Geus EJ, Jansen R, Kluft C, Mangino M, Penninx B, Spector TD, Willemsen G, Hottenga JJ (2017) 2SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio. J Hum Genet 62(11):979–988
27. Abi A, Safavi A (2019) Targeted detection of single-nucleotide variations: Progress and Promise. ACS Sens 4(4):792–807
28. Henricks LM, Lunenburg C, de Man FM, Meulendijks D, Frederix GWJ, Kienhuis E et al (2018) DPYD genotype-guided dose individualisation of fluoropyrimidine therapy in patients with cancer: a prospective safety analysis. Lancet Oncol 19(11):1459–1467
29. Paez D, Tobena M, Fernandez-Plana J, Sebio A, Virgili AC, Cirera L et al (2019) Pharmacogenetic clinical randomised phase II trial to evaluate the efficacy and safety of FOLFIRI with high-dose irinotecan (HD-FOLFIRI) in metastatic colorectal cancer patients according to their UGT1A 1 genotype. Br J Cancer 120(2):190–195
30. Kitano H (2002) Computational systems biology. Nature 420(6912):206–210
31. Forstermann U, Closs EI, Pollock JS, Nakane M, Schwarz P, Gath I et al (1994) Nitric oxide synthase isozymes. Characterization, purification, molecular cloning, and functions. Hypertension 23(6 Pt 2): 1121–1131
32. Yao M, Brummer G, Acevedo D, Cheng N (2016) Cytokine regulation of metastasis and Tumorigenicity. Adv Cancer Res 132:265–367
33. Fridlender ZG, Buchlis G, Kapoor V, Cheng G, Sun J, Singhal S, Crisanti MC, Wang LCS, Heitjan D, Snyder LA, Albelda SM (2010) CCL2 blockade augments cancer immunotherapy. Cancer Res 70(1):109–118
34. Oppmann B, Lesley R, Blom B, Timans JC, Xu Y, Hunte B, Vega F, Yu N, Wang J, Singh K, Zonin F, Vaisberg E, Churakova T, Liu MR, Gorman D, Wagner J, Zurawski S, Liu YJ, Abrams JS, Moore KW, Rennick D, de Waal-Malefyt R, Hannum C, Bazan JF, Kastelein RA (2000) Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. Immunity 13(5):715–725
35. Loft ND, Skov L, Rasmussen MK, Gniadecki R, Dam TN, Brandslund I, Hoffmann HJ, Andersen MR, Dessau RB, Bergmann AC, Andersen NM, Abildtoft MK, Andersen PS, Hetland ML, Glintborg B, Bank S, Vogel U, Andersen V (2018) Genetic polymorphisms associated with psoriasis and development of psoriatic arthritis in patients with psoriasis. PLoS One 13(2): e0192010
36. Cakmak BC, Calkavur S, Ozkinay F, Koroglu OA, Onay H, Itirli G, Karaca E, Yalaz M, Akisu M, Kultursay N (2012) Association between bronchopulmonary dysplasia and MBL2 and IL1-RN polymorphisms. Pediatr Int 54(6):863–868
37. Qin S, Rottman JB, Myers P, Kassam N, Weinblatt M, Loetscher M, Koch AE, Moser B, Mackay CR (1998) The chemokine receptors CXCR3 and CCR5 mark subsets of T cells associated with certain inflammatory reactions. J Clin Invest 101(4):746–754
38. Wang H, Zhang Z, Chu W, Hale T, Cooper JJ, Elbein SC (2005) Molecular screening and association analyses of the interleukin 6 receptor gene variants with type 2 diabetes, diabetic nephropathy, and insulin sensitivity. J Clin Endocrinol Metab 90(2):1123–1129
39. Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, Kuo F, Kendall SM, Requena D, Riaz N, Greenbaum B, Carroll J, Garon E, Hyman DM, Zehir A, Solit D, Berger M, Zhou R, Rizvi NA, Chan TA (2018) Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. Science. 359(6375):582–587
40. Lima L, Oliveira D, Ferreira JA, Tavares A, Cruz R, Medeiros R, Santos L (2015) The role of functional polymorphisms in immune response genes as biomarkers of bacille Calmette-Guerin (BCG) immunotherapy outcome in bladder cancer: establishment of a predictive profile in a southern Europe population. BJU Int 116(5): 753–763
41. Rogado J, Sanchez-Torres JM, Romero-Laorden N, Ballesteros AI, Pacheco-Barcia V, Ramos-Levi A et al (2019) Immune-related adverse events predict the therapeutic efficacy of anti-PD-1 antibodies in cancer patients. Eur J Cancer 109:21–27
42. Maher VE, Fernandes LL, Weinstock C, Tang S, Agarwal S, Brave M, et al. (2019) Analysis of the Association Between Adverse Events and Outcome in Patients Receiving a Programmed Death Protein 1 or Programmed Death Ligand 1 Antibody. J Clin Oncol, JCO1900318. https://doi.org/10.1200/JCO.19.00318
43. Thomas C, Moraga I, Levin D, Krutzik PO, Podoplelova Y, Trejo A, Lee C, Yarden G, Vleck SE, Glenn JS, Nolan GP, Piehler J, Schreiber G, Garcia KC (2011) Structural linkage between ligand discrimination and receptor activation by type I interferons. Cell 146(4):621–632
44. Castro F, Cardoso AP, Goncalves RM, Serre K, Oliveira MJ (2018) Interferon-gamma at the crossroads of tumor immune surveillance or evasion. Front Immunol 9:847
45. Pearl LH (2000) Structure and function in the uracil-DNA glycosylase superfamily. Mutat Res 460(3–4):165–181
46. Krokan HE, Bjoras M (2013) Base excision repair. Cold Spring Harb Perspect Biol 5(4):a012583
47. Zahn A, Daugan M, Safavi S, Godin D, Cheong C, Lamarre A, di Noia JM (2013) Separation of function between isotype switching and affinity maturation in vivo during acute immune responses and circulating autoantibodies in UNG-deficient mice. J Immunol 190(12):5949–5960
48. Nomizo T, Ozasa H, Tsuji T, Funazo T, Yasuda Y, Yoshida H, Yagi Y, Sakamori Y, Nagai H, Hirai T, Kim YH (2017) Clinical impact of single nucleotide polymorphism in PD-L1 on response to Nivolumab for advanced non-small-cell lung Cancer patients. Sci Rep 7:45124
49. Munn DH (2018) The host protecting the tumor from the host - targeting PDL1 expressed by host cells. J Clin Invest 128(2):570–572
50. Bins S, Basak EA, El Bouazzaoui S, Koolen SLW, Oomen-de Hoop E, van der Leest CH et al (2018) Association between single-nucleotide polymorphisms and adverse events in nivolumab-treated non-small cell lung cancer patients. Br J Cancer 118(10): 1296–1301

II.1.3. Discussion

Cette étude s'inscrit dans le cadre du projet IMMUNOCARTA [120-122], développé par le laboratoire d'Oncopharmacologie (Dr Gérard Milano) du Centre Antoine Lacassagne.

Nos résultats ont mis en évidence sept SNPs associés à un risque de non-réponse au traitement. Ces SNPs sont situés sur les gènes CCL2, NOS3, IL12B, IL1RN, CXCR3 et IL6R. Ces six gènes sont essentiellement liés au microenvironnement tumoral. CCL2, situé sur le chromosome 17, est une cytokine de type chimiokine. Les chimiokines sont une famille de protéines impliquées dans les processus immuno-régulateurs et inflammatoires. Sécrété notamment par les macrophages, CCL2 joue le rôle de chemo-attracteur sur les monocytes, les lymphocytes et les basophiles permettant le recrutement et la migration de ces cellules vers le tissu tumoral. NOS3, situé sur le chromosome 7, code pour une enzyme de type oxyde nitrique synthase qui joue un rôle de médiateur biologique dans les activités anti-tumorales. Plus précisément, NOS3 permet la production de NO. IL12B code pour une sous-unité commune aux interleukines 12 et 23. IL12 et IL23 sont des cytokines impliquées dans la différenciation des lymphocytes T en lymphocytes TH1 et du maintien de leur fonction. IL1RN est un gène codant pour la protéine IL-1RA qui appartient à la famille des cytokines de l'interleukine 1. Par son action inhibitrice des activités de IL1A et IL1B, IL-1RA est responsable de la réponse inflammatoire immunitaire induite par l'interleukine 1. CXCR3, situé sur le chromosome X, est un gène qui code pour un récepteur membranaire couplé à la protéine G. Ce récepteur exprimé par les lymphocytes T activés et les cellules NK contribue à réguler le trafic des leucocytes. IL6R est un gène qui code pour une sous-unité du complexe récepteur de l'interleukine 6 (IL6), IL6 étant une cytokine qui régule la croissance et la différenciation des cellules immunitaires.

Les cinq SNPs mis en évidence et liés à la toxicité du traitement de l'IPC sont situés sur des gènes plutôt liés aux cibles thérapeutiques mises en jeu telles que, outre les cibles des anticorps thérapeutiques anti PD -L1 et CTLA4, UNG, IFNW1 et IFNL4. Le gène UNG code pour une uracile-ADN glycosylase. L'une des fonctions importantes de l'uracile-ADN glycosylase est de freiner la mutagenèse. La protéine codée par le gène IFNW1 est un interféron qui aide à réguler positivement l'activité du système immunitaire. Enfin, le gène IFNL4 produit les interférons en réponse à une infection virale et bloquent la réplication et la propagation virales vers des cellules non infectées et en renforcent les gènes antiviraux. De plus, les interférons propagent un signal de régulation de PD-L1 au niveau de la cellule tumorale [123].

Avec 163 SNPs étudié sur 94 patients, nous avons été confrontés à des données dites de grande dimension [124, 125], c'est-à-dire que le nombsre p de variables est supérieur au nombre n de patients ($p >> n$). Dans un tel contexte, les méthodes classiques de sélection de variables ou de modèles pas à pas ne sont pas recommandées [72]. A cela s'ajoute un souci de multi-colinéarité entre les variables [126]. Aussi, nous aurions pu utiliser des méthodes de régularisation classiques comme la régression

sur composantes principales (PCR) ou la régression par les moindres carrés partiels (PLS) qui fournissent des estimateurs efficaces. On présente souvent comme un avantage le fait de conserver toutes les variables. Or, avec 163 SNPs, cette propriété devient un inconvénient car de telles combinaisons auraient été tout simplement ininterprétables d'un point de vue biologique.

C'est pour cela que nous avons utilisé des méthodes de sélection parcimonieuses (sparse) car elles permettent, par définition, une sélection des variables optimales. Les méthodes Lasso [60] et Elastic net [75] effectuent simultanément régularisation et sélection grâce à des pénalisations. Pour cette analyse, nous avons décidé d'utiliser la méthode de régression pénalisée Elastic net car elle combine les points forts de la méthode Lasso et de la méthode Ridge, tout en minimisant leurs inconvénients.

Ces méthodes comme les méthodes de régression « classique » reposent sur la même hypothèse, elles n'acceptent pas de données manquantes ni sur la variable à expliquer ni sur les variables explicatives. Notre base de données présentait 28% de données manquantes parmi les 163 SNPs. L'analyse n'aurait pu se faire que sur 54 patients. Afin de pallier ce problème, nous avons dû réaliser des imputations multiples [127]. Les résultats obtenus lors de l'analyse multivariée ont montré qu'aucun des 163 SNPs n'avait été sélectionné dans le modèle final. Nous avons donc cherché à en comprendre les causes. Les deux raisons potentielles nous semblent être les suivantes : 1-Une non pré-sélection des SNPs; 2-Une multi-colinéarité des SNPs. Nous avons donc réalisé une phase de pré-sélection des SNPs pour s'affranchir de ce problème. Celle-ci a consisté à : 1-Evaluer dans un premier temps, les associations entre les génotypes et la réponse (et la toxicité) sous la forme d'analyses univariées où le seuil de significativité était fixé à 5% ; 2-Eliminer dans un deuxième temps les SNPs dont le déséquilibre de liaison (LD) était supérieur à 0.8. Cette démarche nous a permis de sélectionner 5 SNPs pour la toxicité et 7 SNPs pour la réponse. A la suite de cette phase, une nouvelle analyse, toujours basée sur une méthode de régression pénalisée Elastic net, a été réalisée et cette fois ci, l'ensemble des SNPs pré-sélectionnés (5 SNPs pour la toxicité et 7 SNPs pour la réponse) ont été inclus dans le modèle final. Il faut noter qu'avec un autre seuil de significativité [128-130], le nombre de SNPs pré-sélectionné aurait été supérieur, modifiant peut-être les résultats finaux, mais générant probablement un phénomène de sur-apprentissage. Il est à signaler qu'une analyse de sensibilité à l'aide d'une régression logistique backward a permis d'obtenir les même résultats en termes d'efficacité et de toxicité. Comme nous venons de le constater, la stratégie d'analyse considérée ne prend pas en compte la connaissance des fonctions biologiques des biomarqueurs. Nous avons donc cherché par la suite à savoir s'il existait des méthodes permettant de nous affranchir de cette phase de pré-sélection. Simon *et al.* en 2013, Liquet B *et al.* en 2015 et Samarov *et al.* en 2017 ont développé trois régressions pénalisées de type sparse group Lasso [131], sparse group PLS [132] ou sparse group Elastic net [133]. Ces trois régressions pénalisées sont des extensions des méthodes Lasso, PLS et Elastic net permettant de sélectionner simultanément des variables au sein d'un groupe. Dans le sparse group Lasso ou sparse group Elastic

net, les variables sont regroupées dans des ensembles l'idée étant d'activer ou de désactiver toutes les variables d'un groupe simultanément. Ces algorithmes sont exploités dans le cadre où un groupe d'individus peut avoir des attributs communs qu'il ne partage pas avec d'autres groupes [134-137].

Pour finir, nous avons créé un score génomique prédictif que nous avons divisé en trois catégories pertinentes pour la réponse objective et deux catégories pour la toxicité. Avec une AUC de 0,81 (95% CI: 0,72– 0,9) associée à une sensibilité de 80,5% et une spécificité de 88,5% pour la réponse et une AUC de 0,81 (95% CI: 0,89 - 1,00) associée à une sensibilité de 50% et une spécificité de 95,8% pour la toxicité, nos résultats peuvent être jugés comme satisfaisants. Cependant, notre modèle présente une sensibilité faible, très certainement due aux fréquences relativement faibles des toxicités de grade ≥ 3 (16%).

Dans la continuité de cette étude et sur une série de patients plus importante, nous envisageons de comparer nos résultats avec ceux obtenus en utilisant les méthodes sparse group Lasso, sparse group PLS et sparse group Elastic net.

II.2. Comparaison de méthodes de machine learning non supervisé pour identifier des signatures métabolomiques chez des patientes atteintes d'un cancer du sein localisé

II.2.1. Contexte

La génomique et la transcriptomique ont permis une amélioration des connaissances en biologie moléculaire et une meilleure compréhension de l'oncogénèse. Une taxonomie du cancer du sein a été établie en fonction des caractéristiques génétiques tumorales, permettant une individualisation du pronostic, de la prise de décision thérapeutique et de la prédiction de la réponse au traitement. Cependant, des comportements biologiques hétérogènes persistent. La métabolomique a la particularité d'intégrer l'impact de l'environnement sur la biologie cellulaire. Or l'environnement de la cellule cancéreuse joue un rôle majeur dans le processus d'oncogenèse et dans l'expression du phénotype tumoral. La métabolomique pourrait se révéler être une approche complémentaire de la génomique pour permettre une meilleure compréhension de l'influence du milieu sur le phénotype tumoral exprimé. L'objectif de cette étude a été de chercher à mettre en évidence des sous-groupes de patientes de pronostics différents dans les cancers du sein traités par chimiothérapie adjuvante, en comparant différentes méthodes d'apprentissage non supervisé appliquées à des données de métabolomique. Les résultats ont mis en évidence l'existence de trois groupes de pronostic de patientes (bon, intermédiaire et mauvais) avec des profils cliniques et biologiques différents. Nos résultats ont montré qu'il existait une dysrégulation de trois voies métaboliques (glycolyse ; glutaminolyse ; acides aminés) entre les sous-types des tumeurs du sein. Parmi les cinq méthodes envisagées dans cette étude, K-sparse et SIMLR ont été les méthodes les plus performantes en termes de clustering. Ce travail a été soumis pour publication à la revue *Briefings in Bioinformatics*. Il plaide en faveur de la poursuite de recherches associant apprentissage non supervisé et données de métabolomique dans le but d'améliorer les classifications des tumeurs du sein.

II.2.2. Publication: Comparison of unsupervised machine learning methods to identify metabolomic signatures in patients with localized breast cancer

OXFORD UNIVERSITY PRESS | Briefings in Bioinformatics

Comparison of unsupervised machine learning methods in order to identified metabolomic signatures in patients with localized breast cancer

Journal:	*Briefings in Bioinformatics*
Manuscript ID	Draft
Manuscript Type:	Case Study
Date Submitted by the Author:	n/a
Complete List of Authors:	GAL, Jocelyn; Centre Antoine-Lacassagne, Bailleux, Caroline; Centre Antoine-Lacassagne, Medical Oncology Department Chardin, David; Centre Antoine-Lacassagne, Nuclear Medicine Department Pourcher, Thierry; University of Nice Sophia Antipolis Faculty of Medicine, Laboratory Transporters in Imaging and Radiotherapy in Oncology Gilhodes, Julia; Institut Universitaire du Cancer Toulouse Oncopole Jing, Lun; University of Nice Sophia Antipolis Faculty of Medicine, Laboratory Transporters in Imaging and Radiotherapy in Oncology Guigonis, Jean-Marie; University of Nice Sophia Antipolis Faculty of Medicine, Laboratory Transporters in Imaging and Radiotherapy in Oncology Ferrero, Jean-Marc; Centre Antoine-Lacassagne Milano, Gérard; Centre Antoine-Lacassagne Mograbi, Baharia; University of Nice Sophia Antipolis, CNRS, Inserm, Ircan, FHU-Oncoage Brest, Patrick; University of Nice Sophia Antipolis, CNRS, Inserm, Ircan, FHU-Oncoage Chateau, Yann; Centre Antoine-Lacassagne Humbert, Olivier; Centre Antoine-Lacassagne, Nuclear Medicine Department Chamorey, Emmanuel; Centre Antoine-Lacassagne
Keywords:	Unsupervised machine learning, Metabolomics, Breast neoplasms, Computer simulation

Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer

Jocelyn Gal[1*‖], Caroline Bailleux[2‖], David Chardin[3,4‖], Thierry Pourcher[4], Julia Gilhodes[5], Lun Jing[4], Jean-Marie Guigonis[4], Jean-Marc Ferrero[2], Gerard Milano[6], Baharia Mograbi[7], Patrick Brest[7], Yann Chateau[1], Olivier Humbert[3,4], Emmanuel Chamorey[1].

[1] University Côte d'Azur, Epidemiology and Biostatistics Department, Centre Antoine Lacassagne, Nice, F-06189, France
[2] University Côte d'Azur, Medical Oncology Department Centre Antoine Lacassagne, Nice, F-06189, France
[3] University Côte d'Azur, Nuclear Medicine Department, Centre Antoine Lacassagne, Nice, F-06189, France
[4] University Côte d'Azur, Commissariat à l'Energie Atomique, Institut de Biosciences et Biotechnologies d'Aix-Marseille, Laboratory Transporters in Imaging and Radiotherapy in Oncology, Faculty of Medicine, Nice, F-06100, France
[5] Department of Biostatistics, Institut Claudius Regaud, IUCT-O Toulouse, France.
[6] University Côte d'Azur, Centre Antoine Lacassagne, Oncopharmacology Unit, Nice, F-06189, France
[7] University Côte d'Azur, CNRS UMR7284, INSERM U1081, IRCAN TEAM4; Centre Antoine Lacassagne FHU-Oncoage, Nice, F-06189, France

[*] **Corresponding author:**
Mr Jocelyn Gal
Department of Epidemiology and Biostatistics, Centre Antoine Lacassagne, University Côte d'Azur
33 avenue de Valombrose
06189 Nice, France
Phone : +33-4-92-03-10-31
E-mail : jocelyn.gal@nice.unicancer.fr

Key words: Unsupervised machine learning, Metabolomics, Breast neoplasms, Computer simulation

[‖] *These authors contributed equally to this work*

Jocelyn Gal is a biostatistician in the Epidemiology and Biostatistics Unit at the Antoine Lacassagne Center, Nice, France. He is currently pursuing his PhD in the Laboratory of the Pharmacological Targeting Unit in Oncology, Faculty of Sciences, Université Côte d'Azur, Nice, France. His research interests: bioinformatics, methodology in clinical research, systems biology, data mining and machine learning.

Caroline Bailleux is a medical oncologist specialized in breast cancer. A former student at the Ecole Normale Supérieure - Paris-Saclay in the biology department. A PhD student in metabolomics in breast cancer.

David Chardin is a nuclear medicine physician currently preparing a PHD. His fields of interest include metabolomics and tumor metabolism, Radiomics and Artificial Intelligence.

Thierry Pourcher is the Director of Research heading the laboratory Transporters in Imaging and Radiotherapy in Oncology and the Bernard Rossi Mass Spectrometry Facility for Proteomics at the School of Medicine of Nice. He is skilled in the analysis of metabolomics data.

Julia Gilhodes is currently a PHD student in the biostatistics unit of the Claudius Regaud Institute in Toulouse. She is involved in methodological research and statistical support for translational research and is especially interested in the analysis of omics data and in statistical considerations regarding molecular signatures.

Lun Jing was a PhD student in the laboratory Transporters in Imaging and Radiotherapy in Oncology. She is skilled in metabolomics.

Jean-Marie Guigonis currently works at the Faculty of Medicine of Nice. He heads the Bernard Rossi Proteomics and Metabolomics platform. His current project concerns clinical applications in Mass Spectrometry.

Jean-Marc Ferrero is Professor in Medical Oncology in charge of the Clinical Research Unit at the Antoine Lacassagne Centre, Nice, France. His research interests are: precision medicine in breast cancer and clinical research.

Gerard Milano is head of the Pharmacological Targeting Unit in Oncology at the Antoine Lacassagne Center in Nice, France. He is working on the personalized approach to medicine designed to manage cancer treatment according to specific individual characteristics of the patient's tumor.

Baharia Mograbi has made different contributions to basic and translational research through her work in cell biology, inflammation and cancer research. She is an expert for the French Ministry of Research (ANR, AERES)

Patrick Brest is an INSERM researcher working on RNA regulation in cancer. He has a strong background in genetic disease PB and is a specialist in OMICS analysis.

Yann Chateau is Data Manager in the Epidemiology and Biostatistics Unit at the Antoine Lacassagne Centre in Nice, France. His research interests are: bioinformatics, text mining and clinical research.

Olivier Humbert is an associate professor in Nuclear Medicine and Biophysics, working in the Antoine Lacassagne Centre, Nice, France and the Université Côte d'Azur, Nice, France. His research interests are: cancer imaging, radiomics and metabolomics. He holds a chair of artificial intelligence in medicine at the 3IA Côte d'Azur.

Emmanuel Chamorey is a pharmacist and head of the Epidemiology and Biostatistics Unit at the Antoine Lacassagne Center in Nice, France. His research interests are: onco-pharmacology, methodology in clinical research and statistical analysis.

Abstract

Genomics and transcriptomics have led to the widely-used molecular classification of breast cancer (BC). However, heterogeneous biological behaviors persist within breast cancer subtypes. Metabolomics is a rapidly-expanding field of study dedicated to cellular metabolisms impacted by the environment. The aim of this study was to compare metabolomic signatures of BC obtained by 5 different unsupervised machine learning (ML) methods. Fifty-two consecutive patients with BC with an indication for adjuvant chemotherapy between 2013 and 2016 were retrospectively included. We performed metabolomic profiling of tumor resection samples using liquid chromatography-mass spectrometry. Here, four hundred and forty-nine identified metabolites were selected for further analysis. Clusters obtained using 5 unsupervised ML methods (PCA k-means, sparse k-means, spectral clustering, SIMLR and k-sparse) were compared in terms of clinical and biological characteristics. With an optimal partitioning parameter k=3, the five methods identified three prognosis groups of patients (favorable, intermediate, unfavorable) with different clinical and biological profiles. SIMLR and K-sparse methods were the most effective techniques in terms of clustering. *In-silico* survival analysis revealed a significant difference for 5-year predicted OS between the 3 clusters. Further pathway analysis using the 449 selected metabolites showed significant differences in amino acid and glucose metabolism between BC histologic subtypes. Our results provide proof-of-concept for the use of unsupervised ML metabolomics enabling stratification and personalized management of BC patients. The design of novel computational methods incorporating ML and bioinformatic techniques should make available tools particularly suited to improving the outcome of cancer treatment and reducing cancer-related mortalities.

Introduction

Breast cancer (BC) is the most common type of cancer in women worldwide and the second leading cause of cancer-associated deaths [1]. Treatment strategy may be guided by two classifications indicating the aggressiveness of the tumor. The anatomo-clinical classification is based on age, TNM, histological factors (histological grade, Ki-67) as well as on hormonal-receptor status and Her-2 expression. The molecular classification resulting from genomic [2], transcriptomic [3] and proteomic [4] analyses introduced the concept of luminal A, luminal B, Her-2 and basal-like BC [5-7]. This latter classification from Perou and Sorlie was assessed using unsupervised analyses [6, 8]. Efforts have been made to develop multivariate prognostic models such as AdjuvantOnline®, PREDICT Tool [9, 10] and multigene predictors [11, 12]. The use of biomarker-based tests, including omic-based tests, has steadily increased over the last decade as a result of the need for personalized treatment strategies designed to optimize outcomes [13-18]. Several genomic prognostic markers have been described for BC such as OncotypeDX®, Prosigna®, MammaPrint®, Endopredict® Genomic grade index® and BC Index® [19]. Two markers are commercially available and are increasingly used in clinical practice (21-gene recurrence score OncotypeDX® and 70-gene prognostic signature MammaPrint®). However, heterogeneity persists in biological features within BC subtypes, thus highlighting the need to improve the taxonomy [20]. This heterogeneity may be related to specific combinations of genetic, pathological and environmental factors leading to specific metabolic alterations and interactions [21, 22].

Metabolomics is a new and growing field dedicated to the study of metabolism at overall level that promises to provide new insights into disease mechanisms and drug effects. Indeed, metabolomics may offer a complementary approach to genomics and could be used to better understand the influence of the environment on tumor phenotype [23]. Two distinct approaches characterize metabolomics: a targeted approach aimed at quantifying as accurately as possible a limited number of predefined metabolites of interest [24] and an untargeted approach aimed at measuring, without any a priori, as many metabolites as possible in a sample [25, 26]. As with other omics approaches, metabolomics generates high-dimensional data. The processing of these data can be done applying supervised or unsupervised machine learning (ML) algorithms that are increasingly used for medical diagnosis and therapeutic strategy guidance [27-29]. Unsupervised ML, in which no a priori class label information is given to guide the algorithm [30], seems a suitable alternative to analyse these data and address the problem of BC heterogeneity [6]. The aim of this study was to compare metabolomic signatures of BC obtained using five different unsupervised ML methods. To evaluate the consistency of our results, the clusters obtained by unsupervised ML methods were compared with patients' clinical characteristics and identified metabolic pathways.

Material and methods

Patients

This is a retrospective cohort study based on data and samples from 52 patients already available in the Centre Antoine Lacassagne tumor bank and collected during routine practice between 2013 and 2016. Patient tumor characteristics were: clinical stages I to III_B biopsy-proven BC, with an indication for post-surgery adjuvant therapy. Tumor phenotypes were classified in three subtypes; triple-negative (estrogen receptor, progesterone receptor and Her-2 non-over-expressed); luminal (estrogen receptor and/or progesterone receptor positive and Her-2 non-over-expressed); Her-2 over-expressed (Her-2 over-expressed, estrogen receptor and progesterone receptor either positive or negative) [31]. After surgery, all patients were treated according to current guidelines, with sequential chemotherapy including anthracyclines (epirubicin and cyclophosphamide) and taxanes followed by radiotherapy. Patients with Her-2 over-expressed tumors were treated with trastuzumab concurrently with taxanes and continued for one year. Patients with luminal BC were then treated by endocrine therapy with tamoxifen or an aromatase inhibitor, based on menopausal status. 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 were either extracted from our facility's digital records or retrieved by telephone if patients had changed facilities during surveillance. Written informed consent was obtained from all study participants. All procedures performed in this study 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).

Data-preprocessing, metabolite identification, statistical and pathway analysis

Sample collection, preparation and data-processing using MZmine [32, 33] are shown in Supplementary File S1 and Fig. S1. Metabolites obtained from positive and negative ionization modes were combined. Only metabolites with no null values after pre-processing were selected for analysis. When a metabolite was detected in both positive and negative modes, only the mode offering the highest average intensity was considered. In order to eliminate noisy data, a filtering function was applied before statistical analysis. Statistical analysis was performed on 449 metabolites. Identification of metabolic pathway was performed using MetaboAnalyst database sources [34]. The impact score was determined by the relative pathway topological effect of the metabolites, and *-log(p)* was used as the enrichment score, reflecting the probability of the pathway being identified at random; the number of "hits" was the actual number of matched metabolites in the pathway. For selection of the most relevant pathways, we applied the following criteria: Impact>0, FDR<0.25 and p<0.05 [35].

A Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/) was used to display all possible logical relations between the metabolites or pathways identified by the clustering methods. Differences between clusters regarding the most active metabolites were plotted using boxplots.

Clustering algorithms

Five unsupervised clustering methods were selected and compared: Principal Component Analysis (PCA) k-means, Spectral clustering, Sparse k-means, SIMLR and k-sparse. In order to apply these five unsupervised clustering methods, the optimal number of clusters was determined in advance using five criteria: gap [36], silhouette [37, 38], Davies-Bouldin [39], Calinski-Harabasz [40] and Single-cell Interpretation via Multi-kernel LeaRning (SIMLR) method [41]. PCA k-means clustering, which is frequently used in biology, combines PCA to reduce the number of dimensions of a dataset and the k-means method to minimize the intra-cluster variance for a chosen number of k clusters [42-44]. Spectral clustering [45, 46] is based on graph theory. It consists of identifying dense regions in a multidimensional dataset, i.e. observations that can form a non-convex set but are close to each other. Sparse k-means clustering was developed in 2010 by Witten and Tibshirani [8]. This method is based on a Least Absolute Shrinkage and Selection Operator (LASSO) approach [47] and combines the LASSO approach and the k-means method which simultaneously find the clusters and select features. SIMLR clustering [41] was developed to analyze scRNA-seq data. This method searches for appropriate cell-to-cell similarity metrics to perform dimension reduction and clustering. In multiple-kernel learning frameworks, this method may be especially beneficial for data containing no clearly identifiable clusters. K-sparse clustering [48] is a recent algorithm combining dimension reduction and relevant feature selection using a constraint in L1-norm rather than a lasso-type penalty to select the features. The performance of an unsupervised clustering method is measured by its ability to partition data. Partitioning is considered optimal when it minimizes the average distance between patients within a cluster (homogeneity) and maximizes cluster distances 2 by 2 (separability). The performances of the five methods were compared using the silhouettes index (SI) [38]. The SI ranges between -1 and 1 and assesses whether a patient belongs to the "right" cluster. The closer the index is to 1, the more satisfactory the assignment of a patient to a cluster. The t-SNE method was used for data visualization [49]. Processing times were obtained on a computer using an i5 processor (3.1 GHz).

Clinical evaluation

The relevance of the discovered clusters was assessed by comparing the clinical and survival characteristics between clusters by means of 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. Overall survival (OS) was defined as the time between diagnosis and death due to any cause. Specific survival

(SS) was determined by the time between diagnosis and death due to BC. Recurrence-Free Survival (RFS) was defined as the time between diagnosis and the first recurrence (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, SS, and RFS were estimated using the Kaplan-Meier method. Median follow-up with a 95% confidence interval was calculated by reverse Kaplan–Meier method. All analyses were performed with Matlab® R2018b for PCA k-means, Spectral clustering, SIMLR (https://github.com/BatzoglouLabSU/SIMLR/tree/SIMLR/MATLAB) and k-sparse clustering and R [50] using package Sparcl [51] for sparse k-means clustering. The difference between clusters regarding the most biologically significant metabolites was plotted using boxplots. For clinical and biological analyses, all *p*-values < 0.05 (two-sided) were considered statistically significant.

Prediction for 5- and 10-year overall and specific survival

Online PREDICT tool (https://breast.predict.nhs.uk/tool) [9, 10, 52] was used to estimate predicted OS (pOS) and predicted SS (pSS) at 5 and 10 years, based on several patient and tumor characteristics. For each patient, ten characteristics were entered manually: age at diagnosis, menopausal status, estrogen receptor status, Her-2 status, Ki-67 status, tumor stage, histological grade, mode of detection, number of positive nodes and presence of micrometastases. If information was missing for detection, bisphosphonate therapy or menopausal status, patients were not excluded but the "unknown" category was used. Only one patient was excluded because of missing tumor grade data. A 1000 resamples bootstrap was used to estimate the 95% confidence interval.

Results

Patient characteristics

Tumor and treatment features of the 52 patients are described in Table 1. Median age was 63 years (range: 37-88). The main histological type was invasive ductal carcinoma (92%), and the main tumor stages were T1 (40.5%) and T2 (46%). Twenty-four patients (46%) presented axillary lymph node invasion. Two patients (4%) were oligometastatic at diagnosis. Forty-three percent of patients had histological grade II tumors and 47% had grade III tumors. Half of the patients had negative hormone receptor status (48%) and 24% of patients had Her-2 over-expression. Median follow–up was 48.5 months (95%CI [43-54.5]). Twenty-one patients presented a recurrence: 4 local recurrences (7.5%), 6 regional recurrences (11.5%) and 11 metastatic recurrences (21%). Three-year OS was 90% [82-99], 3-year SS was 92% [85-100] and 3-year RFS was 82% [72-93] (Supplementary Fig. S2). Median OS, SS, and RFS were not reached.

Clustering results

Estimated number of clusters

Using four methods (Gap statistic, Calinski-Harabasz, Silhouette and SIMLR criterion), the optimal number of clusters was equal to three (k=3) (Supplementary Fig. S3). Only for Davies-Bouldin criterion, the optimal number of clusters was equal to four (k=4). It seems reasonable, therefore, to conclude that the optimal number of clusters is equal to 3.

Patient distribution

Three clusters were identified with each of the five clustering methods, (Figure 1). In terms of processing times, PCA k-means was the fastest and K-sparse was the longest (Supplementary Table S1). SIMLR and k-sparse methods were the most discriminant with an average silhouette value of 0.85 and 0.91, respectively (Figure 2). Seventy-three percent of patients (38/52) were ranked in the same clusters by the five methods, 17.5% of patients (9/52) were classified in the same clusters by 4 methods and 9.5% of patients (5/52) were classified in the same clusters by 3 methods.

Comparison of clinical characteristics between clusters

As shown in Table 2, the 5 methods revealed significant inter-cluster differences. Patients in cluster 3 had mainly unfavorable prognostic factors: tumor stage T2/T3, histological grade III, high mitotic score and triple-negative phenotype. In contrast, patients in cluster 1 had mainly favorable prognosis factors: tumor stage T1, histological grade I/II, lower mitotic score and luminal phenotype, whereas patients in cluster 2 constitute an intermediate group presenting both good and poor prognostic factors. Clusters defined by PCA k-means were significantly different for 5 characteristics: tumor stage, mitosis, tumor phenotype, Her-2 status and luminal. Clusters defined by Spectral Clustering were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. Clusters defined by Sparse k-means were significantly different for 4 characteristics: histological grade, tumor phenotype, Her-2 status and luminal. Clusters defined by SIMLR were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. Clusters defined by K-Sparse were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. From a strictly clinical point of view, Spectral clustering, SIMLR and K-sparse are the 3 most discriminating methods. Indeed, for these 3 methods, six prognostic factors (tumor stage, histological grade, mitosis score, Ki-67, tumor phenotype and luminal) were distributed significantly differently between the 3 clusters.

Comparison of survival and predicted survival between clusters

None of the methods created clusters showing significant differences for OS, SS or RFS. Analysis of patients' survival data using PREDICT tool are presented in Table 3 and show a predicted survival gradient for clusters obtained with the 5 methods for OS and SS. There were significant differences for 5-year pOS between clusters obtained with K-sparse (p=0.021), Sparse K-means (p=0.049), Spectral and clustering (p=0.021). The five methods showed a significant difference for 5-year pSS between clusters. In terms of 10-year pOS, there were no significant differences between clusters obtained by any of the 5 methods. In contrast, for 10-year pSS, the 5 methods showed significant differences between clusters. Patients in cluster 3 clearly showed the poorest predicted survival.

Comparison of the most impactful metabolites according to the five methods

To relate the impact of 449 metabolites to cluster construction, we ranked these metabolites extracted from each of the five methods based on their functional contributions to outputs. With this approach, we classified the relative impact of metabolites on cluster construction and on the identification of metabolic signatures. The highest-ranked metabolites were those that provided relevant information to the signature versus those that provided redundant information or no information. Among a total of 449 metabolites, 116 (26%) were selected by K-sparse clustering and 69 (15%) by Sparse K-means clustering. As for the three other methods, which don't select sparse features, the number of metabolites remained equal to 449. The 50 most effective metabolites identified by the five methods are presented in Supplementary Table S2. Furthermore, a comparison of the top 50 metabolites in each of the 5 methods is presented using a Venn diagram (Figure 3) and a bipartite graph (Supplementary Fig. S4). Two metabolites were shared by the 5 methods (Creatine, L-Proline), 9 were shared by 4 methods (Betaine, Glutathione, Humulinic Acid A, Isoleucyl-Methionine, L-Carnitine, L-Methionine, L-Phenylalanine Triethanolamine, Alnustone), 28 were shared by 3 methods and 38 were shared by 2 methods (Table 4).

Comparison between 5 methods of identified metabolic pathways

For a better understanding of metabolic dysregulation among BC subtypes, pathway analysis was performed. Identification of all the metabolic pathways highlighted by each of the 5 methods as shown in Supplementary Table S3. The most relevant pathways for each of the 5 methods are shown in Table 5. Sparse K-means identified only one statistically significant pathways, "cysteine and methionine metabolism", involved in amino acid metabolism. K-Sparse identified 3 different pathways: "glycerolipid metabolism", "Starch and sucrose metabolism" involved in carbohydrates metabolic pathway and "Aminoacyl-tRNA biosynthesis" involved in translation pathway. Spectral clustering identified 17 pathways, the 3 most important being "Glycine, serine and threonine metabolism",

"Alanine, aspartate and glutamate metabolism" and "Histidine metabolism and glutathione metabolism" involved in amino acid metabolic pathway. PCA K-means identified 10 pathways the 3 most important of which are "Alanine, aspartate and glutamate metabolism" involved in amino acid metabolic pathway, "Pyruvate metabolism" involved in carbohydrates metabolic/glucose oxidation pathway and "Citrate cycle (TCA cycle)" involved in energy metabolic pathway.

Finally, with 30 identified pathways, SIMLR is the method that identified the most metabolic pathways. Of these, the 3 most important highlighted metabolic pathways are "arginine and proline metabolism", "glycine, serine and threonine metabolism" and "alanine, aspartate and glutamate metabolism", involved in amino acid metabolic pathways. The Venn diagram (Figure 4) shows the overlap of pathways detected by the five methods. Amino acid metabolism appeared to be the most frequently modified pathway. Enrichment and pathway analyses also showed modifications in glucose metabolism. From the biological point of view, SIMLR and spectral clustering are the two methods that identified the most relevant metabolic pathways.

Comparison of intensity of metabolites between the 5 methods

Among amino acid and glucose metabolisms, fourteen related metabolites were selected as potential biomarkers in BC [53-56]. As shown in Supplementary Fig. S5, the intensities of these 14 metabolites were compared between the 3 clusters for each of the 5 methods. The intensity of Uridine diphosphate (UDP) glucose, Guanine, L-Glutamine, L-Glutamic acid, L-Isoleucine, L-Proline, L-Methionine, L-Phenylalanine, Pyruvic acid, Spermine, Glutathione, Creatine, L-Carnitine and L-Acetylcarnitine were statistically significant between at least one of the clusters. The five methods agree that cluster 3 patients have low levels of Creatine, L-acetylcarnitine, L-Glutamic acid and high levels of Guanine, L-Isoleucine, L-Phenylalanine, Pyruvic acid and Spermine (Figure 5). These metabolite levels seem to be predictive of poor prognosis [54, 57, 58].

Discussion

From a machine learning perspective

To the best of our knowledge, this proof-of-concept study is the first to compare different unsupervised ML methods to identify metabolomics-based prognostic signatures in BC. Analyses were performed intentionally without any prior clinical or biological assumptions. Clinical and biological interpretations were performed only after cluster identification. The objective of our study was to compare different unsupervised ML algorithms for feature selection from untargeted metabolomic data and to evaluate the capacity of these methods to select relevant features for further use in prediction models. This study did not seek to highlight significant differences but rather to assess how unsupervised methods might behave with high-dimension metabolic data and to open up new

perspectives in the particularly active domain of BC phenotype predictors. We demonstrated that the K-sparse and SIMLR methods have a higher clustering performance compared with the three other popular unsupervised ML methods in detecting groups of patients with BC using metabolomic data. Interestingly, even though the spectral method is a little less clinically efficient than the k-sparse and SIMLR methods, it identified relevant metabolic pathways.

Our study suffers from various limitations, namely the relatively small number of patients and the monocentric and retrospective nature of the study. Besides, our results could not be validated on an external cohort. The clustering performances were assessed only by internal validation based on silhouette value. Indeed, we could not compare the labels obtained from our classification with the true labels to calculate the accuracy of the classification since the true labels were unknown. Other unsupervised ML methods such as model-based clustering, bi-clustering and deep learning may be of value in this analysis and should be further explored.

These considerations raise important questions: in future, on what basis should decisions be made? On results from a single method? Or on results provided by several methods? In view of the findings we have highlighted, it seems that decisions should be taken collegially, i.e. based on the results of a set of methods, as at multidisciplinary consultation meetings involving health professionals from different disciplines and whose skills are essential to take decisions ensuring patients the best possible care according to the state of the science.

From a clinical perspective

From a clinical point of view, the methods were able to highlight three distinct groups of patients with different clinical profiles. Patients identified in cluster 1 may be considered to have the best prognosis, patients in cluster 2 an intermediate prognosis, while patients in cluster 3 may be considered to have the worst prognosis. The results in Table 2 show that the tumors of patients in cluster 1 were predominantly non-invasive and non-proliferative, whereas the tumors of cluster 3 patients were mainly invasive and proliferative. Tumors in cluster 2 were rather invasive but not proliferative, hence the intermediate prognosis. We hypothesize that these patients would have an intermediate (atypical) biological profile, which is why the methods are discordant. We further evidence heterogeneity within the triple-negative BC subpopulation with most of the patients classified in cluster 3. However, a third of the triple-negative patients were in cluster 1 Recent molecular profiling studies of triple-negative BC using parallel sequencing and other “omics” technologies have also uncovered an unexpectedly high level of heterogeneity as well as a number of common features [59, 60].

In addition, no significant difference between clusters could be demonstrated in terms of age, histologic type, lymph node involvement, metastasis or survival (OS, SS or RFS). Indeed, with a median follow-up of only 48.5 months, this duration is insufficient to demonstrate a significant difference in

terms of OS, SS, or RFS. Nevertheless, from our study, it is quite easy to predict that patients in cluster 3 have the highest risk of progression and that, conversely, patients in cluster 1 have the lowest risk of progression. With a 5-year pOS rate at around 75% for cluster 1, 70% for cluster 2 and 60% for cluster 3, *in-silico* analyses have demonstrated their high potential value [28, 61, 62] and confirmed that patients in cluster 3 have a poorer prognosis [63, 64]. One limit of our study could be the representativity of our population, e.g. it is recognized that BCs in younger patients (< 40 years) are more aggressive [65]. Our study did not include a large number of young patients, which could explain why no significant difference was demonstrated in terms of age between clusters. Similarly, with only three patients with invasive lobular carcinoma (6%), our results did not identify a metabolic signature associated with this phenotype. Previous studies have shown a survival benefit in favor of invasive lobular carcinoma [66, 67] and metabolomic studies focused on this particular type of BC could provide valuable biological information. Furthermore, due to over-representation of hormonal-receptor negative tumors (48%) in our population compared to the literature [68], our population could have had an unfavorable prognosis. This bias may result from our method of tumor selection. We decided to analyze frozen samples available in our biobank. Obviously, hormonal-receptor negative, triple-negative, Her-2-positive tumors are more often frozen and stored for further molecular testing and inclusion in clinical trials. In the present study, it is interesting to note that the five methods classified 73% of the patients in the same cluster. Among the 27% of patients classified differently by at least one of the methods, 9.5% of patients were classified heterogeneously by the five methods. Indeed, for each of these 5 patients, three methods classified them in one cluster and 2 others in another cluster without any connection between the types of method used. Moreover, it is interesting to note that the different methods classified patients, on the one hand, in either the good prognostic cluster or the intermediate prognostic cluster or, on the other, in either the intermediate prognostic cluster or the poor prognostic cluster, but never in the good prognostic cluster or the poor prognostic cluster. A clinical analysis of these 5 patients showed that they had atypical clinical profiles, probably due to particular biological profiles. These atypical profiles would explain why no classification consensus could be highlighted. Overall, ML methods must remain a decision-making tool for the clinician, especially in cases where patients have particular clinical and biological characteristics. To avoid possible medical errors, the final responsibility for the decision lies with the clinician [69].

From a biological perspective

From a physiological point-of-view, this study extends the molecular stratification of BC to metabolomic profiles. Indeed, our results suggest that dysregulation of metabolic pathways exists between BC subtypes and that a particular amino acid profile characterizes the different BC histologic subtypes. Dysregulations of amino acid metabolism are well-known key events during cancer

development [70] and are emerging hallmarks of cancers [71, 72]. Amino acids serve not only as building blocks in protein synthesis but also as energy sources favoring cancer cell proliferation and growth [73]. Of interest, we identified significant differences between the BC subtypes of three metabolic pathways (i.e. Glycolysis and lactate production, Glutaminolysis, and amino acid) that play a pivotal role in BC growth [74, 75]. Using the five methods, we consistently found that patients in cluster 3 showed higher levels of Guanine, L-Isoleucine, L Methionine, L-Phenylalanine, Pyruvic acid, Spermine and low levels of Creatine, L-Acetylcarnitine and L-Glutamic acid. Our results suggested that these metabolites could be candidate biomarker predictors of poorer prognosis [76-80]. All these results are consistent with the literature [54, 81-84].

Indeed, to meet the biosynthetic needs associated with rapid proliferation, cancer cells must increase the import of nutrients. Two main metabolites are essential for biosynthesis and survival in mammalian cells, and particularly in cancer cells: glucose [85] and glutamine [86]. The increased glucose uptake in tumors compared to other healthy and non-proliferative tissues was first described more than 90 years ago by Otto Warburg [87]. Glucose is the primary energy source of all cells because of its involvement in many processes such as glycolysis or the Krebs cycle [88] in mitochondria. Unlike healthy cells that adapt to available substrates (glucose/fatty acids/proteins), some tumor cells are addicted to glucose. The other important point is that, once metabolized, tumor cells will prefer lactic fermentation to the Krebs cycle.

Lastly, the precise etiology of BC is still unknown even though some genetic, epigenetic and environmental factors have been identified [89]. It has been conclusively demonstrated that cancer cell metabolism is heavily influenced by microenvironmental factors, including nutrient availability. Sullivan and coworkers [90] found that diet affects local nutrient availability. This effect can lead to substantial changes in the metabolism of tumor cells, thereby modifying the response of these cells to drugs targeting metabolism. Drugs capable of inhibiting tumor proliferation may then become ineffective. Therefore, knowledge of microenvironmental nutrient levels is essential to a better understanding of tumor metabolism.

Outcomes for cancer patients vary greatly. The classification of BC into subtypes has been was defined in the literature on the basis of molecular characterization of proteomics (single omic). This has helped improve prognosis and personalized treatment. These considerations have motivated efforts to produce large amounts of multi-omic data such as TCGA [91] and ICGC [92]. However, current algorithms still face challenges and need to integrate omic data [93-96]. Defining BC subtypes using multi-omic data could help to better understand some of the dark areas that still persist in the field of tumor mechanisms in order to offer even more personalized treatments.

Conclusion

In the era of personalized medicine, OMICS science (genomics, transcriptomics, proteomics, and metabolomics) must contribute to the quest for cancer-specific biomarkers. The present study argues in favor of further research in this domain. Metabolomics is emerging as a relevant and promising tool for the classification of BC to enable more precise diagnosis [53, 97-99]. Even though it is less accurate than the targeted approach, untargeted metabolomics nevertheless permits identification and quantification of a vast number of major metabolites. Thus, this approach presents a particular interest in the search for new candidate biomarkers [100-102] and could be applied in everyday medical practice given that the cost and duration of metabolomic analyses are relatively low. However, due to the retrospective design of our study and the small number of patients recruited, our results need to be validated in a larger cohort and in the context of a prospective clinical trial.

Key points

- 1st study combining non-supervised machine learning approach and non-targeted metabolomic.
- Performance of unsupervised machine learning applied to small data.
- K-sparse and SIMLR seem to be the most discriminating methods, they were able to identify, according to the metabolome of the tumor, 3 populations whose clinico-biological characteristics are distinct.
- Metabolomic approach seems to be a relevant and promising in the classification of breast cancers

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest to disclose.

Funding

The authors declare no competing financial interests.

Acknowledgements

The authors acknowledge support from the Centre Antoine Lacassagne, TIRO Unit, University Côte d'Azur, and the Departmental Council of the Alpes Maritimes, France.

The authors sincerely thank Mrs. Clair Della Vedova for her help in developing the figures.

References

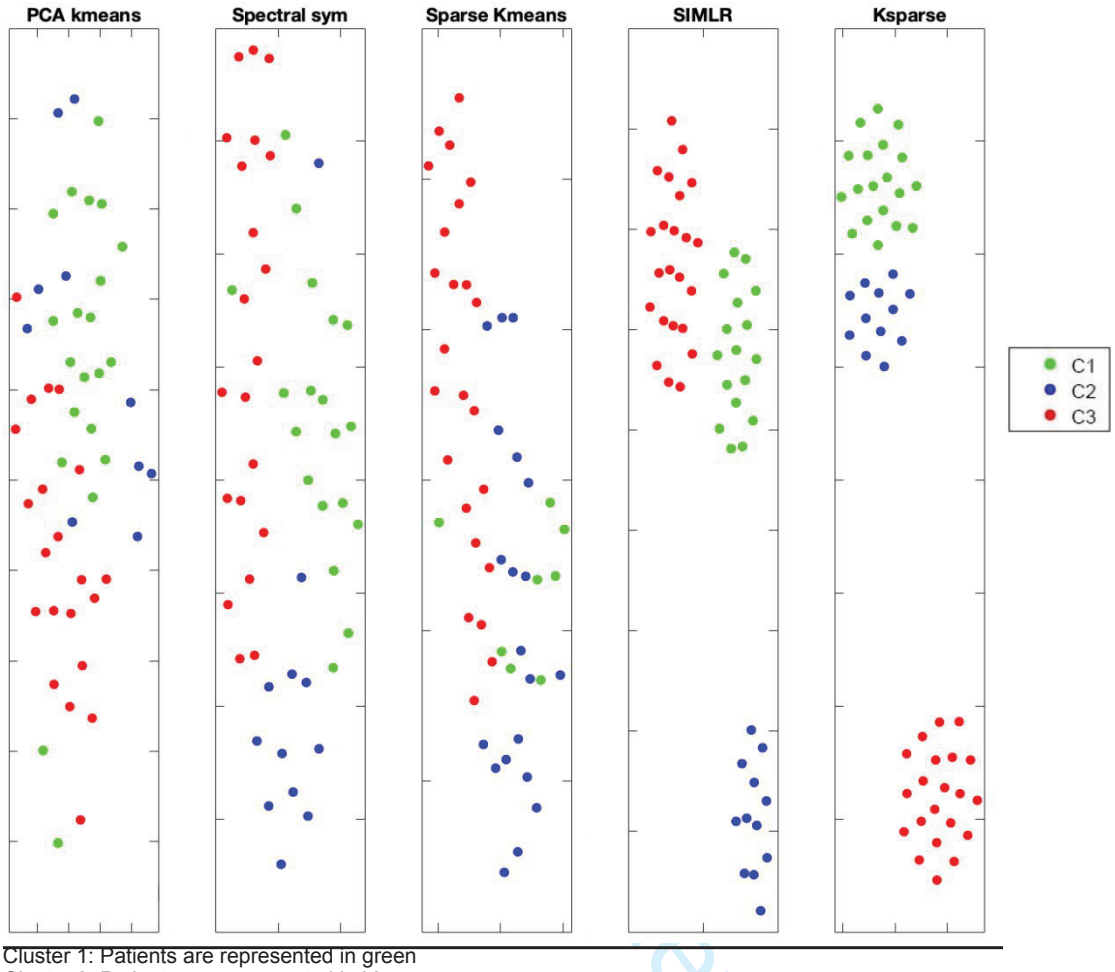
1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017, CA Cancer J Clin 2017; 67:7-30.
2. Perou CM, Jeffrey SS, van de Rijn M et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Proc Natl Acad Sci U S A 1999; 96:9212-9217.
3. Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays, Nature 2000; 405:827-836.
4. Pandey A, Mann M. Proteomics to study genes and genomes, Nature 2000; 405:837-846.
5. Perou CM, Sorlie T, Eisen MB et al. Molecular portraits of human breast tumours, Nature 2000; 406:747-752.
6. Sorlie T, Perou CM, Tibshirani R et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, Proc Natl Acad Sci U S A 2001; 98:10869-10874.
7. Sorlie T, Tibshirani R, Parker J et al. Repeated observation of breast tumor subtypes in independent gene expression data sets, Proc Natl Acad Sci U S A 2003; 100:8418-8423.
8. Witten DM, Tibshirani R. A framework for feature selection in clustering, Journal of the American Statistical Association 2010; 105:713-726.
9. Candido Dos Reis FJ, Wishart GC, Dicks EM et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation, Breast Cancer Res 2017;19:58.
10. Wishart GC, Azzato EM, Greenberg DC et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer, Breast Cancer Res 2010;12:R1.
11. Ross JS. Multigene predictors in early-stage breast cancer: moving in or moving out?, Expert Rev Mol Diagn 2008;8:129-135.
12. Ross JS, Hatzis C, Symmans WF et al. Commercialized multigene predictors of clinical outcome for breast cancer, Oncologist 2008;13:477-493.
13. Cao Y, DePinho RA, Ernst M et al. Cancer research: past, present and future, Nat Rev Cancer 2011;11:749-754.
14. McShane LM, Polley MY. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility, Clin Trials 2013;10:653-665.
15. Ehmann F, Caneva L, Prasad K et al. Pharmacogenomic information in drug labels: European Medicines Agency perspective, Pharmacogenomics J 2015;15:201-210.
16. Buyse M, Loi S, van't Veer L et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer, J Natl Cancer Inst 2006;98:1183-1192.
17. Wang Y, Klijn JG, Zhang Y et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Lancet 2005;365:671-679.
18. van de Vijver MJ, He YD, van't Veer LJ et al. A gene-expression signature as a predictor of survival in breast cancer, N Engl J Med 2002;347:1999-2009.
19. Wesolowski R, Ramaswamy B. Gene expression profiling: changing face of breast cancer classification and management, Gene Expr 2011;15:105-115.
20. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer?, Nat Rev Cancer 2012;12:323-334.
21. Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond, Cell 2008;134:703-707.
22. McClellan J, King MC. Genetic heterogeneity in human disease, Cell 2010;141:210-217.
23. Cannon WB. The wisdom of the body, 2nd ed. Oxford, England: Norton & Co., 1939.
24. Roberts LD, Souza AL, Gerszten RE et al. Targeted metabolomics, Curr Protoc Mol Biol 2012;Chapter 30:Unit 30 32 31-24.
25. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD et al. Untargeted Metabolomics Strategies-Challenges and Emerging Directions, J Am Soc Mass Spectrom 2016;27:1897-1905.
26. Vinayavekhin N, Saghatelian A. Untargeted metabolomics, Curr Protoc Mol Biol 2010;Chapter 30:Unit 30 31 31-24.
27. Camacho DM, Collins KM, Powers RK et al. Next-Generation Machine Learning for Biological Networks, Cell 2018;173:1581-1592.

28. Gal J, Milano G, Ferrero JM et al. Optimizing drug development in oncology by clinical trial simulation: Why and how?, Brief Bioinform 2017.
29. Yu MK, Ma J, Fisher J et al. Visible Machine Learning for Biomedicine, Cell 2018; 173:1562-1565.
30. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects, Science 2015; 349:255-260.
31. Tang P, Tse GM. Immunohistochemical Surrogates for Molecular Classification of Breast Carcinoma: A 2015 Update, Arch Pathol Lab Med 2016; 140:806-814.
32. Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data, BMC Bioinformatics 2005; 6:179.
33. Pluskal T, Castillo S, Villar-Briones A et al. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics 2010;11:395.
34. Xia J, Mandal R, Sinelnikov IV et al. MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis, Nucleic Acids Res 2012; 40:W127-133.
35. Irizarry RA, Wang C, Zhou Y et al. Gene set enrichment analysis made simple, Stat Methods Med Res 2009; 18:565-575.
36. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2001; 63:411-423.
37. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009.
38. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 1987; 20:53-65.
39. Davies DL, Bouldin DW. A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence 1979:224-227.
40. Caliński T, Harabasz J. A dendrite method for cluster analysis, Communications in Statistics-theory and Methods 1974; 3:1-27.
41. Wang B, Zhu J, Pierson E et al. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning, Nat Methods 2017; 14:414-416.
42. Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. 2007, p. 1027-1035. Society for Industrial and Applied Mathematics.
43. Lloyd S. Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM, IEEE Trans. Inform. Theor.(1957/1982) Google Scholar 1957.
44. Steinhaus H. Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV: 801-804 1956.
45. Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. In: Advances in neural information processing systems. 2002, p. 849-856.
46. Von Luxburg U. A tutorial on spectral clustering, Statistics and computing 2007;17:395-416.
47. Tibshirani R. Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) 1996:267-288.
48. Gilet C, Deprez M, Caillau J-B et al. Clustering with feature selection using alternating minimization, Application to computational biology, arXiv preprint arXiv:1711.02974 2017.
49. Maaten Lvd, Hinton G. Visualizing data using t-SNE, Journal of machine learning research 2008;9:2579-2605.
50. Team RC. R: A language and environment for statistical computing 2013.
51. Witten DM, Tibshirani R. sparcl: Perform sparse hierarchical clustering and sparse k-means clustering, R package version 2013;1.
52. Wishart GC, Bajdik CD, Azzato EM et al. A population-based validation of the prognostic model PREDICT for early breast cancer, Eur J Surg Oncol 2011;37:411-417.
53. Beger RD. A review of applications of metabolomics in cancer, Metabolites 2013;3:552-574.

54. Silva C, Perestrelo R, Silva P et al. Breast Cancer Metabolomics: From Analytical Platforms to Multivariate Data Analysis. A Review, Metabolites 2019;9.
55. McCartney A, Vignoli A, Biganzoli L et al. Metabolomics in breast cancer: A decade in review, Cancer Treat Rev 2018;67:88-96.
56. Gunther UL. Metabolomics Biomarkers for Breast Cancer, Pathobiology 2015;82:153-165.
57. Cardoso MR, Santos JC, Ribeiro ML et al. A Metabolomic Approach to Predict Breast Cancer Behavior and Chemotherapy Response, Int J Mol Sci 2018;19.
58. Asiago VM, Alvarado LZ, Shanaiah N et al. Early detection of recurrent breast cancer using metabolite profiling, Cancer Res 2010;70:8309-8318.
59. Bianchini G, Balko JM, Mayer IA et al. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease, Nat Rev Clin Oncol 2016;13:674-690.
60. Mills MN, Yang GQ, Oliver DE et al. Histologic heterogeneity of triple negative breast cancer: A National Cancer Centre Database analysis, Eur J Cancer 2018;98:48-58.
61. Belkacemi Y, Hanna NE, Besnard C et al. Local and Regional Breast Cancer Recurrences: Salvage Therapy Options in the New Era of Molecular Subtypes, Front Oncol 2018;8:112.
62. Buonaguro FM, Caposio P, Tornesello ML et al. Cancer Diagnostic and Predictive Biomarkers 2018, Biomed Res Int 2019;2019:3879015.
63. Senkus E, Kyriakides S, Ohno S et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up, Ann Oncol 2015;26 Suppl 5:v8-30.
64. Ponde NF, Zardavas D, Piccart M. Progress in adjuvant systemic therapy for breast cancer, Nat Rev Clin Oncol 2019;16:27-44.
65. Assi HA, Khoury KE, Dbouk H et al. Epidemiology and prognosis of breast cancer in young women, J Thorac Dis 2013;5 Suppl 1:S2-8.
66. Wasif N, Maggard MA, Ko CY et al. Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes, Ann Surg Oncol 2010;17:1862-1869.
67. Wang K, Zhu GQ, Shi Y et al. Long-Term Survival Differences Between T1-2 Invasive Lobular Breast Cancer and Corresponding Ductal Carcinoma After Breast-Conserving Surgery: A Propensity-Scored Matched Longitudinal Cohort Study, Clin Breast Cancer 2019;19:e101-e115.
68. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications, World J Clin Oncol 2014;5:412-424.
69. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence, Nat Med 2019;25:44-56.
70. Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism, Cell Metab 2016;23:27-47.
71. Hainaut P, Plymoth A. Targeting the hallmarks of cancer: towards a rational approach to next-generation cancer therapy, Curr Opin Oncol 2013;25:50-51.
72. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation, Cell 2011;144:646-674.
73. Li Z, Zhang H. Reprogramming of glucose, fatty acid and amino acid metabolism for cancer progression, Cell Mol Life Sci 2016;73:377-392.
74. Haukaas TH, Euceda LR, Giskeodegard GF et al. Metabolic Portraits of Breast Cancer by HR MAS MR Spectroscopy of Intact Tissue Samples, Metabolites 2017;7.
75. DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism, Sci Adv 2016;2:e1600200.
76. Jeon H, Kim JH, Lee E et al. Methionine deprivation suppresses triple-negative breast cancer metastasis in vitro and in vivo, Oncotarget 2016;7:67223-67234.
77. Zuo Y, Ulu A, Chang JT et al. Contributions of the RhoA guanine nucleotide exchange factor Net1 to polyoma middle T antigen-mediated mammary gland tumorigenesis and metastasis, Breast Cancer Res 2018;20:41.
78. Xiao F, Wang C, Yin H et al. Leucine deprivation inhibits proliferation and induces apoptosis of human breast cancer cells via fatty acid synthase, Oncotarget 2016;7:63679-63689.
79. Thomas TJ, Thomas T. Cellular and Animal Model Studies on the Growth Inhibitory Effects of Polyamine Analogues on Breast Cancer, Med Sci (Basel) 2018;6.

80. Melone MAB, Valentino A, Margarucci S et al. The carnitine system and cancer metabolic plasticity, Cell Death Dis 2018;9:228.
81. Lecuyer L, Dalle C, Lyan B et al. Plasma metabolomic signatures associated with long-term breast cancer risk in the SU.VI.MAX prospective cohort, Cancer Epidemiol Biomarkers Prev 2019.
82. Oikari S, Kettunen T, Tiainen S et al. UDP-sugar accumulation drives hyaluronan synthesis in breast cancer, Matrix Biol 2018;67:63-74.
83. Pan H, Xia K, Zhou W et al. Low serum creatine kinase levels in breast cancer patients: a case-control study, PLoS One 2013;8:e62112.
84. Phannasil P, Ansari IH, El Azzouny M et al. Mass spectrometry analysis shows the biosynthetic pathways supported by pyruvate carboxylase in highly invasive breast cancer cells, Biochim Biophys Acta Mol Basis Dis 2017;1863:537-551.
85. Mason EF, Rathmell JC. Cell metabolism: an essential link between cell growth and apoptosis, Biochim Biophys Acta 2011;1813:645-654.
86. Hensley CT, Wasti AT, DeBerardinis RJ. Glutamine and cancer: cell biology, physiology, and clinical opportunities, J Clin Invest 2013;123:3678-3684.
87. Warburg O, Wind F, Negelein E. The Metabolism of Tumors in the Body, J Gen Physiol 1927;8:519-530.
88. Anderson NM, Mucka P, Kern JG et al. The emerging role and targetability of the TCA cycle in cancer metabolism, Protein Cell 2018;9:216-237.
89. Fernandez MF, Reina-Perez I, Astorga JM et al. Breast Cancer and Its Relationship with the Microbiota, Int J Environ Res Public Health 2018;15.
90. Sullivan MR, Danai LV, Lewis CA et al. Quantification of microenvironmental metabolites in murine cancers reveals determinants of tumor nutrient availability. Elife 2019; 8.
91. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature 2008;455:1061-1068.
92. Zhang J, Baran J, Cros A et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data, Database (Oxford) 2011;2011:bar026.
93. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark, Nucleic Acids Res 2019;47:1044.
94. Mitra S, Saha S. A multiobjective multi-view cluster ensemble technique: Application in patient subclassification, PLoS One 2019;14:e0216904.
95. Wu C, Zhou F, Ren J et al. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection, High Throughput 2019;8.
96. Ramazzotti D, Lal A, Wang B et al. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival, Nat Commun 2018;9:4453.
97. Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: current trends and future perspectives, J Pharm Biomed Anal 2014;87:1-11.
98. Bennett DA, Waters MD. Applying biomarker research, Environ Health Perspect 2000;108:907-910.
99. Vermeersch KA, Styczynski MP. Applications of metabolomics in cancer research, J Carcinog 2013;12:9.
100. Jacob M, Lopata AL, Dasouki M et al. Metabolomics toward personalized medicine, Mass Spectrom Rev 2017.
101. Trivedi DK, Hollywood KA, Goodacre R. Metabolomics for the masses: The future of metabolomics in a personalized world, New Horiz Transl Med 2017;3:294-305.
102. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine, Nat Rev Drug Discov 2016;15:473-484.

Figure 1: Visualization of each cluster by clustering method using T-sne



Cluster 1: Patients are represented in green
Cluster 2: Patients are represented in blue
Cluster 3: Patients are represented in red

Figure 2: Silhouette value (SI) representation for each patient by clustering method

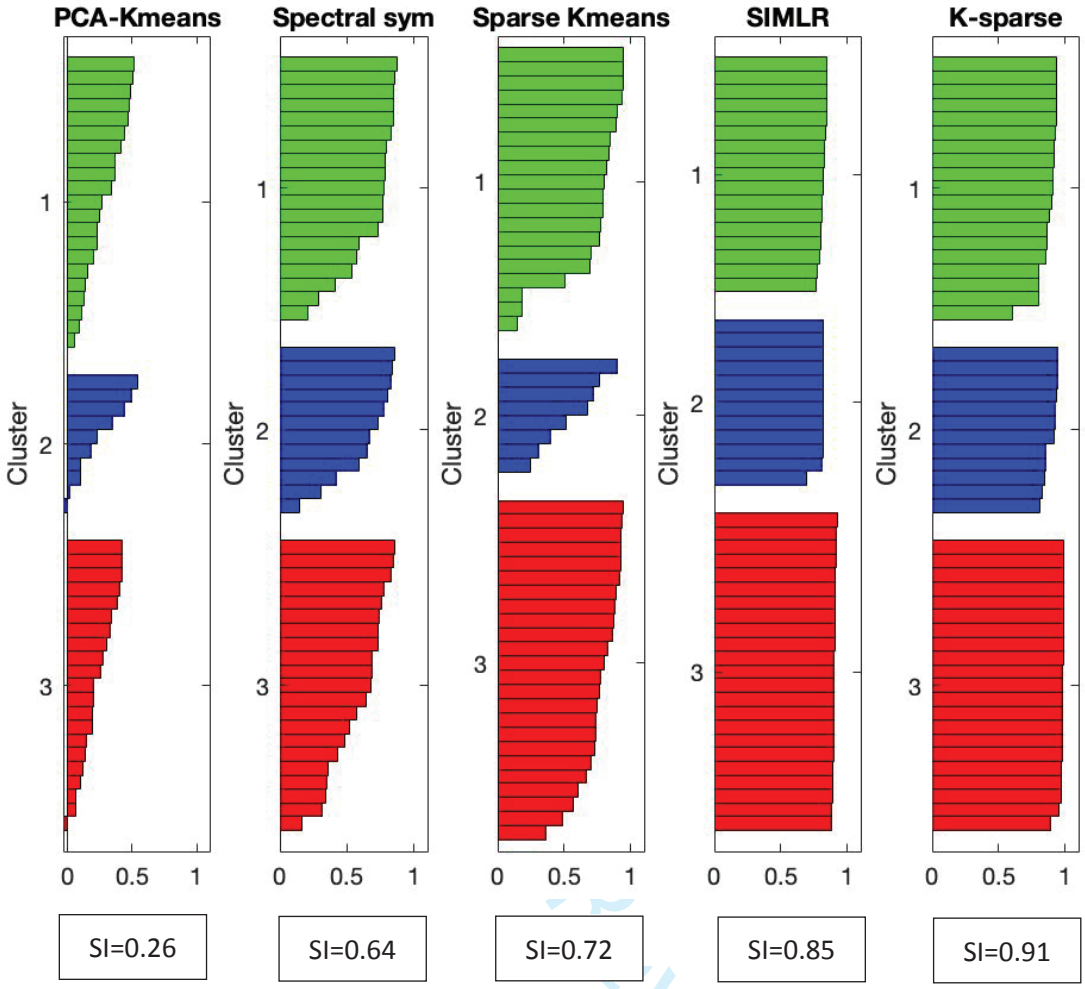


Figure 3: Venn diagram of metabolic that were in common or unique to the five clustering methods

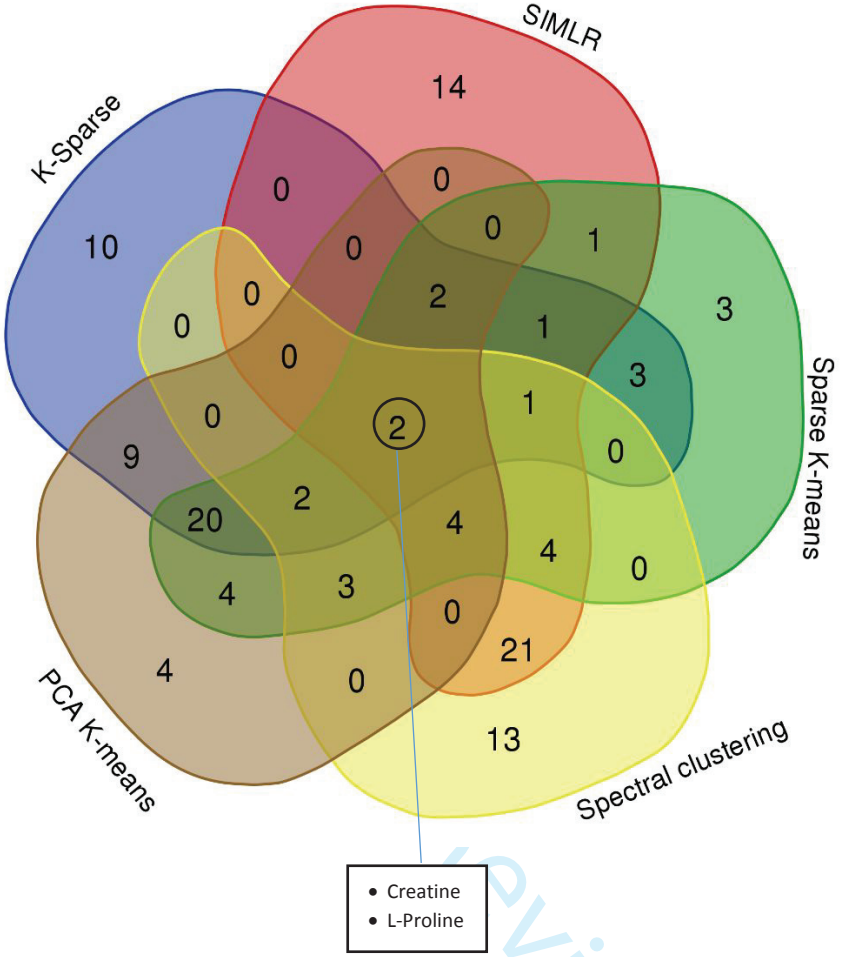


Figure 4: Venn diagram of pathways that were in common or unique to the five clustering methods.

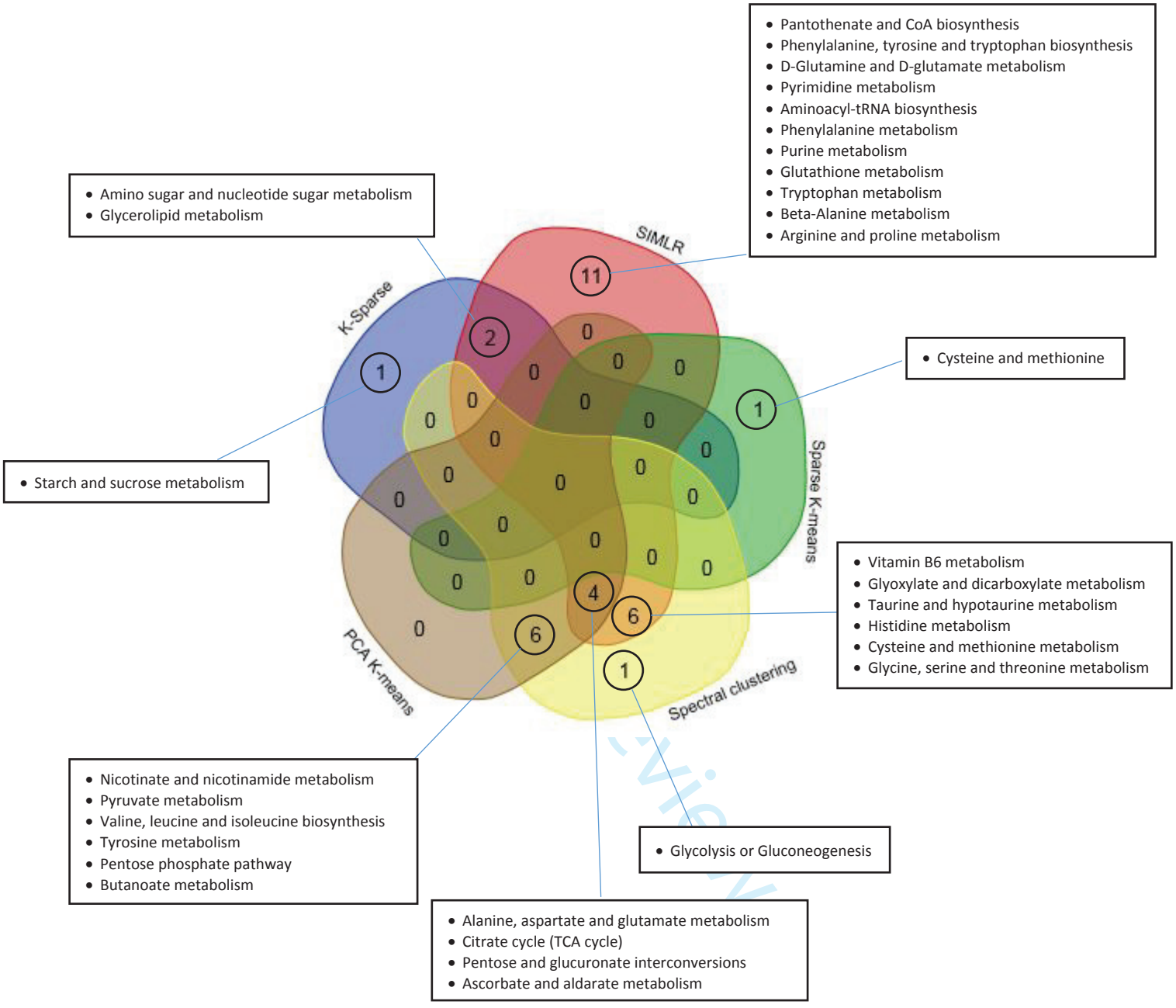


Figure 5: Boxplot of the 8 metabolites extracted from 5 ML methods

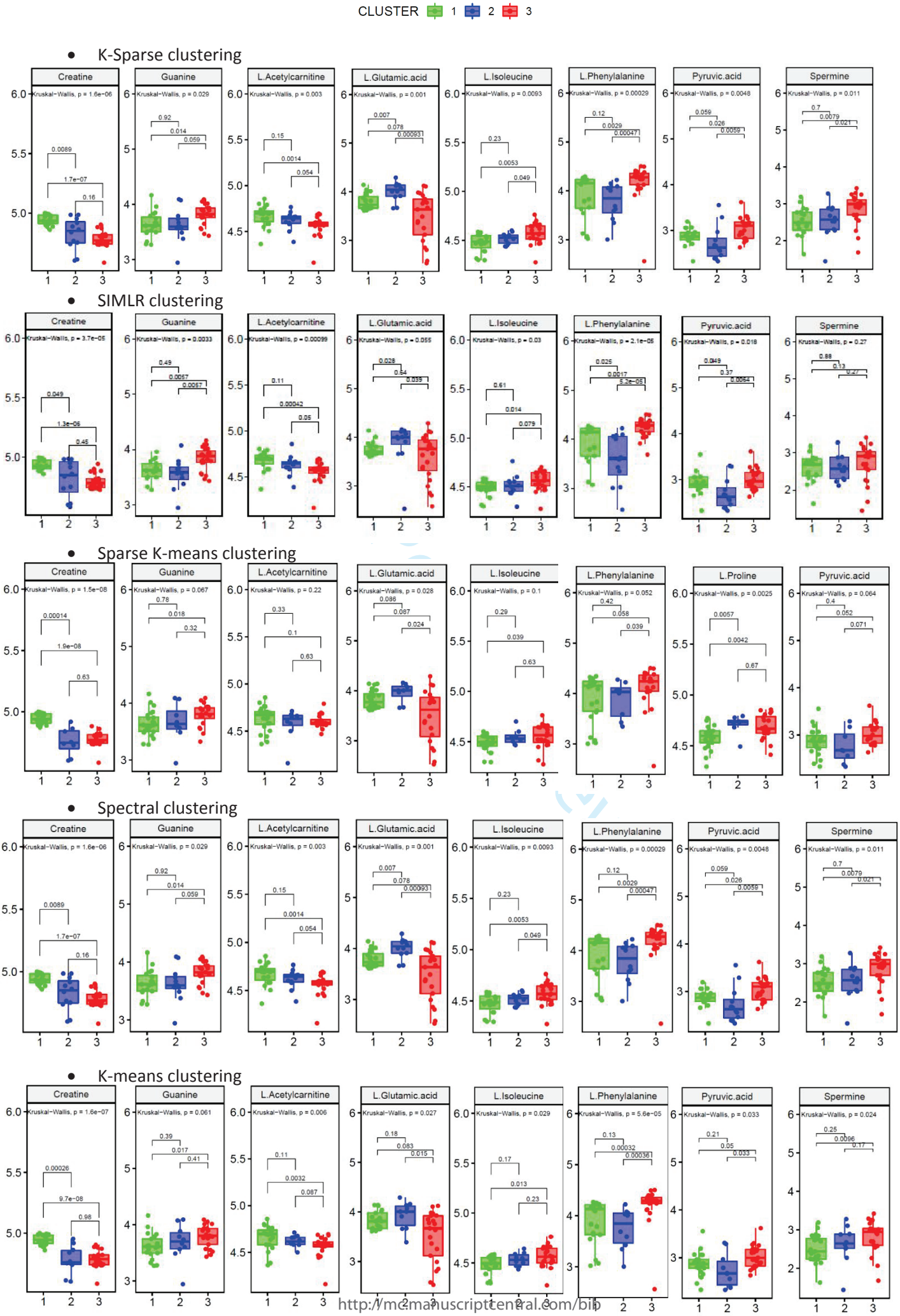


Table 1: Patients' demographics and treatment characteristics.

Clinical characteristic	No. of patients	%
Age (median min – max)	63.2 (37-88)	
Histology type		
Invasive ductal carcinoma	48	92
Invasive lobular carcinoma	3	6
Microinvasive carcinoma	1	2
Tumor stage		
T1	21	40.5
T2	24	46
T3	7	13.5
Axillary lymph node status		
N0	28	54
N+	24	46
Metastasis		
M0	50	96
M1	2	4
Histological grade		
I	5	10
II	22	43
III	24	47
Hormonal receptors status*		
Negative	25	48
Positive	27	52
Her-2 status		
Non-over-expressed	40	74
Over-expressed	12	24
Triple-negative status		
No	37	71
Yes	15	29
Tumor phenotype		
Her2	12	23
Luminal	25	48
Triple-Negative	15	29
Adjuvant Chemotherapy		
No	13	25
Yes	39	75
Adjuvant Radiotherapy		
No	9	17
Yes	43	83
Adjuvant Hormonotherapy		
No	24	46
Yes	28	54

* Oestrogen and/or progesterone

Table 2: Clinical comparison of 52 patients between clusters

Clinical characteristic	PCA-K-means				Spectral Clustering				Sparse K-means				SIMLR				K-Sparse			
	C1 (N=21)	C2 (N=10)	C3 (N=21)	p-value	C2 (N=19)	C1 (N=12)	C3 (N=21)	p-value	C1 (N=24)	C2 (N=8)	C3 (N=20)	p-value	C1 (N=17)	C2 (N=12)	C3 (N=23)	p-value	C1(N=19)	C2 (N=12)	C3 (N=21)	p-value
Age [a]	62.7(15.2)	64.8(16)	62.9(15)	0.93	64.8(14.3)	62.5(16.5)	62 (15.3)	0.8	64.1(15)	60.5 (17.2)	63 (14.9)	0.85	64.3(14.1)	64.9 (16.1)	61.4 (15.6)	0.755	64.8(14.3)	62.5(16.5)	62(15.3)	0.827
Histology type				1				0.392				0.106				0.752				0.392
Ductal carcinoma	19(90.5)	10(100)	19(90.5)		17(89.5)	11(91.7)	20(95.2)		21(87.5)	7(87.5)	20(100)		15(88.2)	12(100)	21(91.3)		17(89.5)	11(91.7)	20(95.2)	
Lobular carcinoma	2(9.5)	0(0)	1(4.8)		2(10.5)	1(8.3)	0(0)		3(12.5)	0(0)	0(0)		2(11.8)	0(0)	1(4.3)		2(10.5)	1(8.3)	0(0)	
Microinvasive carcinoma	0(0)	0(0)	1(4.8)		0(0)	0(0)	1(4.8)		0(0)	1(12.5)	0(0)		0(0)	0(0)	1(4.3)		0(0)	0(0)	1(4.8)	
Tumor stage				**0.005**				**0.018**				0.063				**0.045**				**0.018**
T1	14(66.7)	3(30)	4(19)		12(63.2)	5(41.7)	4(19)		14(58.3)	2(25)	5(25)		10(58.8)	6(50)	5(21.7)		12(63.2)	5(41.7)	4(19)	
T2/T3	7(33.3)	7(70)	17(81)		7(36.8)	7(58.3)	17(81)		10(41.7)	6(75)	15(75)		7(41.2)	6(50)	18(78.3)		7(36.8)	7(58.3)	17(81)	
Axillary lymph node				0.162				0.075				0.526				0.387				0.075
N0	14(66.7)	6(60)	8(38.1)		14(73.7)	6(50)	8(38.1)		15(62.5)	4(50)	9(45)		11(64.7)	7(58.3)	10(43.5)		14(73.7)	6(50)	8(38.1)	
N+	7(33.3)	4(40)	13(61.9)		5(26.3)	6(50)	13(61.9)		9(37.5)	4(50)	11(55)		6(35.3)	5(41.7)	13(56.5)		5(26.3)	6(50)	13(61.9)	
Metastasis				0.667				1				1				0.497				1
M0	21(100)	10(100)	19(90.5)		18(94.7)	12(100)	20(95.2)		23(96)	8(100)	19(95)		17(100)	12(100)	21(86.9)		18(94.7)	12(100)	20(95.2)	
M1	0(40)	0(0)	2(9.5)		1(5.3)	0(0%)	1(4.8)		1(4)	0(0%)	1(5)		0(0%)	0(0%)	2(13.1)		1(5.3)	0(0)	1(50)	
Histological grade				0.109				**0.025**				**0.008**				**0.007**				**0.025**
I/II	13(61.9)	7(70)	7(35)		12(63.2)	9(75)	6(30)		15(62.5)	5(71.4)	7(35)		11(64.7)	9(75)	7(31.8)		12(63.2)	9(75)	6(30)	
III	8(38.1)	3(30)	13(75)		7(36.8)	3(25)	14(70)		9(37.5)	2(28.6)	13(65)		6(35.3)	3(25)	15(68.2)		7(36.8)	3(25)	14(70)	
Mitosis				**0.024**				**0.016**				0.133				**0.005**				**0.016**
1	11(52.4)	4(40)	2(10)		10 (52.6)	5 (41.7)	2 (10)		11 (45.8)	2 (28.6)	4 (20)		10 (58.8)	5 (41.7)	2 (9.1)		10 (52.6)	5 (41.7)	2 (10)	
2	3(14.3)	4(40)	7(35)		3 (15.8)	5 (41.7)	6 (30)		4 (16.7)	4 (57.1)	6 (30)		2 (11.8)	5 (41.7)	7 (31.8)		3 (15.8)	5 (41.7)	6 (30)	
3	7(33.3)	2(20)	11(55)		6 (31.6)	2 (16.7)	10 (60)		9 (37.5)	1 (14.3)	10 (50)		5 (29.4)	2 (16.7)	13 (59.1)		6 (31.6)	2 (16.7)	12 (60)	
Ki67 [a]	25(5,100)	27.5(10,90)	60(10,90)	0.066	41.1 (30.6)	33(22.6)	58.8 (27.2)	**0.027**	30 (19.2,80)	35 (23.8,45)	60 (28.8,90)	0.196	38 (31)	32.8 (22.7)	59.7 (25.9)	**0.009**	41.1 (30.6)	33 (22.6)	58.8(27.2)	**0.027**
Tumour phenotype				**0.024**				**0.012**				**0.006**				**0.018**				**0.012**
Her-2 over-expressed	1(4.8)	4(40)	7(33.3)		1(5.3)	4(33.3)	7(33.3)		2(8.3)	4(50)	6(30)		1(5.9)	4(33.3)	7(30.4)		1(5.3)	4(33.3)	7(33.3)	
Luminal	14(66.7)	5(50)	6(28.6)		13(68.4)	7(58.3)	5(23.8)		16(66.7)	4(50)	5(25)		12(70.6)	7(58.3)	6(26.1)		13(68.4)	7(58.3)	5(23.8)	
Triple-Negative	6(28.6)	1(10)	8(38.1)		5(26.3)	1(8.3)	9(42.9)		6(25)	0(0)	9(45)		4(23.5)	1(8.3)	10(43.5)		5(26.3)	1(8.3)	9(42.9)	
Hormonal receptors status				0.178				0.075				0.112				0.071				0.075
Negative	7(33.3)	5(50)	13(61.9)		6(31.6)	5(41.7)	14(66.7)		8(33.3)	4(50)	13(65)		5(29.4)	5(41.7)	15(65.2)		6(31.6)	5(41.7)	14(66.7)	
Positive	14(66.7)	5(50)	7(38.1)		13(68.4)	7(58.3)	7(33.3)		16(66.7)	4(50)	7(35)		12(70.6)	7(58.3)	8(34.8)		13(68.4)	7(58.3)	7(33.3)	
Her-2 status				**0.028**				0.061				**0.031**				0.115				0.061
Non-over-expressed	20(95.2)	6(60)	13(66.7)		18(94.7)	8(66.7)	14(66.7)		22(91.7)	4(50)	14(70)		16(94.1)	8(66.7)	16(69.6)		18(94.7)	6(66.7)	14(66.7)	
Over-expressed	1(4.8)	5(40)	6(33.3)		1(5.3)	4(33.3)	7(33.3)		2(8.3)	4(50)	6(30)		1(5.9)	4(33.3)	7(30.4)		1(5.3)	4(33.3)	7(33.3)	
Triple-Negative status				0.272				0.104				0.051				0.087				0.104
No	15(71.4)	9(90)	13(61.9)		14(73.7)	11(91.7)	12(57.1)		18(75)	8(100)	11(55)		13(76.5)	11(91.7)	13(56.5)		14(73.7)	11(91.7)	12(57.1)	
Yes	6(28.6)	1(10)	8(38.1)		5(26.3)	1(8.3)	9(42.9)		6(25)	0(0)	9(45)		4(23.5)	1(8.3)	10(43.5)		5(26.3)	1(8.3)	9(42.9)	
Luminal				**0.047**				**0.014**				**0.018**				**0.015**				**0.014**
No	7(33.3)	5(50)	15(71.4)		6(31.6)	5(41.7)	16(76.2)		8(33.3)	4(50)	15(75)		5(29.4)	5(41.7)	17(73.9)		6(31.6)	5(41.7)	16(76.2)	
Yes	14(66.7)	5(50)	6(28.6)		13(68.4)	7(58.3)	5(23.8)		16(66.7)	4(50)	5(25)		12(70.6)	7(58.3)	6(26.1)		13(68.4)	7(58.3)	5(28.8)	
Adjuvant Chemotherapy				0.52				0.423				0.459				0.459				0.423
No	7(33.3)	3(30)	4(19)		7(36.8)	2(16.7)	4(19)		6(25)	2(25)	5(25)		6(35.3)	3(25)	4(17.4)		7(36.8)	2(16.7)	4(19)	
Yes	14(85.7)	7(70)	17(81)		12(63.2)	10(83.3)	17(81)		18(75)	6(75)	1575)		11(64.7)	9(75)	19(82.6)		12(63.2)	10(83.3)	17(81)	
Adjuvant Radiotherapy				0.561				0.803				0.69				1				0.803
No	3(14.3)	3(30)	3(14.3)		3(15.8)	3(25)	3(14.3)		3(12.5)	2(25)	4(20)		3(17.6)	2(16.7)	4(17.4)		3(15.8)	3(25)	3(14.3)	
Yes	18(85.7)	7(70)	18(85.7)		16(84.2)	9(75)	18(85.7)		21(87.5)	6(75)	16(80)		14(82.4)	10(83.3)	19(82.6)		16(84.2)	9(75)	18(85.7)	

C1: cluster 1; C2: cluster 2; C3: cluster 3; [a]: mean (sd) or median (min, max),

Table 3: Comparison of prediction for overall and specific survival between clusters at 5 and 10-year

		Predict 5-year				Predict 10-year			
		Overall Survival		Specific Survival		Overall Survival		Specific Survival	
Methods	**No. of patients**	**% [95% CI]**	**P-value**		**P-value**		**P-value**		**P-value**
K-sparse			**0.021**		**0.002**		0.077		**0.004**
	Cluster 1 (n = 19)	77% [67 - 82]		87% [80 - 91]		58% [48 - 65]		80% [73 - 86]	
	Cluster 2 (n = 12)	71% [57 - 82]		81% [69 - 90]		53% [38 - 66]		75% [60 - 85]	
	Cluster 3 (n = 20)	59% [47 - 69]		68% [60 - 74]		41% [29 - 52]		62% [53 - 69]	
SIMLR			0.1		**0.011**		0.241		**0.009**
	Cluster 1 (n = 17)	75% [64 - 82]		85% [77 - 91]		55% [45 - 64]		77% [65 - 84]	
	Cluster 2 (n = 12)	72% [56 - 82]		83% [69 - 91]		55% [40 - 67]		79% [65 - 87]	
	Cluster 3 (n = 22)	61% [50 -70]		71% [63 - 77]		43% [32 - 53]		64% [55 - 70]	
Sparse K-means			**0.049**		**0.027**		0.203		**0.024**
	Cluster 1 (n = 24)	74% [64 - 80]		84% [76 - 89]		54% [43 - 63]		80% [73 - 86]	
	Cluster 2 (n = 7)	72% [58 - 87]		83% [70 - 94]		56% [37 - 72]		75% [60 - 85]	
	Cluster 3 (n = 20)	61% [49 - 69]		70% [61 - 78]		42% [32 - 52]		62% [53 - 69]	
Spectral clustering			**0.021**		**0.002**		0.077		**0.004**
	Cluster 1 (n = 19)	77% [68 - 83]		77% [80 - 91]		58% [48 - 65]		82% [73 - 86]	
	Cluster 2 (n = 12)	71% [57 - 81]		71% [69 - 90]		52% [32 - 64]		75% [60 - 85]	
	Cluster 3 (n = 20)	59% [47 - 68]		69% [60 - 76]		41% [29 - 52]		62% [53 - 69]	
PCA K-means			0.055		**0.009**		0.085		**0.008**
	Cluster 1 (n = 21)	77% [67 - 81]		86% [79 - 91]		58% [48 - 65]		79% [71 - 85]	
	Cluster 2 (n = 10)	69% [53 - 81]		80% [66 - 90]		52% [32 - 64]		77% [63 - 86]	
	Cluster 3 (n = 20)	60% [47 - 69]		69% [61 - 78]		41% [29 - 52]		63% [54 - 70]	

Table 4: Table indicating which metabolites are in each intersection or are unique to a certain list

	Clustering Methods	Nbr	Metabolites
5	K-Sparse PCA K-means SIMLR Sparse K-means Spectral clustering	2	Creatine; L-Proline ;
4	K-Sparse SIMLR Sparse K-means Spectral clustering	1	Triethanolamine ;
	K-Sparse PCA K-means SIMLR Sparse K-means	2	L-Methionine ; L-Phenylalanine
	K-Sparse PCA K-means Sparse K-means Spectral clustering	2	L-Carnitine ; Betaine ;
	PCA K-means SIMLR Sparse K-means Spectral clustering	4	Glutathione; Isoleucyl-Methionine; Humulinic acid A; Alnustone;
3	K-Sparse SIMLR Sparse K-means	1	Hydroxyprolyl-Valine ;
	K-Sparse PCA K-means Sparse K-means	20	Aminoadipic acid; Methylmalonic acid; 1b-Furanoeudesm-4(15)-en-1-ol acetate; Glycerophosphocholine; Lidocaine; Adenosine monophosphate ; 2-Methyl-3-ketovaleric acid; Liqcoumarin; p-Cresol sulfate; 2-Methylbutyroylcarnitine; Methoxsalen; Citramalic acid; Hypoxanthine; L-Acetylcarnitine; Ethyl aconitate; Guanine; L-Glutamic acid; Uridine 5'-monophosphate; N1,N12-Diacetylspermine; 5-Aminoimidazole ribonucleotide;
	SIMLR Sparse K-means Spectral clustering	4	2,5-Dichloro-4-oxohex-2-enedioate; Histidinyl-Isoleucine; 3-(4-Methyl-3-pentenyl)thiophene; (-)-Epigallocatechin
	PCA K-means Sparse K-means Spectral clustering	3	L-Isoleucine ; Ascorbic acid ; Neurine ;
2	K-Sparse Sparse K-means	3	5-Hydroxyisourate ; Hexanoylcarnitine ; L-Glutamine ;
	K-Sparse PCA K-means	9	Creatinine; Proline; betaine; Erythronic acid; Garcinia acid; Thiolutin; 4-Chloro-1H-indole-3-acetic acid; Niacinamide 3-Dehydroxycarnitine; Dihydrothymine;
	SIMLR Spectral clustering	21	5b-Cyprinol sulfate; 2',4-Dihydroxy-4',6'-dimethoxychalcone; Propenoylcarnitine; 5-Hydroxyindoleacetic acid; Phaseolic acid Lisuride; 2-Bromophenol; (alpha-D-mannosyl)7-beta-D-mannosyl-diacetylchitobiosyl-L-asparagine isoform B (protein); Plastoquinone 3; 2,2,4,4,-Tetramethyl-6-(1-oxopropyl)-1,3,5-cyclohexanetrione; 1-Pyrroline; Gingerol; Prehumulinic acid; 1-Methylpyrrolo[1,2-a]pyrazine; 5-(methylthio)-2,3-Dioxopentyl phosphate; Propionic acid; Isosakuranin; Phenmetrazine; Methionine sulfoxide; Glycerol; Carboxyphosphamide;
	SIMLR Sparse K-means	1	Phosphoric acid ;
	PCA K-means Sparse K-means	4	I(-) ; L-Tyrosine ; Gravelliferone ; Valganciclovir ;
1	K-Sparse	10	Prolylhydroxyproline; Guanidoacetic acid; Histamine; PC-M6; L-Histidine; N-Acetyl-L-aspartic acid; 3-Mercaptohexyl hexanoate; Trimethylamine N-oxide; Pantothenic acid; Flunitrazepam;
	SIMLR	14	3-Hydroxy-6,8-dimethoxy-7(11)-eremophilen-12,8-olide; Glycerol tripropanoate; Alanyl-Isoleucine; 1-(2,4,6-Trimethoxyphenyl)-1,3-butanedione; 1-Oxo-1H-2-benzopyran-3-carboxaldehyde; 1,3,11-Tridecatriene-5,7,9-triyne; N-Acetyl-L-methionine; 3-Methyl sulfolene; 5-(4-Acetoxy-3-oxo-1-butynyl)-2,2'-bithiophene; Ac-Ser-Asp-Lys-Pro-OH; Cyclic AMP; Benzothiazole; (±)-2-Methylthiazolidine; 2-Methylcitric acid;
	Spectral clustering	13	2,3-diketogulonate; 2,5-Furandicarboxylic acid; Pyrrolidine; Piperidine; Beta-Alanine; Aspartyl-L-proline; Erythro-5-hydroxy-L-lysinium(1+); Acrylamide; 5-Hydroxylysine; S-Nitrosoglutathione; 2,2-dichloro-1,1-ethanediol; Valerenic acid; Dichloromethane;
	Sparse K-means	3	Erinapyrone C; Ergothioneine; N-Methylethanolaminium phosphate;
	PCA K-means	4	Dimethylglycine; Pipecolic acid; Methyl (9Z)-10'-oxo-6,10'-diapo-6-carotenoate; N-Desmethylvenlafaxine

Table 5: List of significant relevant pathways identified by 5 methods

Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd[a]	Match Status[b]	Raw *P*[c]	*-log*(p)	Impact[d]
K-Sparse method							
C1 vs C3	UDP - glucose	Starch and sucrose metabolism	50	1	0,0107	4,5388	0,1390
	UDP - glucose	Amino sugar and nucleotide sugar metabolism	88	1	0,0107	4,5388	0,0928
	UDP - glucose; Glyceric acid	Glycerolipid metabolism	32	2	0,0153	4,1831	0,0206
SIMLR method							
Clusters Comparaison	**Interaction metabolite**	**Pathway Name**	**Total Cmpd**	**Match Status**	**P Value**	***-log*(p)**	**Impact**
C1 VS C2	Glutathione; Oxidized glutathione; Glycine; L-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine;	Glutathione metabolism	38	12	0	12,826	0,3628
	Cadaverine; Aminopropylcadaverine; Ascorbic acid;						
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0	12,469	0,1383
	L-Tryptophan; N-Acetylserotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; L-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine;	Tryptophan metabolism	79	8	0,0001	9,1233	0,2741
	5'-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0008	7,1674	0,2509
	L-Glutamine; Phosphoribosylformylglycineamidine; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; 5-Hydroxyisourate; Guanosine; Adenosine diphosphate ribose; 5-Aminoimidazole ribonucleotide; Glyoxylic acid; Glycine; Adenosine 3',5'-diphosphate;	Purine metabolism	92	17	0,0011	6,8091	0,2048
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0027	5,9281	0,268
	L-Glutamine; Ornithine; Citrulline; L-Arginine; L-Glutamic acid; N-Acetylornithine; L-Proline; Hydroxyproline; Guanidoacetic acid; Creatine; 4-Guanidinobutanoic acid; N2-Succinyl-L-ornithine; Putrescine; Spermidine; N-Acetylputrescine; Pyruvic acid; Glyoxylic acid; Spermine;	Arginine and proline metabolism	77	19	0,0053	5,238	0,6514
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,0075	4,8991	0,176
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0076	4,8821	0,0394
	2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0154	4,1754	0,0324
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; Pyruvic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,018	4,0154	0,46986
	Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; N-Acetyl-D-Glucosamine 6-Phosphate; Uridine diphosphate-N-acetylglucosamine ; Cytidine monophosphate N-acetylneuraminic acid; D-Glucose; D-Xylose	Amino sugar and nucleotide sugar metabolism	88	7	0,0187	3,9783	0,1417
	Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0412	3,1903	0,3705

	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0412	3,1898	0,0773
C1 VS C3	Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0139	4,2752	0,3705
	Phenylpyruvic acid; L-Phenylalanine; L-Tyrosine; 3-Dehydroquinate; L-Tryptophan;	Phenylalanine, tyrosine and tryptophan biosynthesis	27	5	0,0189	3,9687	0,099
	L-Tryptophan; N-Acetylserotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; L-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine;	Tryptophan metabolism	79	8	0	16,409	0,2741
C2 VS C3	Glutathione; Oxidized glutathione; Glycine; L-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine;	Glutathione metabolism	38	12	0	16,133	0,3628
	Cadaverine; Aminopropylcadaverine; Ascorbic acid;						
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid	Ascorbate and aldarate metabolism	45	5	0	13,096	0,1383
	5'-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0001	9,8548	0,2509
	Phenylpyruvic acid; L-Phenylalanine; L-Tyrosine; 3-Dehydroquinate; L-Tryptophan;	Phenylalanine, tyrosine and tryptophan biosynthesis	27	5	0,0001	8,9814	0,099
	L-Histidine; L-Phenylalanine; L-Arginine; L-Glutamine; Glycine; L-Methionine; L-Lysine; L-Isoleucine; L-Threonine; L-Tryptophan; L-Tyrosine; L-Proline; L-Glutamic acid; Phosphoserine;	Aminoacyl-tRNA biosynthesis	75	14	0,0002	8,758	0,1127
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0004	7,7271	0,268
	L-Glutamine; Phosphoribosylformylglycineamidine; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; 5-Hydroxyisourate; Guanosine; Adenosine diphosphate ribose; 5-Aminoimidazole ribonucleotide; Glyoxylic acid; Glycine; Adenosine 3',5'-diphosphate;	Purine metabolism	92	17	0,0007	7,306	0,2048
	Malonic acid; Beta-Alanine; Spermine; Spermidine; Dihydrouracil; Pantothenic acid; Uracil; L-Histidine	beta-Alanine metabolism	28	8	0,0012	6,7568	0,3577
	Uridine 5'-monophosphate; L-Glutamine; Dihydrouracil; Cytidine monophosphate; Cytidine; Cytosine; Uracil; Dihydrothymine; Uridine diphosphate glucose; Malonic acid; Ureidosuccinic acid; Beta-Alanine; Methylmalonic acid;	Pyrimidine metabolism	60	13	0,0014	6,5817	0,2756
	Pantothenic acid; Dihydrouracil; Beta-Alanine; Pyruvic acid; Adenosine 3',5'-diphosphate; Uracil;	Pantothenate and CoA biosynthesis	27	6	0,0023	6,0879	0,2736
	L-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; L-Tyrosine;	Phenylalanine metabolism	45	6	0,0072	4,9364	0,2468
	L-Glutamic acid; L-Glutamine; Oxoglutaric acid	D-Glutamine and D-glutamate metabolism	11	3	0,0124	4,39	0,139
	L-Glutamine; Ornithine; Citrulline; L-Arginine; L-Glutamic acid; N-Acetylornithine; L-Proline; Hydroxyproline; Guanidoacetic acid; Creatine; Creatinine; 4-Guanidinobutanoic acid; N2-Succinyl-L-ornithine; Putrescine; Spermidine; N-Acetylputrescine; Pyruvic acid; Glyoxylic acid; Spermine;	Arginine and proline metabolism	77	19	0,0169	4,082	0,6514
	2-Hydroxyethanesulfonate ; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0215	3,8411	0,0324

	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,0221	3,8108	0,4122
	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0267	3,6235	0,0773
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid	Citrate cycle (TCA cycle)	20	3	0,0302	3,5015	0,176
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,0372	3,2914	0,4699
	Uridine diphosphate glucose; Glycerol 3-phosphate; Glycerol; Glyceric acid; Galactosylglycerol;	Glycerolipid metabolism	32	5	0,0427	3,1546	0,2162
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0427	3,1536	0,0394
	Sparse K-means method						
Clusters Comparaison	**Interaction metabolite**		**Total Cmpd**	**Match Status**	**Raw p**	***-log*(p)**	**Impact**
C1 VS C2	L-Methionine; Glutathione	Cysteine and methionine metabolism	56	2	0.007	4.9	0.0454
C1 VS C3	L-Methionine; Glutathione;	Cysteine and methionine metabolism	56	2	0.0020	6.2	0.00454
	Spectral clustering method						
Clusters Comparaison	**Interaction metabolite**	**Pathway Name**	**Total Cmpd**	**Match Status**	**Raw p**	***-log*(p)**	**Impact**
C1 VS C3	Iminoaspartic acid; Quinolinic acid; Niacinamide; Pyruvic acid; Propionic acid;	Nicotinate and nicotinamide metabolism	44	5	0,0024	6,0206	0,0712
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,0040	5,5100	0,4699
	5'-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0098	4,6232	0,2509
	Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0101	4,5961	0,3705
	xoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,0171	4,0710	0,1760
	Pyruvic acid; L-Threonine; L-Isoleucine;	Valine, leucine and isoleucine biosynthesis	27	3	0,0178	4,0277	0,0350
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0210	3,8609	0,0394
	D-Glucose; Glyceric acid; Pyruvic acid;	Pentose phosphate pathway	32	3	0,0232	3,7622	0,0218
	Pyruvic acid; L-Lactic acid; D-Glucose;	Glycolysis or Gluconeogenesis	31	3	0,0249	3,6928	0,0953
	Pyruvic acid; L-Lactic acid;	Pyruvate metabolism	32	2	0,0274	3,5955	0,3201
	L-Glutamic acid; Pyruvic acid; Butyric acid; Oxoglutaric acid;	Butanoate metabolism	40	4	0,0283	3,5644	0,0852
	2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0287	3,5525	0,0324
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0303	3,4966	0,2680
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0,0330	3,4104	0,1383
	Epinephrine; Dopamine; L-Tyrosine; Homovanillic acid; Pyruvic acid;	Tyrosine metabolism	76	5	0,0385	3,2580	0,1750
	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,0390	3,2431	0,4122

	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0447	3,1074	0,0773
PCA K-means method							
Clusters Comparaison	**Interaction metabolite**	**Pathway Name**	**Total Cmpd**	**Match Status**	**Raw p**	***-log*(p)**	**Impact**
C1 vs C3	Iminoaspartic acid; Quinolinic acid; Niacinamide; Pyruvic acid; Propionic acid;	Nicotinate and nicotinamide metabolism	44	5	0,003	5,9412	0,0712
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,011	4,4865	0,1760
	Epinephrine; Dopamine; L-Tyrosine; Homovanillic acid; Pyruvic acid;	Tyrosine metabolism	76	5	0,024	3,7311	0,1750
	Pyruvic acid; L-Lactic acid;	Pyruvate metabolism	32	2	0,043	3,1507	0,3201
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate ;Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,044	3,1214	0,0394
	Pyruvic acid; L-Threonine; L-Isoleucine;	Valine, leucine and isoleucine biosynthesis	27	3	0,045	3,1107	0,0350
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0,045	3,0926	0,1383
	L-Glutamic acid; Pyruvic acid; Butyric acid; Oxoglutaric acid;	Butanoate metabolism	40	4	0,046	3,0843	0,0852
	D-Glucose; Glyceric acid; Pyruvic acid;	Pentose phosphate pathway	32	3	0,046	3,0769	0,0218
	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,048	3,0446	0,4122

[a] Total cmpd is the total number of compounds in the pathway

[b] Hits is the actual matched number from the uploaded data.

[c] Raw p is the original *p*-value calculated from the pathway analysis.

[d] Impact is the pathway impact value calculated from pathway topology analysis.

Supplementary Information

Comparison of unsupervised machine learning methods in order to identified metabolomic signatures in patients with localized breast cancer

Jocelyn Gal[1*‖] , Caroline Bailleux[2‖] , David Chardin[3,4‖] , Thierry Pourcher[4], Julia Gilhodes[5], Lun Jing[4], Jean-Marie Guigonis[4], Jean-Marc Ferrero[2], Gerard Milano[6], Baharia Mograbi[7], Patrick Brest[7], Yann Chateau[1], Olivier Humbert[3,4], Emmanuel Chamorey[1].

[1] University Côte d'Azur, Epidemiology and Biostatistics Department, Centre Antoine Lacassagne, Nice, F-06189, France
[2] University Côte d'Azur, Medical Oncology Department Centre Antoine Lacassagne, Nice, F-06189, France
[3] University Côte d'Azur, Nuclear medicine Department, Centre Antoine Lacassagne, Nice, F-06189, France
[4] University Côte d'Azur, Commissariat à l'Energie Atomique, Institut de biosciences et biotechnologies d'Aix-Marseille, Laboratory Transporters in Imaging and Radiotherapy in Oncology, Faculty of medicine, Nice, F-06100, France
[5] Department of Biostatistics, Institut Claudius Regaud, IUCT-O Toulouse, France.
[6] University Côte d'Azur, Centre Antoine Lacassagne, Oncopharmacology Unit, Nice, F-06189, France
[7] University Côte d'Azur, CNRS UMR7284, INSERM U1081, IRCAN TEAM4; Centre Antoine Lacassagne FHU-Oncoage, Nice, F-06189, France

[*]Corresponding author:
Mr Jocelyn Gal
Department of Epidemiology and Biostatistics, Centre Antoine Lacassagne, University Côte d'Azur
33 avenue de Valombrose
06189 Nice, France
Phone : +33-4-92-03-10-31
E-mail : jocelyn.gal@nice.unicancer.fr

[‖]*These authors contributed equally to this work*

Table of contents:

Supplementary File S1: Sample collection, preparation and liquid chromatography-mass spectrometry analysis.

Tumor samples were collected during breast surgery and quickly stored at -80°C until analysis in our facility's biobank. Freeze-dried samples were processed by methanol extraction and analyzed by liquid chromatography-mass spectrometry (LC-MS) for metabolite characterization [1]. Liquid chromatographic analysis was performed using a DIONEX Ultimate 3000 HPLC system (Thermo Fisher Scientific). 10µL of each sample was injected onto a Synergi 4µm Hydro-RP 80 Å, 250 x 3.0 mm column (Phenomenex, Le Pecq, France). The mobile phases were composed of 0.1% formic acid in water (A) and 0.1% formic acid in acetonitrile (B). The gradient was set as follows with a flow rate of 0.9 mL/min: 0% phase B from 0 to 5 min, 0-95% B from 5 to 21min, holding at 95% B to 21.5min, 95-0% B from 21.5 to 22min, holding at 0% B until 25min for column equilibration. Mass spectrometry analysis was carried out on a Q Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific) with a heated electrospray ionization source, HESI II, operating in both positive and negative modes. High-resolution accurate-mass full-scan MS and top 5 MS2 spectra were collected in a data-dependent fashion at a resolving power of 70 000 and 35 000 at m/z 400, respectively.

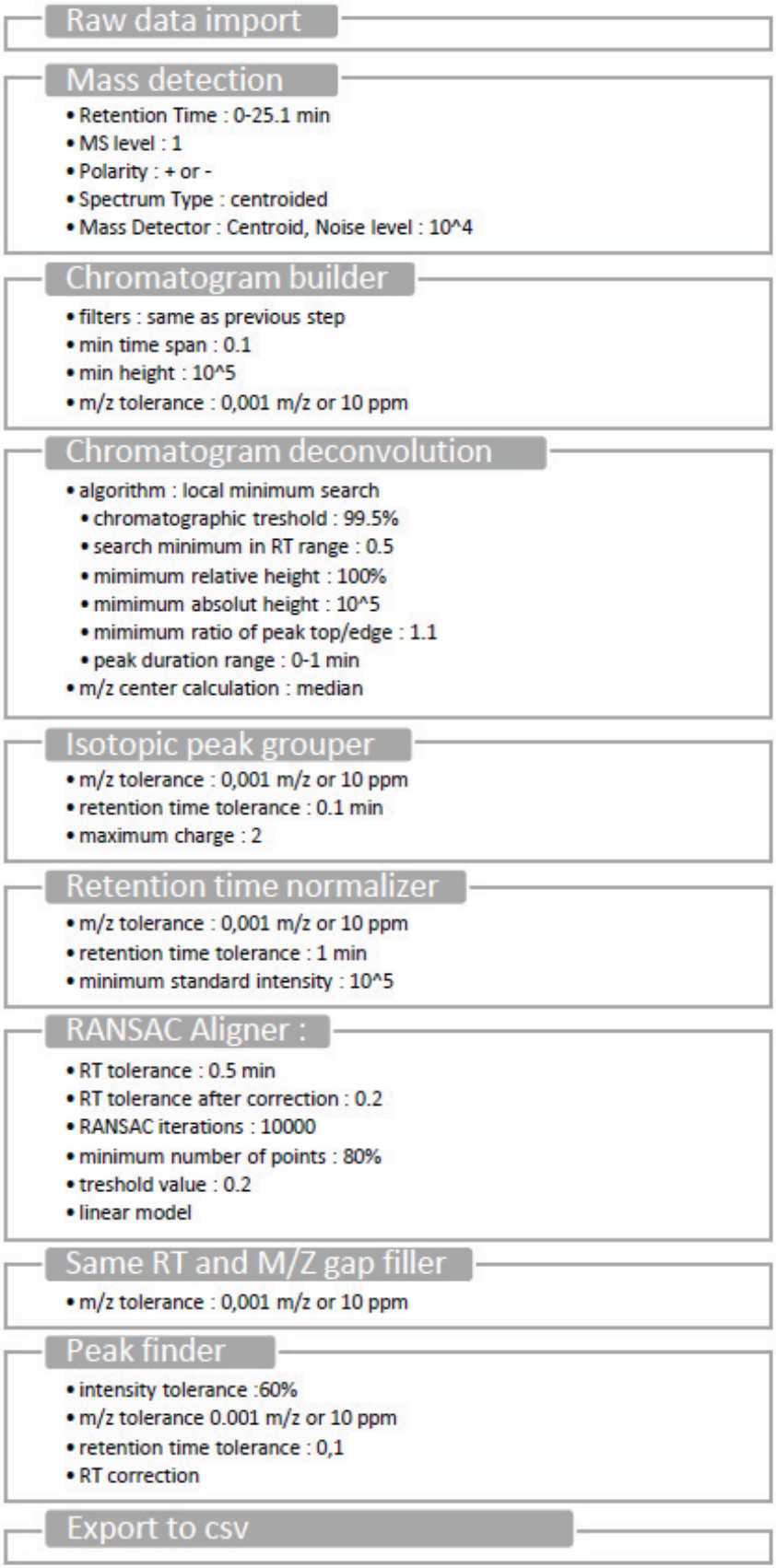
MSconvert (Version2.1, ProteoWizard) was used to convert raw data files obtained from LC-MS/MS to centroided mzXML files. The data collected from positive and negative ionization modes were analyzed separately using MzMine® (Version 2.38) [2, 3]. Isolated chromatograms were built for each mass with a noise threshold of 10^5. A local minimum search algorithm was used to select the validated peaks. Peaks were then aligned by RANSAC (random sample consensus) algorithm with a tolerance of 10 ppm in m/z and 1 min of retention time. Missing values were filled in using the same m/z and RT range as observed in detected samples, where possible. Only peaks with no missing values after gap-filling were kept. Peaks were then identified using the Human Metabolome DataBase (HMDB, version 3.0) by searching for $M+H^+$ and $M-H^+$ ion forms in positive and negative mode, respectivey, with 15ppm of mass tolerance. Linear normalization was performed using the average intensity in each sample as a normalization factor.

[1] Maharjan RP, Ferenci T. Global metabolite analysis: the influence of extraction methodology on metabolome profiles of Escherichia coli, Anal Biochem 2003;313:145-154.

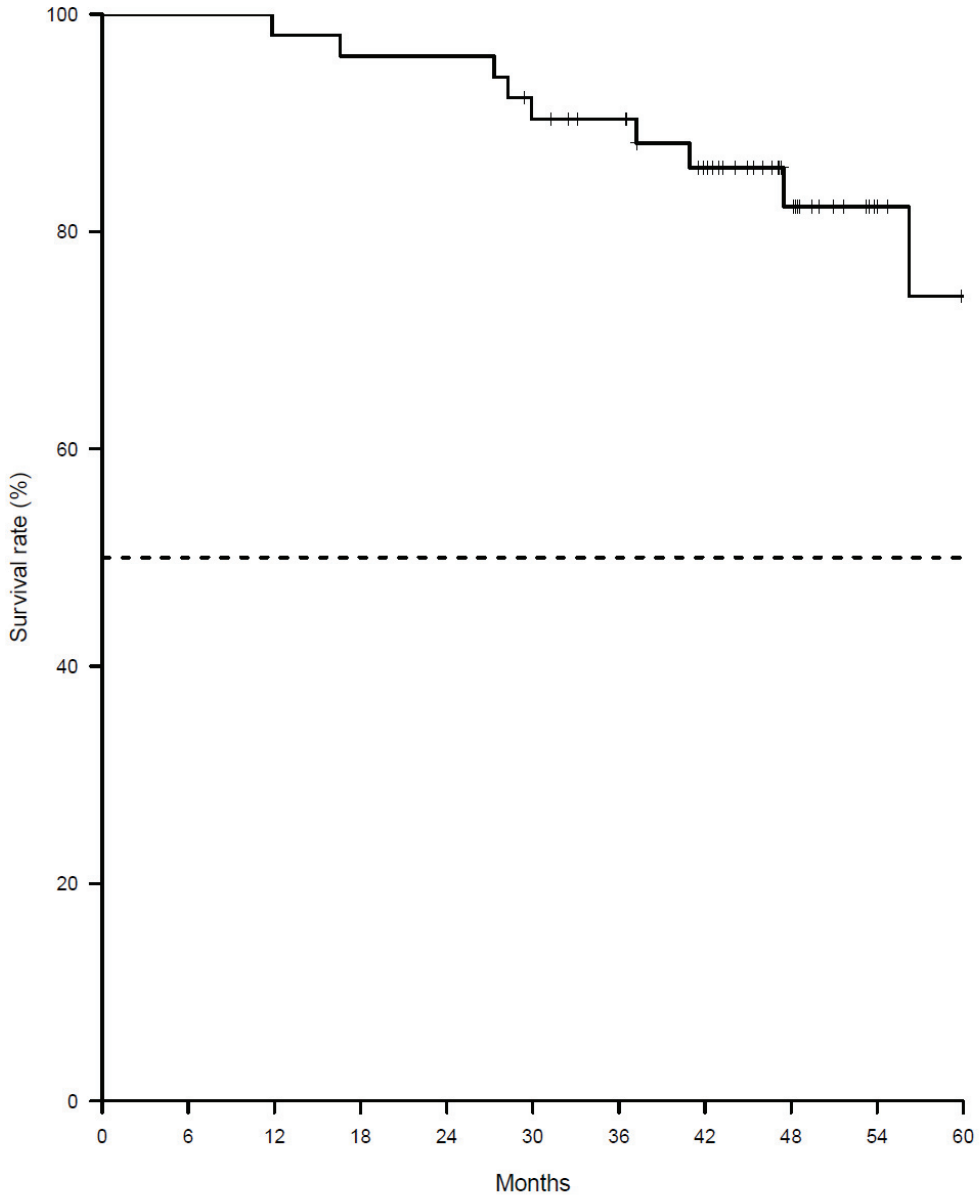
[2] Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data, BMC Bioinformatics 2005;6:179.

[3] Pluskal T, Castillo S, Villar-Briones A et al. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics 2010;11:395.

Supplementary Fig. S1: Protocol used in MZmine for the treatment of metabolomic data

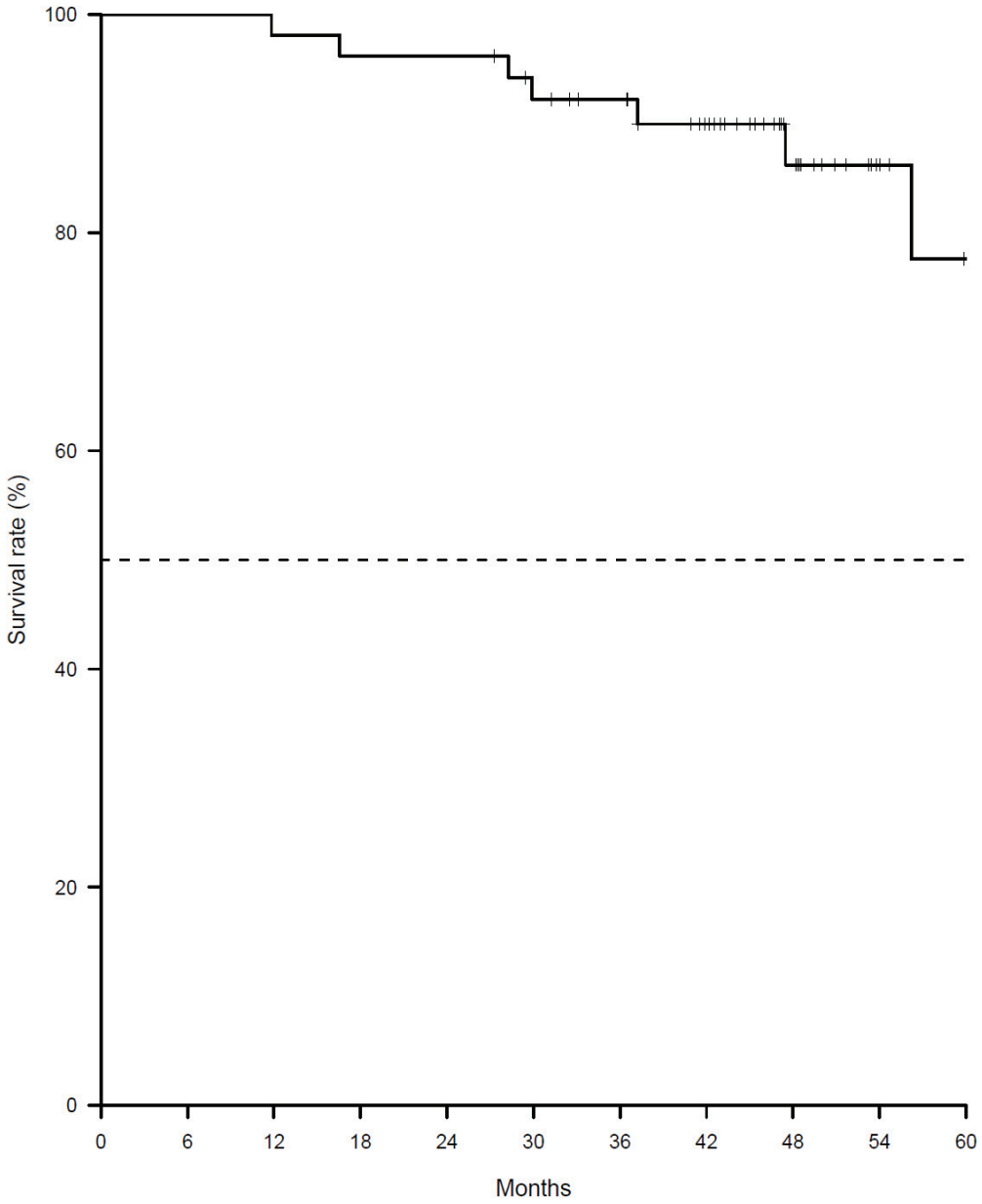


Supplementary Fig. S2a: Overall survival



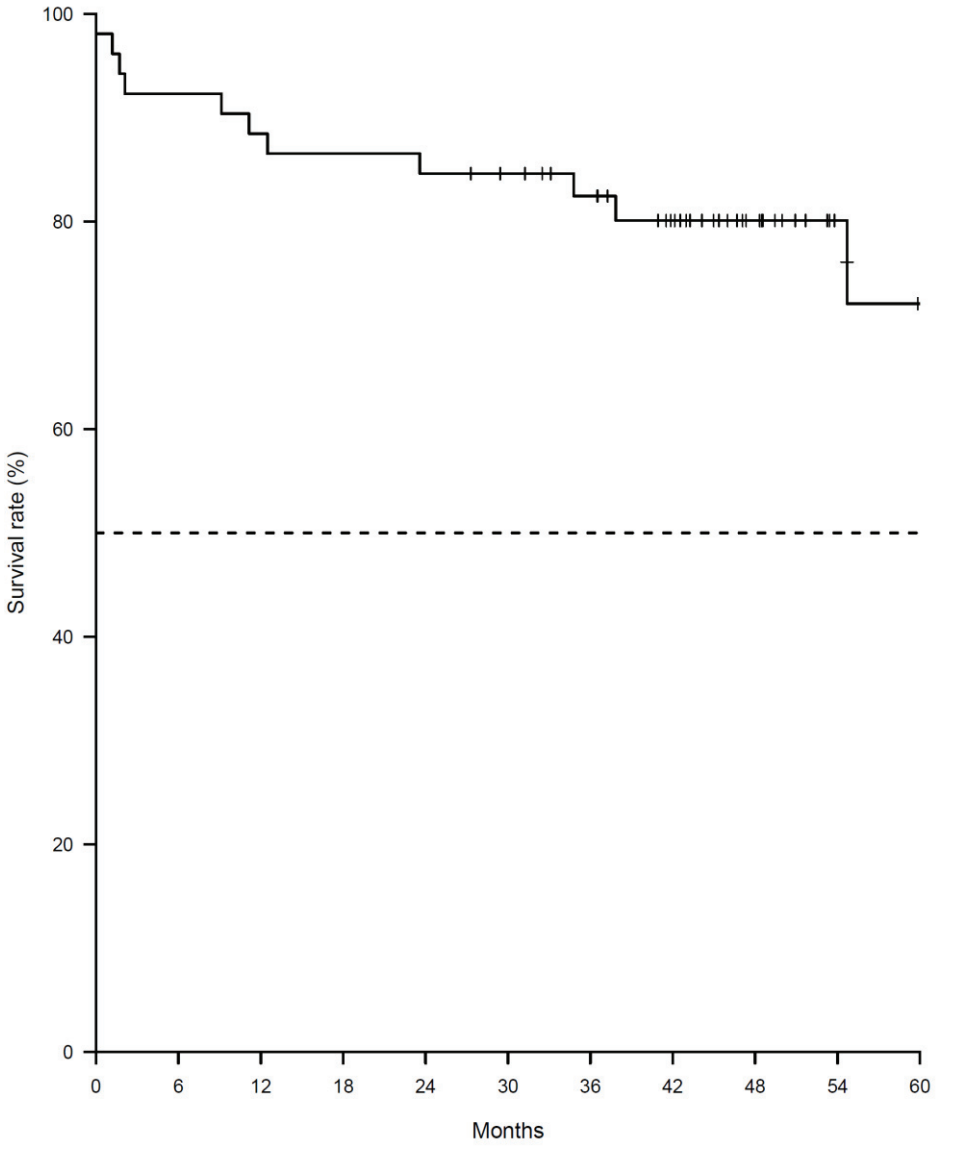
	Months					
time	0	12	24	36	48	60
n.risk	52	51	50	43	23	8
n.event	0	1	1	3	3	1
surv	100	98.1	96.2	90.3	82.3	74.1
IC 95% lower	100	94.4	91.1	82.6	71.5	57.7
IC 95% upper	100	100	100	98.8	94.7	95.7

Supplementary Fig. S2b: Specific survival



	Months					
time	0	12	24	36	48	60
n.risk	52	51	50	43	23	8
n.event	1	5	2	1	1	1
surv	100	98.1	96.2	92.2	86.2	77.6
IC 95% lower	100	94.4	91.1	85.1	76.1	60.9
IC 95% upper	100	100	100	99.8	97.7	98.8

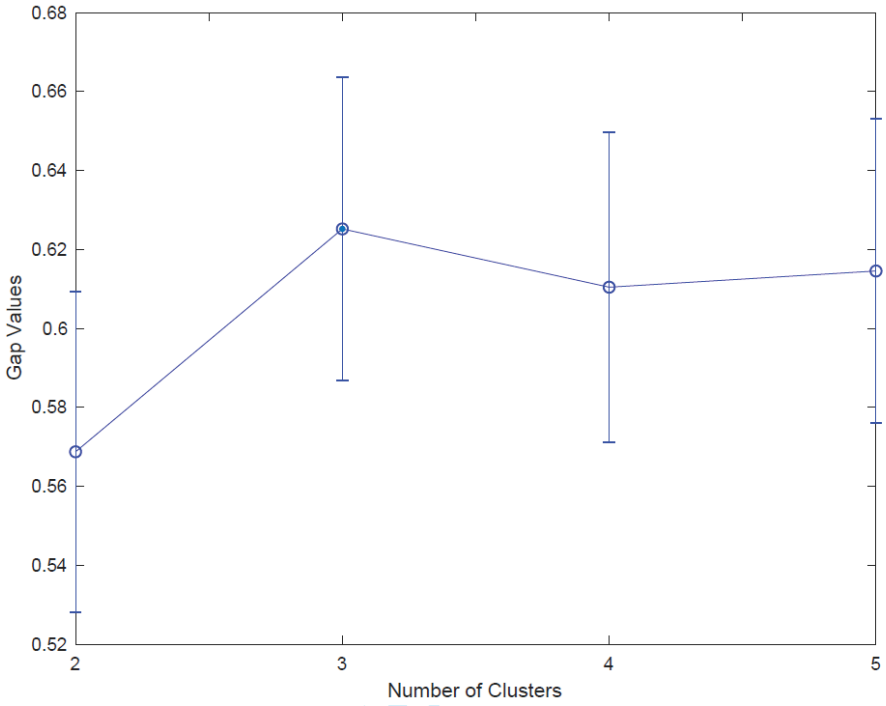
Supplementary Fig. S2c: Recurrence free survival



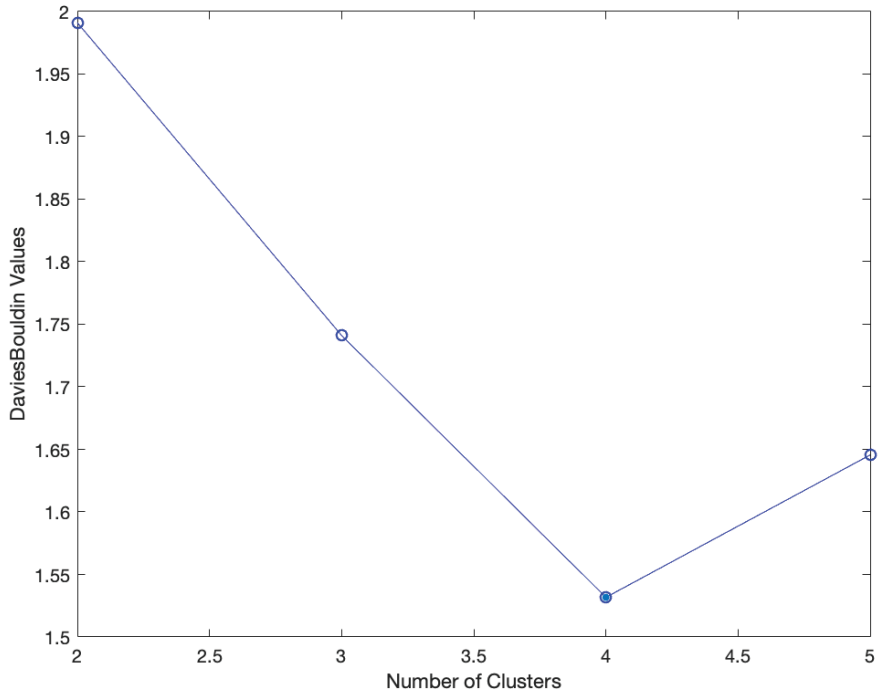
	Months					
time	0	12	24	36	48	60
n.risk	52	51	50	43	23	8
n.event	1	5	2	1	1	1
surv	98.1	85.5	84.6	82.4	80.1	72.1
IC 95% lower	94.4	80.2	75.4	72.6	69.7	56.2
IC 95% upper	100	97.6	95.0	93.6	92.0	92.4

Supplementary Fig. S3: Estimate optimal number of clusters.

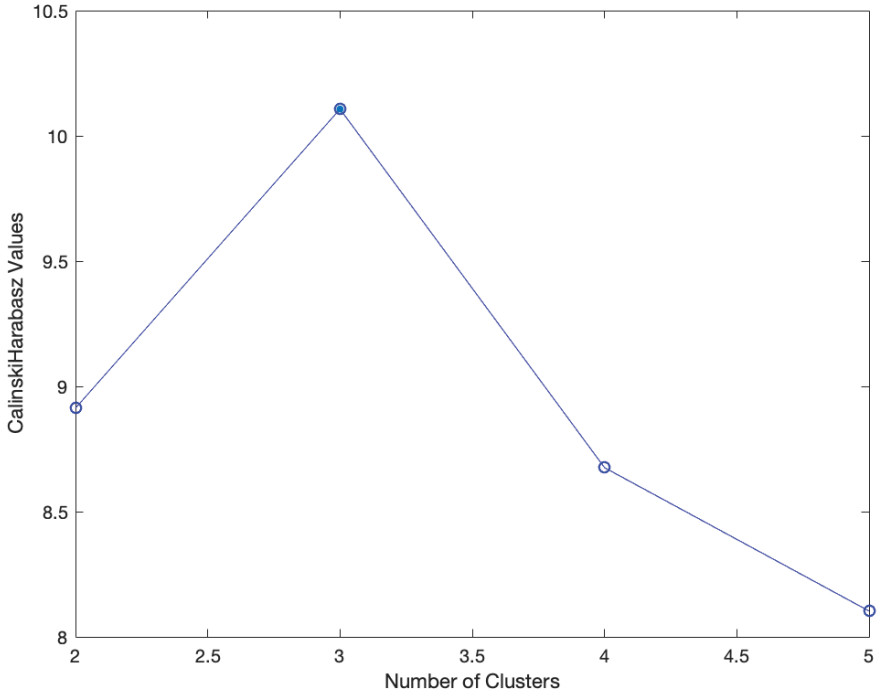
A: Gap statistic criterion



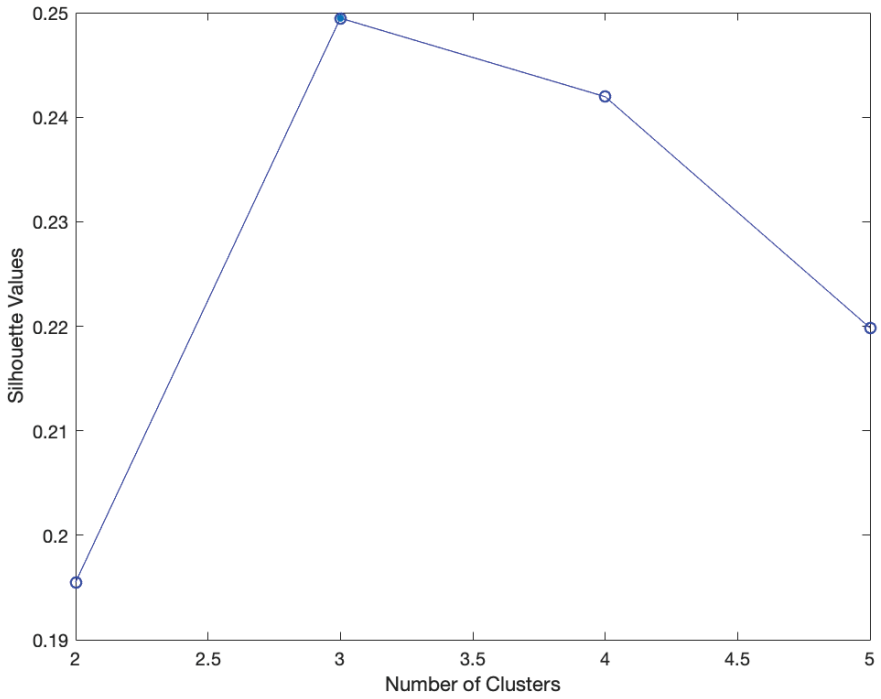
B: Davies-Bouldin criterion



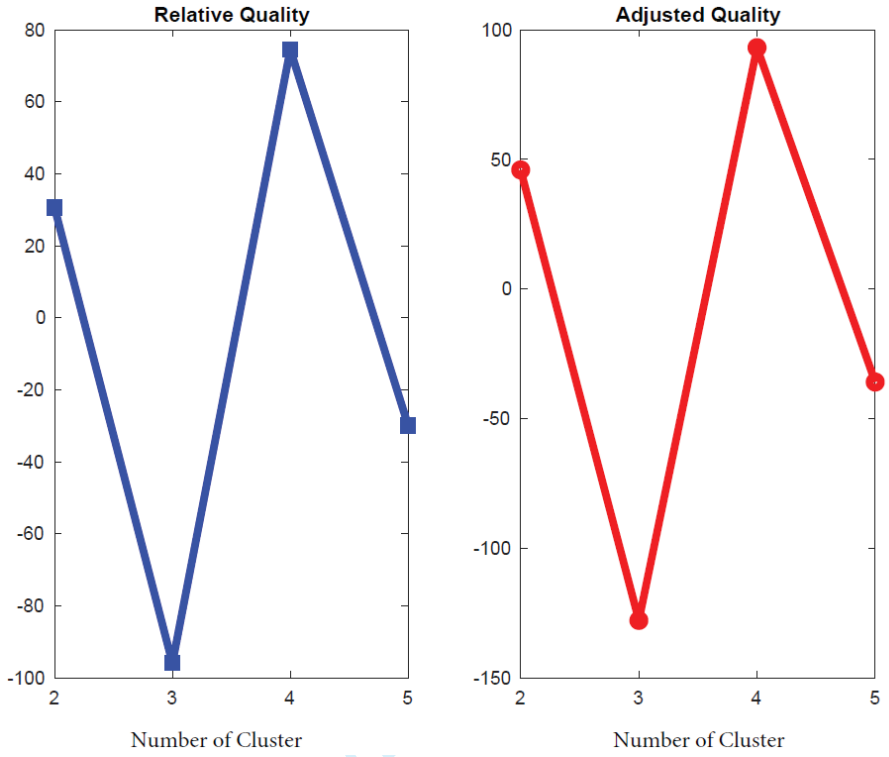
C: Calinski-Harabasz criterion



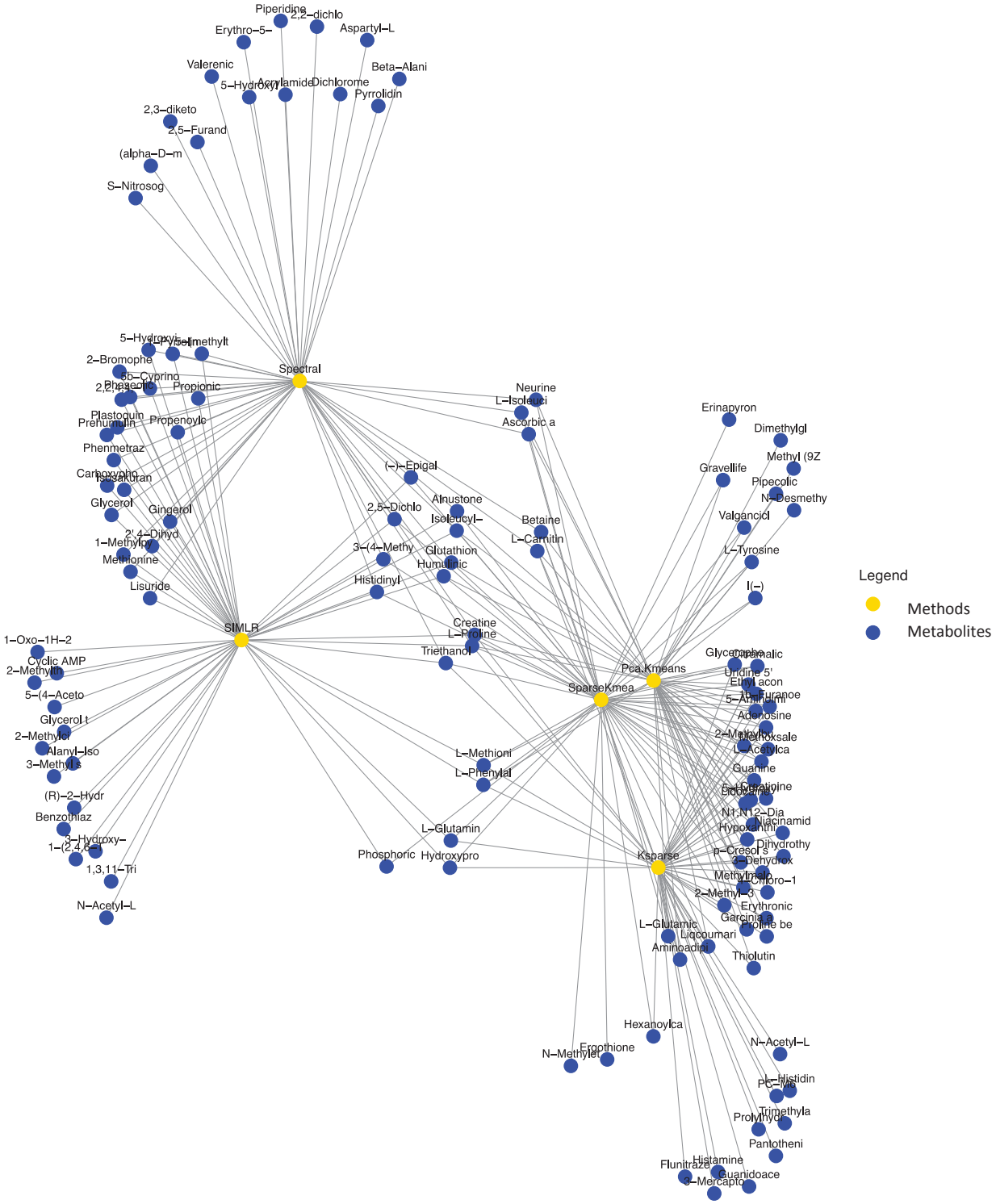
D: Silhouette criterion



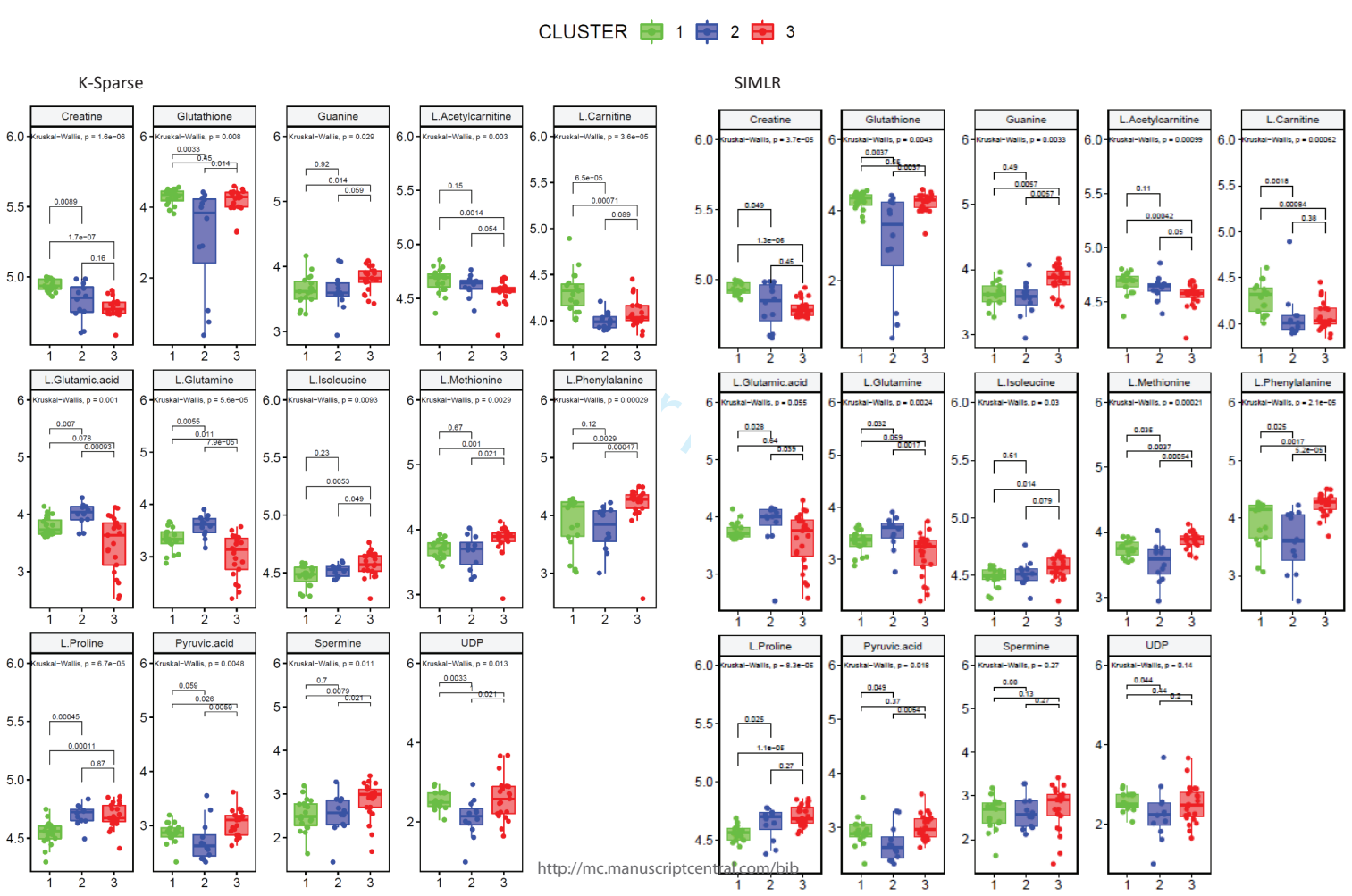
E: SIMLR criterion

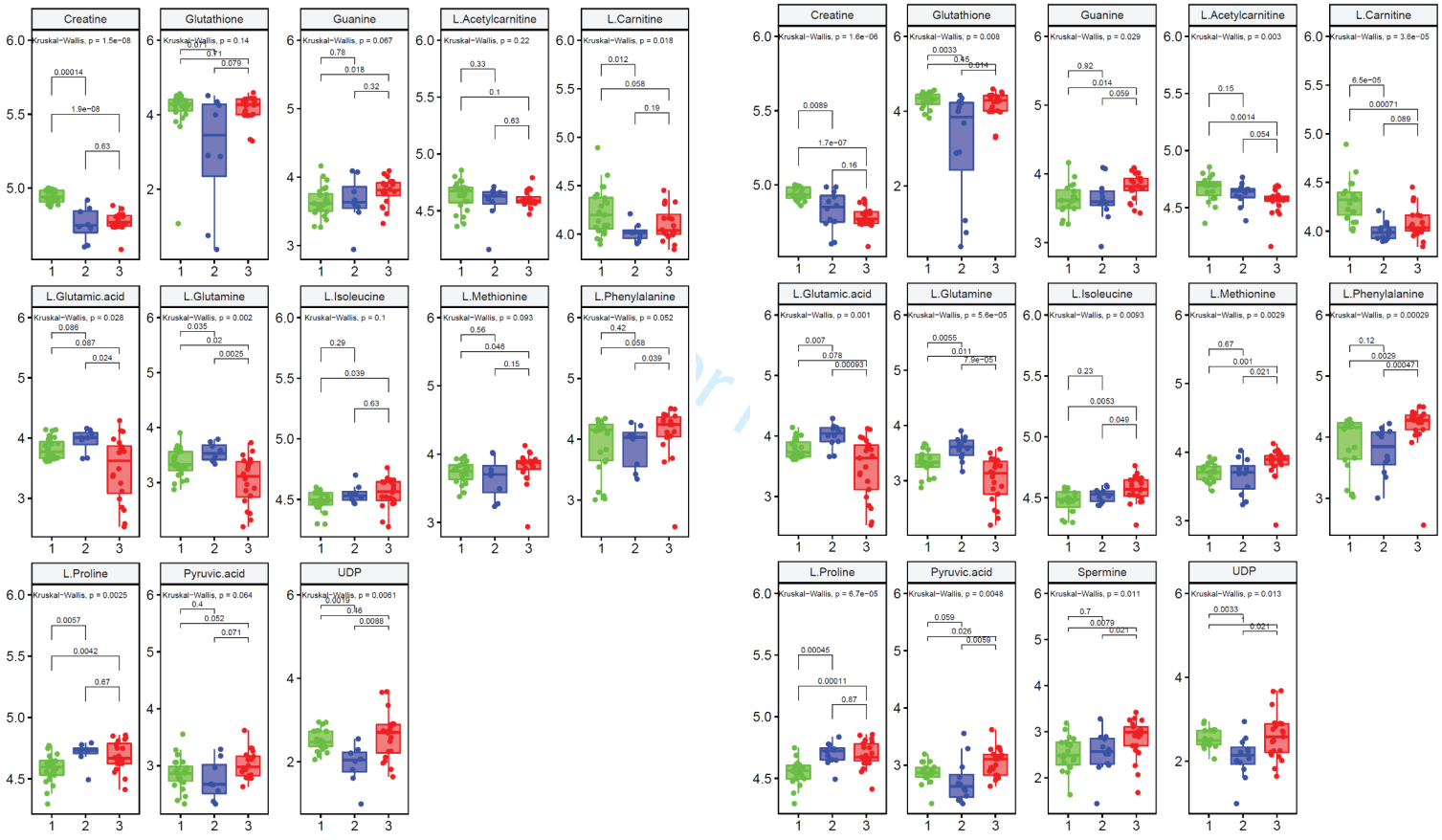


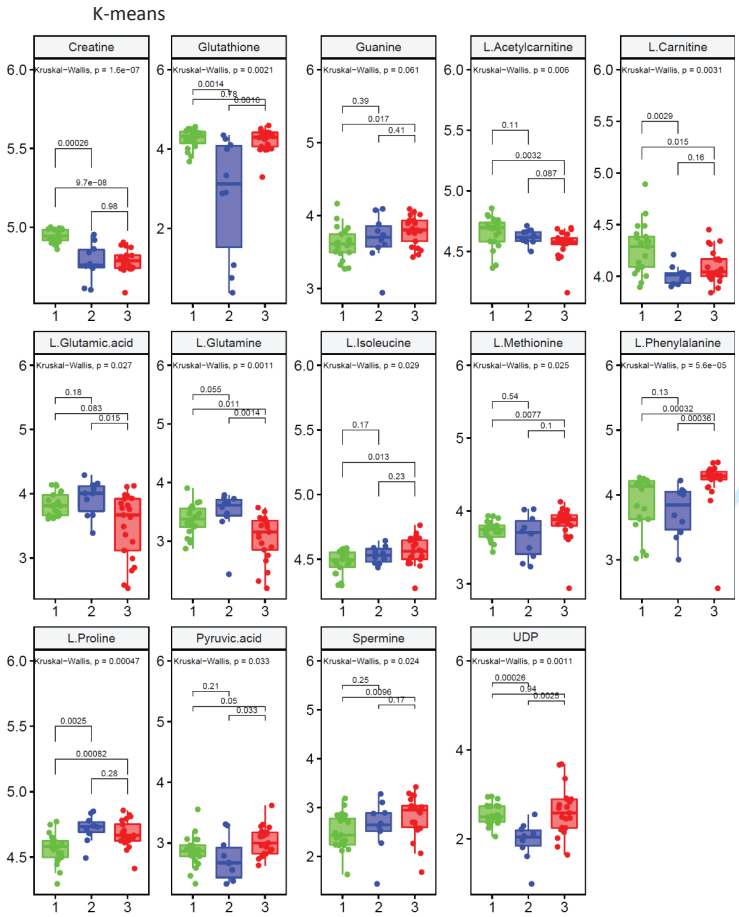
Supplementary Fig. S4: Bipartite graph of the top 50 metabolites extracted from 5 machine learning methods



Supplementary Fig. S5: Boxplot of the 14 metabolites extracted from 5 ML methods




Sparse K-means
Creatine
Kruskal–Wallis, p = 1.5e−08
Glutathione
Kruskal–Wallis, p = 0.14
Guanine
Kruskal–Wallis, p = 0.067
L.Acetylcarnitine
Kruskal–Wallis, p = 0.22
L.Carnitine
Kruskal–Wallis, p = 0.018
L.Glutamic.acid
Kruskal–Wallis, p = 0.028
L.Glutamine
Kruskal–Wallis, p = 0.002
L.Isoleucine
Kruskal–Wallis, p = 0.1
L.Methionine
Kruskal–Wallis, p = 0.093
L.Phenylalanine
Kruskal–Wallis, p = 0.052
L.Proline
Kruskal–Wallis, p = 0.0025
Pyruvic.acid
Kruskal–Wallis, p = 0.064
UDP
Kruskal–Wallis, p = 0.0061
Spectral Clustering
Creatine
Kruskal–Wallis, p = 1.6e−06
Glutathione
Kruskal–Wallis, p = 0.008
Guanine
Kruskal–Wallis, p = 0.029
L.Acetylcarnitine
Kruskal–Wallis, p = 0.003
L.Carnitine
Kruskal–Wallis, p = 3.6e−05
L.Glutamic.acid
Kruskal–Wallis, p = 0.001
L.Glutamine
Kruskal–Wallis, p = 5.6e−05
L.Isoleucine
Kruskal–Wallis, p = 0.0093
L.Methionine
Kruskal–Wallis, p = 0.0029
L.Phenylalanine
Kruskal–Wallis, p = 0.00029
L.Proline
Kruskal–Wallis, p = 6.7e−05
Pyruvic.acid
Kruskal–Wallis, p = 0.0048
Spermine
Kruskal–Wallis, p = 0.011
UDP
Kruskal–Wallis, p = 0.013


K-means
Creatine
Kruskal−Wallis, p = 1.6e−07
0.00026
9.7e−08
0.98
Glutathione
Kruskal−Wallis, p = 0.0021
0.0014
0.0016
Guanine
Kruskal−Wallis, p = 0.061
0.39
0.017
0.41
L.Acetylcarnitine
Kruskal−Wallis, p = 0.006
0.11
0.0032
0.087
L.Carnitine
Kruskal−Wallis, p = 0.0031
0.0029
0.015
0.16
L.Glutamic.acid
Kruskal−Wallis, p = 0.027
0.18
0.083
0.015
L.Glutamine
Kruskal−Wallis, p = 0.0011
0.055
0.011
0.0014
L.Isoleucine
Kruskal−Wallis, p = 0.029
0.17
0.013
0.23
L.Methionine
Kruskal−Wallis, p = 0.025
0.54
0.0077
0.1
L.Phenylalanine
Kruskal−Wallis, p = 5.6e−05
0.13
0.00032
0.00036
L.Proline
Kruskal−Wallis, p = 0.00047
0.0025
0.00082
0.28
Pyruvic.acid
Kruskal−Wallis, p = 0.033
0.21
0.05
0.033
Spermine
Kruskal−Wallis, p = 0.024
0.25
0.0096
0.17
UDP
Kruskal−Wallis, p = 0.0011
0.00026
0.84
0.0025

Supplementary Table S1: Processing time's comparison between 5 clustering methods

	PCA K-means	SIMLR	Spectral clustering	Sparse K-means	K-sparse
Times (s)	0.04	0.25	0.30	0.31	0.48

Supplementary Table S2: The 50 most effective metabolites identified by the 5 ML methods

Importance	K-sparse	SIMLR	Sparse K-means	Spectral clustering	PCA K-means
1	Thiolutin	2,2,4,4,-Tetramethyl-6-(1-oxopropyl)-1,3,5-cyclohexanetrione	Creatine	Creatine	Creatine
2	L-Glutamic acid	Histidinyl-Isoleucine	Humulinic acid A	1-Methylpyrrolo[1,2-a]pyrazine	Ascorbic acid
3	Lidocaine	Humulinic acid A	Ascorbic acid	Humulinic acid A	L-Acetylcarnitine
4	1b-Furanoeudesm-4(15)-en-1-ol acetate	5b-Cyprinol sulfate	L-Proline	Histidinyl-Isoleucine	L-Proline
5	Citramalic acid	Isosakuranin	L-Carnitine	Ascorbic acid	Glutathione
6	2-Methyl-3-ketovaleric acid	Lisuride	Methoxsalen	Phaseolic acid	Humulinic acid A
7	Betaine	Prehumulinic acid	Glutathione	1-Pyrroline	L-Carnitine
8	Hexanoylcarnitine	Phosphoric acid	L-Phenylalanine	L-Proline	Lidocaine
9	L-Proline	5-(methylthio)-2,3-Dioxopentyl phosphate	L-Isoleucine	5b-Cyprinol sulfate	L-Phenylalanine
10	Flunitrazepam	Cyclic AMP	Betaine	2,5-Dichloro-4-oxohex-2-enedioate	N-Desmethylvenlafaxine
11	Liqcoumarin	Carboxyphosphamide	L-Acetylcarnitine	2,2,4,4,-Tetramethyl-6-(1-oxopropyl)-1,3,5-cyclohexanetrione	Methoxsalen
12	Methoxsalen	L-Phenylalanine	Lidocaine	Glycerol	I(-)
13	Ethyl aconitate	Plastoquinone 3	Neurine	Propionic acid	L-Isoleucine
14	Methylmalonic acid	Gingerol	Hypoxanthine	Triethanolamine	Betaine
15	Niacinamide	(-)-Epigallocatechin	Ethyl aconitate	2',4-Dihydroxy-4',6'-dimethoxychalcone	Aminoadipic acid
16	Guanine	Hydroxyprolyl-Valine	L-Glutamic acid	Isoleucyl-Methionine	Liqcoumarin
17	Pantothenic acid	Triethanolamine	Aminoadipic acid	Alnustone	Dihydrothymine
18	L-Methionine	Phaseolic acid	Isoleucyl-Methionine	Plastoquinone 3	Hypoxanthine
19	L-Glutamine	Glycerol	Alnustone	S-Nitrosoglutathione	Glycerophosphocholine
20	Uridine 5'-monophosphate	Propionic acid	5-Aminoimidazole ribonucleotide	Prehumulinic acid	1b-Furanoeudesm-4(15)-en-1-ol acetate
21	L-Carnitine	Propenoylcarnitine	1b-Furanoeudesm-4(15)-en-1-ol acetate	2-Bromophenol	Proline betaine
22	Histamine	Isoleucyl-Methionine	Triethanolamine	5-(methylthio)-2,3-Dioxopentyl phosphate	Citramalic acid
23	Dihydrothymine	Alnustone	Citramalic acid	Propenoylcarnitine	5-Aminoimidazole ribonucleotide
24	L-Phenylalanine	2,5-Dichloro-4-oxohex-2-enedioate	Liqcoumarin	Phenmetrazine	2-Methylbutyroylcarnitine
25	Triethanolamine	2',4-Dihydroxy-4',6'-dimethoxychalcone	L-Tyrosine	(-)-Epigallocatechin	Niacinamide
26	L-Histidine	1,3,11-Tridecatriene-5,7,9-triyne	Methylmalonic acid	Lisuride	L-Glutamic acid
27	N1,N12-Diacetylspermine	N-Acetyl-L-methionine	2-Methylbutyroylcarnitine	Acrylamide	Guanine
28	Glycerophosphocholine	1-Pyrroline	Guanine	Betaine	Erythronic acid
29	3-Dehydroxycarnitine	L-Methionine	p-Cresol sulfate	3-(4-Methyl-3-pentenyl)thiophene	L-Tyrosine
30	Adenosine monophosphate	Methionine sulfoxide	L-Methionine	(alpha-D-mannosyl)7-beta-D-mannosyl-diacetylchitobiosyl-L-asparagine, isoform B (protein)	Adenosine monophosphate
31	2-Methylbutyroylcarnitine	L-Proline	I(-)	Carboxyphosphamide	Methylmalonic acid
32	Creatine	3-Methyl sulfolene	L-Glutamine	L-Isoleucine	Neurine
33	N-Acetyl-L-aspartic acid	2-Methylcitric acid	2-Methyl-3-ketovaleric acid	Methionine sulfoxide	p-Cresol sulfate
34	5-Hydroxyisourate	1-Oxo-1H-2-benzopyran-3-carboxaldehyde	N1,N12-Diacetylspermine	Dichloromethane	Creatinine
35	Hydroxyprolyl-Valine	(±)-2-Methylthiazolidine	Gravelliferone	5-Hydroxylysine	Dimethylglycine
36	Aminoadipic acid	3-Hydroxy-6,8-dimethoxy-7(11)-eremophilen-12,8-olide	(-)-Epigallocatechin	Pyrrolidine	2-Methyl-3-ketovaleric acid
37	5-Aminoimidazole ribonucleotide	1-Methylpyrrolo[1,2-a]pyrazine	Glycerophosphocholine	Gingerol	Ethyl aconitate
38	4-Chloro-1H-indole-3-acetic acid	1-(2,4,6-Trimethoxyphenyl)-1,3-butanedione	Adenosine monophosphate	2,3-diketogulonate	N1,N12-Diacetylspermine
39	Creatinine	5-(4-Acetoxy-3-oxo-1-butynyl)-2,2'-bithiophene	Uridine 5'-monophosphate	2,2-dichloro-1,1-ethanediol	Thiolutin
40	PC-M6	Ac-Ser-Asp-Lys-Pro-OH	5-Hydroxyisourate	L-Carnitine	Uridine 5'-monophosphate
41	Trimethylamine N-oxide	Alanyl-Isoleucine	2,5-Dichloro-4-oxohex-2-enedioate	Erythro-5-hydroxy-L-lysinium(1+)	L-Methionine
42	3-Mercaptohexyl hexanoate	5-Hydroxyindoleacetic acid	Ergothioneine	Neurine	Gravelliferone
43	Prolylhydroxyproline	Benzothiazole	Hexanoylcarnitine	Isosakuranin	Pipecolic acid
44	p-Cresol sulfate	Glutathione	Histidinyl-Isoleucine	Aspartyl-L-proline	Methyl (9Z)-10'-oxo-6,10'-diapo-6-carotenoate
45	L-Acetylcarnitine	Phenmetrazine	Valganciclovir	Valerenic acid	4-Chloro-1H-indole-3-acetic acid
46	Hypoxanthine	Glycerol tripropanoate	Phosphoric acid	5-Hydroxyindoleacetic acid	Isoleucyl-Methionine
47	Garcinia acid	Creatine	Hydroxyprolyl-Valine	Beta-Alanine	Alnustone
48	Erythronic acid	2-Bromophenol	3-(4-Methyl-3-pentenyl)thiophene	Glutathione	3-Dehydroxycarnitine
49	Guanidoacetic acid	3-(4-Methyl-3-pentenyl)thiophene	Erinapyrone C	Piperidine	Garcinia acid
50	Proline betaine	(alpha-D-mannosyl)7-beta-D-mannosyl-diacetylchitobiosyl-L-asparagine, isoform B (protein)	N-Methylethanolaminium phosphate	2,5-Furandicarboxylic acid	Valganciclovir

Supplementary Table S3: List of all pathways identified 5 methods

Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	Raw p	-*log*(p)	Impact
K-Sparse method							
C1 vs C2		Porphyrin and chlorophyll metabolism	104	2	0,0101	4,5996	0,0000
	L-Histidine; L-Phenylalanine; L-Glutamine; L-Methionine; L-Isoleucine; L-Threonine; L-Tyrosine; L-Proline; L-Glutamic acid; Phosphoserine;	Aminoacyl-tRNA biosynthesis	75	10	0,0453	3,0950	0,0563
C1 vs C3	UDP- glucose	Starch and sucrose metabolism	50	1	0,0107	4,5388	0,1390
	UDP- glucose	Amino sugar and nucleotide sugar metabolism	88	1	0,0107	4,5388	0,0928
	UDP- glucose	Galactose metabolism	41	1	0,0107	4,5388	0,0009
	UDP-glucose; Glyceric acid	Glycerolipid metabolism	32	2	0,0153	4,1831	0,0206
	Isoleucine; Methylmalonic acid	Valine, leucine and isoleucine degradation	40	2	0,0264	3,6350	0,0000
C2 vs C3	Formiminoglutamic acid; L-Glutamic acid; L-Histidine; Histamine; Ergothioneine;	Histidine metabolism	44	5	0,0299	3,5114	0,1977
	Triethanolamine; Glycerylphosphorylethanolamine; Glycerophosphocholine;	Glycerophospholipid metabolism	39	3	0,0366	3,3065	0,0263
	Pipecolic acid; Aminoadipic acid;	Lysine degradation	47	2	0,0490	3,0160	0,0161
SIMLR method							
Clusters Comparaison	**Interaction metabolite**	**Pathway Name**	**Total Cmpd**	**Match Status**	**Raw p**	**-*log*(p)**	**Impact**
C1 VS C2	Glutathione; Oxidized glutathione; Glycine; L-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine; Cadaverine; Aminopropylcadaverine; Ascorbic acid;	Glutathione metabolism	38	12	0	12,826	0,3628
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0	12,469	0,1383
	L-Tryptophan; N-Acetylserotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; L-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine;	Tryptophan metabolism	79	8	0,0001	9,1233	0,2741
	5'-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0008	7,1674	0,2509
	L-Glutamine; Phosphoribosylformylglycineamidine; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; 5-Hydroxyisourate; Guanosine; Adenosine diphosphate ribose; 5-Aminoimidazole ribonucleotide; Glyoxylic acid; Glycine; Adenosine 3',5'-diphosphate;	Purine metabolism	92	17	0,0011	6,8091	0,2048
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0027	5,9281	0,268
	L-Glutamine; Ornithine; Citrulline; L-Arginine; L-Glutamic acid; N-Acetylornithine; L-Proline; Hydroxyproline; Guanidoacetic acid; Creatine; 4-Guanidinobutanoic acid; N2-Succinyl-L-ornithine; Putrescine; Spermidine; N-Acetylputrescine; Pyruvic acid; Glyoxylic acid; Spermine;	Arginine and proline metabolism	77	19	0,0053	5,238	0,6514
	Arginine; Ornithine;	D-Arginine and D-ornithine metabolism	8	2	0,0054	5,2304	0
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,0075	4,8991	0,176
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0076	4,8821	0,0394
	Glycine;5b-Cyprinol sulfate;	Primary bile acid biosynthesis	47	2	0,0123	4,4004	0,0082
	2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0154	4,1754	0,0324

		Glycerylphosphorylethanolamine; Glycerophosphocholine;	Ether lipid metabolism	23	2	0,0162	4,1223	0
		Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; Pyruvic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,018	4,0154	0,46986
		Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; N-Acetyl-D-Glucosamine 6-Phosphate; Uridine diphosphate-N-acetylglucosamine ; Cytidine monophosphate N-acetylneuraminic acid; D-Glucose; D-Xylose	Amino sugar and nucleotide sugar metabolism	88	7	0,0187	3,9783	0,1417
		Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0412	3,1903	0,3705
		Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0412	3,1898	0,0773
	C1 VS C3	Isoleucine; Methylmalonic acid	Valine, leucine and isoleucine degradation	40	2	0,0067	4,9992	0
		Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0139	4,2752	0,3705
		L-Phenylalanine; L-Tyrosine; L-Tryptophan; L-Glutamic acid; L-Glutamine; L-Homocysteine; L-Histidine; Glycine; Adenosine monophosphate;	Nitrogen metabolism	39	9	0,0152	4,1851	0
		Phenylpyruvic acid; L-Phenylalanine; L-Tyrosine; 3-Dehydroquinate; L-Tryptophan;	Phenylalanine, tyrosine and tryptophan biosynthesis	27	5	0,0189	3,9687	0,099
		L-Histidine; L-Phenylalanine; L-Arginine; L-Glutamine; Glycine; L-Methionine; L-Lysine; L-Isoleucine; L-Threonine; L-Tryptophan; L-Tyrosine; L-Proline; L-Glutamic acid; Phosphoserine;	Aminoacyl-tRNA biosynthesis	75	14	0,0245	3,7085	0,1127
		L-Glutamine; Phosphoribosylformylglycineamidine; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; Guanosine; Adenosine diphosphate ribose; 5-Aminoimidazole ribonucleotide; Glyoxylic acid; Glycine; Adenosine 3',5'-diphosphate;	Purine metabolism	92	17	0,0328	3,4159	0,2048
		Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0434	3,1375	0,0773
		Glycerol 3-phosphate; Triethanolamine; Citicoline; O-Phosphoethanolamine; Glycerylphosphorylethanolamine; CDP-Ethanolamine; Glycerophosphocholine;	Glycerophospholipid metabolism	39	7	0,0491	3,0137	0,2464
	C2 VS C3	L-Tryptophan; N-Acetylserotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; L-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine;	Tryptophan metabolism	79	8	0	16,409	0,2741
		Glutathione; Oxidized glutathione; Glycine; L-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine;	Glutathione metabolism	38	12	0	16,133	0,3628
		Cadaverine; Aminopropylcadaverine; Ascorbic acid;						
		Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid	Ascorbate and aldarate metabolism	45	5	0	13,096	0,1383
		5'-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0001	9,8548	0,2509
		L-Phenylalanine; L-Tyrosine; L-Tryptophan; L-Glutamic acid; L-Glutamine; L-Homocysteine; L-Histidine; Glycine; Adenosine monophosphate;	Nitrogen metabolism	39	9	0,0001	9,1978	0

	Phenylpyruvic acid; L-Phenylalanine; L-Tyrosine; 3-Dehydroquinate; L-Tryptophan;	Phenylalanine, tyrosine and tryptophan biosynthesis	27	5	0,0001	8,9814	0,099
	L-Histidine; L-Phenylalanine; L-Arginine; L-Glutamine; Glycine; L-Methionine; L-Lysine; L-Isoleucine; L-Threonine; L-Tryptophan; L-Tyrosine; L-Proline; L-Glutamic acid; Phosphoserine;	Aminoacyl-tRNA biosynthesis	75	14	0,0002	8,758	0,1127
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0004	7,7271	0,268
	Glycine;Cyprinol sulfate;	Primary bile acid biosynthesis	47	2	0,0006	7,4259	0,0082
	L-Glutamine; Phosphoribosylformylglycineamidine; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; 5-Hydroxyisourate; Guanosine; Adenosine diphosphate ribose; 5-Aminoimidazole ribonucleotide; Glyoxylic acid; Glycine; Adenosine 3',5'-diphosphate;	Purine metabolism	92	17	0,0007	7,306	0,2048
	Malonic acid; Beta-Alanine; Spermine; Spermidine; Dihydrouracil; Pantothenic acid; Uracil; L-Histidine	beta-Alanine metabolism	28	8	0,0012	6,7568	0,3577
	Uridine 5'-monophosphate; L-Glutamine; Dihydrouracil; Cytidine monophosphate; Cytidine; Cytosine; Uracil; Dihydrothymine; Uridine diphosphate glucose; Malonic acid; Ureidosuccinic acid; Beta-Alanine; Methylmalonic acid;	Pyrimidine metabolism	60	13	0,0014	6,5817	0,2756
	Pantothenic acid; Dihydrouracil; Beta-Alanine; Pyruvic acid; Adenosine 3',5'-diphosphate; Uracil;	Pantothenate and CoA biosynthesis	27	6	0,0023	6,0879	0,2736
	L-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; L-Tyrosine;	Phenylalanine metabolism	45	6	0,0072	4,9364	0,2468
	L-Glutamic acid; L-Glutamine; Oxoglutaric acid	D-Glutamine and D-glutamate metabolism	11	3	0,0124	4,39	0,139
	Isoleucine;Methylmalonic acid	Valine, leucine and isoleucine degradation	40	2	0,015	4,1984	0
	L-Glutamine; Ornithine; Citrulline; L-Arginine; L-Glutamic acid; N-Acetylornithine; L-Proline; Hydroxyproline; Guanidoacetic acid; Creatine; Creatinine; 4-Guanidinobutanoic acid; N2-Succinyl-L-ornithine; Putrescine; Spermidine; N-Acetylputrescine; Pyruvic acid; Glyoxylic acid; Spermine;	Arginine and proline metabolism	77	19	0,0169	4,082	0,6514
	2-Hydroxyethanesulfonate ; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0215	3,8411	0,0324
	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,0221	3,8108	0,4122
	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0267	3,6235	0,0773
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid	Citrate cycle (TCA cycle)	20	3	0,0302	3,5015	0,176
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,0372	3,2914	0,4699
	Uridine diphosphate glucose; Glycerol 3-phosphate; Glycerol; Glyceric acid; Galactosylglycerol;	Glycerolipid metabolism	32	5	0,0427	3,1546	0,2162
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0427	3,1536	0,0394

Sparse Kmeans method

Clusters Comparaison	Interaction metabolite		Total Cmpd	Match Status	Raw p	*-log*(p)	Impact
C1 VS C2 C1 VS C3	L-Methionine;Glutathione	Cysteine and methionine metabolism	56	2	0.007	4.9	0.0454

C1 VS C3	L-Methionine;Glutathione;	Cysteine and methionine metabolism	56	2	0.0020	6.2	0.00454

Spectral clustering method

Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	Raw p	*-log*(p)	Impact
C1 VS C2	Glycine; L-Glutamic acid; L-Threonine;	Porphyrin and chlorophyll metabolism	104	3	0,0240	3,7290	0,0000
	Uridine diphosphate glucose; Glycerol 3-phosphate; Glycerol; Glyceric acid; Galactosylglycerol;	Glycerolipid metabolism	32	5	0,0334	3,4003	0,2162
	Galactitol; Galactosylglycerol; Uridine diphosphate glucose; D-Galactonate; D-Glucose; Glycerol;	Galactose metabolism	41	6	0,0363	3,3154	0,1539
C1 VS C3	Iminoaspartic acid; Quinolinic acid; Niacinamide; Pyruvic acid; Propionic acid;	Nicotinate and nicotinamide metabolism	44	5	0,0024	6,0206	0,0712
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,0040	5,5100	0,4699
	5'-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0098	4,6232	0,2509
	Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0101	4,5961	0,3705
	xoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,0171	4,0710	0,1760
	Pyruvic acid; L-Threonine; L-Isoleucine;	Valine, leucine and isoleucine biosynthesis	27	3	0,0178	4,0277	0,0350
	Pyruvic acid;	Terpenoid backbone biosynthesis	33	1	0,0207	3,8797	0,0000
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0210	3,8609	0,0394
	D-Glucose; Glyceric acid; Pyruvic acid;	Pentose phosphate pathway	32	3	0,0232	3,7622	0,0218
	Pyruvic acid; L-Lactic acid; D-Glucose;	Glycolysis or Gluconeogenesis	31	3	0,0249	3,6928	0,0953
	Pyruvic acid; L-Lactic acid;	Pyruvate metabolism	32	2	0,0274	3,5955	0,3201
	L-Glutamic acid; Pyruvic acid; Butyric acid; Oxoglutaric acid;	Butanoate metabolism	40	4	0,0283	3,5644	0,0852
	2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0287	3,5525	0,0324
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0303	3,4966	0,2680
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0,0330	3,4104	0,1383
	Epinephrine; Dopamine; L-Tyrosine; Homovanillic acid; Pyruvic acid;	Tyrosine metabolism	76	5	0,0385	3,2580	0,1750
	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,0390	3,2431	0,4122
	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0447	3,1074	0,0773
C2 VS C3	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,0209	3,8659	0,4122
	L-Glutamic acid; L-Glutamine; Oxoglutaric acid	D-Glutamine and D-glutamate metabolism	11	3	0,0275	3,5922	0,1390

PCA-Kmeans method

Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	Raw p	*-log*(p)	Impact
C1 vs C2	Iminoaspartic acid; Quinolinic acid; Niacinamide; Pyruvic acid; Propionic acid;	Nicotinate and nicotinamide metabolism	44	5	0,020	3,9278	0,0712

		Galactitol; Galactosylglycerol; Uridine diphosphate glucose; D-Galactonate; D-Glucose; Glycerol;	Galactose metabolism	41	6	0,025	3,6893	0,1539
		Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0,046	3,0797	0,1383
		Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,048	3,0407	0,3705
		L-Glutamic acid; L-Glutamine; Oxoglutaric acid	D-Glutamine and D-glutamate metabolism	11	3	0,049	3,0236	0,1390
	C1 vs C3	Iminoaspartic acid; Quinolinic acid; Niacinamide; Pyruvic acid; Propionic acid;	Nicotinate and nicotinamide metabolism	44	5	0,003	5,9412	0,0712
		Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,011	4,4865	0,1760
		Epinephrine; Dopamine; L-Tyrosine; Homovanillic acid; Pyruvic acid;	Tyrosine metabolism	76	5	0,024	3,7311	0,1750
		Pyruvic acid;	Terpenoid backbone biosynthesis	33	1	0,031	3,4834	0,0000
		Pyruvic acid;L-Lactic acid;	Pyruvate metabolism	32	2	0,043	3,1507	0,3201
		D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate ;Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,044	3,1214	0,0394
		Pyruvic acid; L-Threonine; L-Isoleucine;	Valine, leucine and isoleucine biosynthesis	27	3	0,045	3,1107	0,0350
		Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0,045	3,0926	0,1383
		L-Glutamic acid; Pyruvic acid; Butyric acid; Oxoglutaric acid;	Butanoate metabolism	40	4	0,046	3,0843	0,0852
		D-Glucose; Glyceric acid; Pyruvic acid;	Pentose phosphate pathway	32	3	0,046	3,0769	0,0218
		N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,048	3,0446	0,4122
	C2 vs C3	L-Glutamic acid; L-Glutamine; Oxoglutaric acid	D-Glutamine and D-glutamate metabolism	11	3	0,012	4,4588	0,1390
		N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,046	3,0796	0,4122

II.2.3. Discussion

Cette étude s'inscrivait dans le cadre du projet EMMEEA, développé par le Centre Antoine Lacassagne et le laboratoire TIRO de la faculté de médecine de Nice.

Les résultats que nous avons mis en évidence sont doublement intéressants d'un point de vue clinique. D'une part, l'analyse des métabolites a permis d'identifier trois groupes de patientes qui se sont révélés correspondre à des profils de pronostic bon, intermédiaire ou mauvais définis selon les facteurs cliniques et biologiques reconnus. Cela confirme le potentiel discriminatif de la métabolomique et laisse entrevoir des applications concrètes dans le diagnostic des patients. D'autre part, les analyses ont retrouvé une hétérogénéité au sein des tumeurs triple-négatives. En effet, la majorité de ces tumeurs ont été classées dans le groupe pronostic mauvais mais il ressort qu'environ un tiers de ces tumeurs ont été classées dans le groupe bon pronostic. Le groupe des tumeurs triple-négatives est connu pour contenir des sous-types hétérogènes au niveau génétique et moléculaire avec des répercussions sur le pronostic [138, 139]. La métabolomique pourrait donc être un outil déterminant pour aider à affiner la classification des cancers du sein triple-négatifs au-delà des autres "omics" utilisés actuellement. Ce résultat fait écho à l'utilisation de ces méthodes dans l'établissement de la classification moléculaire initiale du cancer du sein (luminal A et B; Her2; Normal-like et Basale-like) [43-45]. La métabolomique possède des avantages comparé à la transcriptomique et à la protéomique Tout d'abord, la métabolomique est le dernier maillon au niveau de la cascade des omiques. Elle est le reflet le plus proche de la fonctionnalité et du phénotype d'une cellule [140]. En comparaison à la génomique, la métabolomique a la particularité d'intégrer l'impact de l'environnement tumoral devenu une nouvelle cible thérapeutique [141, 142]. Enfin, il faut considérer que le coût et la durée des analyses en métabolomique sont peu élevés tout en restant non invasive [143], rendant ces approches compatibles à une routine clinique contrairement à la transcriptomique. L'inconvénient de la métabolomique réside en deux points : 1- la non publication des données conduisant par conséquent à une certaine réserve quant à la fiabilité des résultats ; 2- un besoin d'une standardisation des pratiques entre les plateformes afin que cette approche puisse être utilisée en routine clinique. La validation interne de cette classification a été effectuée à l'aide de l'indice de silhouette. En raison du caractère inconnu des clusters, aucune validation externe n'a pu être réalisée. Néanmoins une validation « clinique » a été réalisée en comparant nos résultats avec les caractéristiques cliniques et biologiques des patientes incluses dans cette étude [144]. On notera que les méthodes utilisées dans ce travail sont sensibles aux outliers[1], qu'elles ne tolèrent pas les données manquantes et qu'elles attendent en entrée uniquement des données quantitatives. Bien que les cinq méthodes testées aient montré des résultats globalement similaires, nous avons pu constater que certaines méthodes sont plus performantes que d'autres. Avec un indice de silhouette de 0,91 et de 0,85 respectivement les

[1] Outliers ou données aberrantes sont des données qui sont « distantes » des autres données. Elles contrastent grandement avec les valeurs « normalement » mesurées.

méthodes K-sparse et SIMLR sont les méthodes les plus performantes en termes de clustering. Ces deux méthodes apportent chacune des informations différentes mais complémentaires.

Il est important également de souligner l'efficacité et la sensibilité de ces méthodes qui ont permis de mettre en évidence trois groupes de patientes avec des profils de pronostic différents à partir d'une cohorte de seulement 52 patientes et 449 métabolites. A notre connaissance, cette étude est la première à comparer différentes méthodes de ML non supervisées afin d'identifier des signatures pronostiques basées sur la métabolomique dans le cancer du sein. La métabolomique offre un accès à un grand nombre de données intégrant l'influence de l'ensemble des facteurs internes et externes modifiant les processus cellulaires. La compréhension des phénomènes complexes influençant ces données nécessite des transformations de données et des approches statistiques innovantes. La métabolomique semble être donc un outil pertinent et prometteur dans la classification des cancers du sein. Cependant, la pertinence clinique de cette signature nécessite d'être validée sur des cohortes indépendantes comprenant un nombre important de patients. Dans la continuité de ces travaux, nous allons très prochainement comparer les résultats obtenus entre ces cinq méthodes chez 39 patientes atteintes d'un cancer d'un cancer du sein traité par chimiothérapie néoadjuvante.

Partie III : Prolongements et perspectives

III. Partie III : Prolongements et perspectives

III.1. Optimiser le développement des médicaments en oncologie par simulation d'essais cliniques : pourquoi et comment ?

III.1.1. Contexte

A côté des apports potentiels indéniables offerts par l'IA au niveau de la biologie du cancer, il y a un domaine tout autant prometteur qui est celui des essais thérapeutiques simulés (ETS). Depuis de nombreuses années, les acteurs de la recherche clinique s'efforcent d'optimiser les essais thérapeutiques en développant diverses techniques méthodologiques et statistiques afin de prédire les résultats de ces essais. Dans ce cadre, des méthodologies alternatives telles que l'IA et les ETS peuvent être des options intéressantes. Pour en démontrer l'intérêt, nous avons réalisé une revue systématique de la littérature afin de mieux cerner l'importance de ces essais *in silico* dans un contexte de développement d'un médicament. Ce travail s'est basé sur la lecture critique de 139 articles dont seulement 10 étaient réellement des ETS. Une réactualisation de cette recherche en 2019 a pu mettre en évidence que seulement 2 nouveau ETS « vrais » ont été publié depuis 2017. Nos résultats ont montré que la réalisation d'un ETS dans lequel serait inclus des patients entièrement simulés relève plus du concept que de la réalité. A la date de publication de cet article dans *Briefing in Bioinformatics* [85], le concept et les méthodes d'ETS restaient principalement du domaine de la pharmacologie (pharmacocinétique et pharmacodynamie). Nos résultats ont mis en évidence que l'apport des méthodes d'IA appliquées à des données de génomiques pourrait permettre d'améliorer nos connaissances dans la compréhension des mécanismes des cancers et être complémentaires aux ETS. En effet, ces méthodes détectent des groupes de gènes avec une expression similaire et établissent le profil génétique de la tumeur permettant ainsi d'identifier les profils de répondeurs et / ou de patients tolérants aux médicaments et ainsi fournir des informations pour l'élaboration des modèles mathématiques. Quoi qu'il en soit, ces essais *in silico* devront toujours être validés dans le cadre d'un essai thérapeutique prospectif. A l'avenir, ils devront faire partie intégrante des futurs programmes de développement des médicaments afin de fournir un soutien quantitatif lors de la prise de décision. Cette médecine *in silico* a ouvert la voie au développement de la médecine 4P : prédictive, préventive, personnalisée et participative.

III.1.2. Publication: Optimizing drug development in oncology by clinical trial simulation: Why and how?

Briefings in Bioinformatics, 2017, 1–15

doi: 10.1093/bib/bbx055
Paper

Optimizing drug development in oncology by clinical trial simulation: Why and how?

Jocelyn Gal, Gérard Milano, Jean-Marc Ferrero, Esma Saâda-Bouzid, Julien Viotti, Sylvie Chabaud, Paul Gougis, Christophe Le Tourneau, Renaud Schiappa, Agnes Paquet and Emmanuel Chamorey

Corresponding author: Jocelyn Gal, Epidemiology and Biostatistics Unit, Centre Antoine Lacassagne, 33 avenue de Valombrose, 06189 Nice, France. Tel.: +33-4-9203-1031; E-mail: jocelyn.gal@nice.unicancer.fr

Abstract

In therapeutic research, the safety and efficacy of pharmaceutical products are necessarily tested on humans via clinical trials after an extensive and expensive preclinical development period. Methodologies such as computer modeling and clinical trial simulation (CTS) might represent a valuable option to reduce animal and human assays. The relevance of these methods is well recognized in pharmacokinetics and pharmacodynamics from the preclinical phase to postmarketing. However, they are barely used and are poorly regarded for drug approval, despite Food and Drug Administration and European Medicines Agency recommendations. The generalization of CTS could be greatly facilitated by the availability of software for modeling biological systems, by clinical trial studies and hospital databases. Data sharing and data merging raise legal, policy and technical issues that will need to be addressed. Development of future molecules will have to use CTS for faster development and thus enable better patient management. Drug activity modeling coupled with disease modeling, optimal use of medical data and increased computing speed should allow this leap forward. The realization of CTS requires not only bioinformatics tools to allow interconnection and global integration of all clinical data but also a universal legal framework to protect the privacy of every patient. While recognizing that CTS can never replace 'real-life' trials, they should

Jocelyn Gal is a biostatistician in the Epidemiology and Biostatistics Unit at the Antoine Lacassagne Center, Nice, France. He is currently pursuing his PhD in the Laboratory of Pharmacological Targeting Unit in Oncology, Faculty of Sciences, University of Nice. His research interests are bioinformatics, methodology in clinical research, systems biology, data mining and machine learning.
Gerard Milano is the head of the Pharmacological Targeting Unit in Oncology at the Antoine Lacassagne Center in Nice, France. He is working on the personalized approach to medicine designed to manage cancer treatment according to specific individual characteristics of the patient's tumor.
Jean-Marc Ferrero is Professor in Medical Oncology in charge of the Clinical Research Unit at the Antoine Lacassagne Centre, Nice, France. His research interests are precision medicine in breast cancer and clinical research.
Esma Saâda-Bouzid, oncologist, received her PhD degree in Molecular and Cellular Interactions in 2015 from the University of Nice, France. Her research interests are molecular biology and clinical research in oncology.
Julien Viotti received his PhD degree in Molecular and Cellular Interactions in 2014 from the University of Nice, France. His research interests are molecular biology and statistical analysis.
Sylvie Chabaud is a biostatistician in charge of the Biostatistics and Therapy Evaluation Unit at the Centre Léon Bérard, Lyons, France. Her research interests are simulation and modeling, experimental design, statistical analysis and therapeutic evaluation. She is a member of the Child Rare European Simulation (CRESim) Project Group.
Paul Gougis is resident physician at the Pitié-Salpêtrière Hospital in Paris, France. He is a specialist in Clinical Pharmacology and Therapeutic Assessment.
Christophe Le Tourneau is Professor in Medical Oncology at the Institute Curie of Paris, France. His research interests are early trials and precision medicine.
Renaud Schiappa is Data Manager in the Epidemiology and Biostatistics Unit at the Antoine Lacassagne Centre in Nice, France. His research interests are bioinformatics, text mining and clinical research.
Agnes Paquet is a biostatistician on the Functional Genomics Platform at the Molecular and Cellular Pharmacology Institute of Sophia Antipolis, Valbonne, France. She works on the analysis of gene expression of individual cells to find and/or characterize the various subpopulations of cells.
Emmanuel Chamorey is a pharmacist and head of the Epidemiology and Biostatistics Unit at the Antoine Lacassagne Center in Nice, France. His research interests are oncopharmacology, methodology in clinical research and statistical analysis.
Submitted: 7 February 2017; **Received (in revised form):** 19 April 2017

be implemented in future drug development schemes to provide quantitative support for decision-making. This *in silico* medicine opens the way to the P4 medicine: predictive, preventive, personalized and participatory.

Key words: clinical trial simulation; drug development; machine learning; big data; oncology

Introduction

A leading objective of therapeutic research is to obtain, as early as possible, the US Food and Drug Administration (FDA) approval of new drugs for the benefit of patients as well as to address public health issues [1]. The average time lag between discovery of a molecule and granting of marketing authorization for the drug is about 10 years [2, 3]. In oncology, as in most other therapeutic domains, the development of a new molecular entity (NME) follows a mandatory, standard, two-phase procedure. First, a preclinical phase assesses the toxicity and efficacy of the drug by *in vitro* and *in vivo* tests, and then, a second clinical phase is performed comprising four phases of development. Phase I evaluates the drug's toxicity, Phase II assesses drug efficacy, Phase III compares the efficacy of the NME with the standard treatment and Phase IV assesses the safety profile and efficacy in real life (postmarketing authorization). This development process faces three types of constraints. The first is ethical because of the potential toxicity and possible ineffectiveness of the tested drug, especially during the dose escalation in Phase I trials [4, 5]. The second is organizational with the increasing complexity of clinical trials (design, personnel, clinical operations, stakeholders, payers, geography and regulatory agencies) [6, 7]. The third is economic, as therapeutic development is increasingly costly (spending increased from $800 million to $2.6 billion between 2003 and 2014 [6]) and returns on investment do not always meet expectations, with the exception of cancer immunology. The approval success rate for anticancer drugs following Phase I was estimated at 7% and at about 50% after Phase III [8, 9]. In the United States, of the 71 drugs approved by the FDA between 2002 and 2014, the overall survival (OS) improvement, all cancers combined, was 2.1 months and 2.5 months for progression-free survival (PFS) [10]. Concerning solid tumors, a review of the European Medicines Agency (EMA) shows an improvement of 1.5 months [11].

The challenge for the pharmaceutical industry is to succeed in reconciling all these constraints [12, 13]. For many years, clinical researchers have been striving to optimize clinical trials by developing various methodological and statistical techniques designed to predict the results of these trials. Alternative methodologies such as computer modeling and clinical trial simulation (CTS) can be valuable options. For instance, the Radiotherapy Department at University College London Hospitals NHS Foundation Trust has recently partnered with Google DeepMind [14], with the aim of applying machine learning (ML) to radiotherapy planning for head and neck cancer. The artificial intelligence computer system known as Watson designed by IBM is being designed to help medical decisions [15, 16]. TensorFlow, the open-source software library for machine intelligence developed by Google can be used for computational biology [17, 18]. In 2009, the European Commission authorized France to award aid amounting to €46.3 million for the BioIntelligence program developed by the Dassault Systems company [19].

In December 2016, the US Congress voted the 21st Century Cures Act [20, 21], which facilitates FDA approval on NME or new indications for existing drugs to maximize the return on investment by reducing the time spent and the number of patients enrolled in clinical trials. In this case, CTS could be used to obtain FDA approval. We propose to discuss the principle behind, and the interest of, CTS approaches in drug development in oncology. We will show that CTS can be used in oncology to optimize dose selection and to characterize drug effect on tumor growth, OS and safety or to optimize designs of clinical trials because even though cancers affect an increasing number of patients, clinical trial eligibility criteria are increasingly targeting specific patients (e.g. selection based on a specific histology or really rare genetic mutations). Throughout this article, we will use the term CTS as a generic expression encompassing the terms modeling and simulation and *in silico* (key points).

Search strategy and article extraction

We performed a review of the published literature using one electronic literature database (Medline®) applying both validated terminology [Medical Subject Headings/MeSH (MeSH)] and free text words. The search included the terms: 'clinical trial simulation' and oncology. Only studies fully published in the English language with authors listed were included. No restriction was placed on publication dates. Data were extracted by two independent reviewers (J. G., E. C.). First, we screened titles alone for eligibility, and then, in a second step, the abstracts [22]. Only full-text articles were included in the final list. Disagreements between the two reviewers were resolved by a second joint, consensual review of these publications. Additional publications were retrieved by hand-search reviewing references in the included publications (Figure 1). Finally, 10 CTS projects in oncology were identified.

A brief history of CTS

During the 1970–80s, Peck, Sheiner and colleagues [23–27] were the first to show interest in CTS by proposing advanced pharmacokinetic/pharmacodynamic (PK/PD) analyses and nonlinear mixed effect modeling to improve patient care with drugs that were already commercialized. During the 1990s, several scientists, mainly European and American, proposed to expand its use to drug development during the 1990s and published several papers [28–34]. In 1997, an important study by Sheiner [35] proposed the 'learn and confirm' concept within the drug development process to explain in simple terms how modeling and simulations needed to be done throughout the drug development process. This eventually led to the development of the first FDA guidance on population pharmacokinetics published in 1999 [36]. This same year, Holford [37] published a guideline on the Center for Drug Development Science Web site entitled 'Simulation in Drug Development: Good Practices' to help develop and build a model linked to simulation of clinical trials. A year later, three others papers were published by Holford *et al.* [38], Bonate [39] and Gieschke and Steimer [40]. The first was a guideline article, while the second and the third discussed the potential of CTS in drug development. In 2003, Kimko and Dufull published a book entitled 'Simulation for Designing

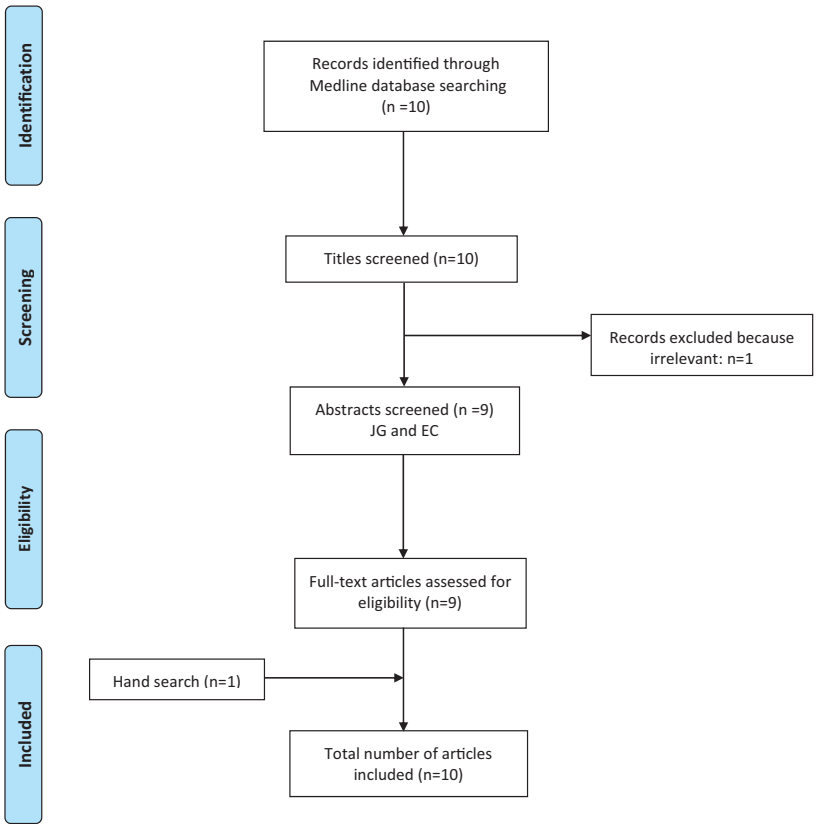


Figure 1. Flow diagram illustrating flow of studies identified from the search strategy.

Clinical Trials' [41]. More recently, in 2010, Holford *et al.* [42] published 'Clinical Trial Simulation: A Review' reviewing the current role of CTS in drug development and clinical use and providing an update on developments in the use and methodology of CTS since their own review in 2000 [38]. Parallel to this pharmacological approach, some authors investigated the role of computers in clinical research [43–50].

Could CTS improve clinical research in oncology?

Currently, the number of NME placed on the market is too small, compared with the 'me-too drugs' [10] (drugs whose therapeutic uses are similar in efficacy but with side-effect profiles not completely superposable or improving quality of life). CTS could lower development costs of 'me-too drugs'. Pharmaceutical laboratories could then invest more in research and development of innovative molecules. CTS can be applied in both exploratory and confirmatory drug development by using specific cases in which modern trial designs and a statistical approach have been successful (Figure 2).

During exploratory studies

The pharmacological aspects of cancer treatment have changed profoundly over the past 10 years, and the advent of targeted therapies has made it even more difficult to adapt doses [51]. Application of PK/PD modeling at each separate stage of drug development in oncology is summarized in Table 1. Inter-patient variability [52], pharmacogenetic implications [53] and major pharmacokinetic and pharmacodynamic [54–56] changes must be taken into account even more thoroughly than in the past, as many targeted treatments are effective only if they are administered on an ongoing basis. The development of biomarkers and related diagnostic tests entails the implementation of a clearly defined and validated method of analysis and the demonstration, via well-conducted studies, of the clinical validity of the biomarker. Hence, joint evaluation of targeted therapies and their biomarkers requires us to rethink the design and analysis of randomized clinical trials to obtain marketing authorization [53, 57–61]. Two distinct types of personalized medicine trials were identified [59]: stratified clinical trials that can be stratified according to either molecular alterations [61, 62] or tumor types (V-BASKET trial NCT01524978) [63], and algorithm-testing trials that evaluate a treatment algorithm instead of drugs' efficacy. Algorithm-testing trials include non-randomized trials (WINTER trial NCT01856296) that usually use patients as their own control to assess efficacy, and randomized trials [60] that address various questions. Furthermore, Gosho *et al.* [64] described seven different study designs using biomarkers.

During this stage, CTS could be used as early as possible [65–71] to ensure dose safety [72–76] and to allow a drug's dose–response [77–80] characteristics to be better understood at every stage of its development. CTS, for example, can provide model-based tumor growth inhibition metrics as biomarkers to capture treatment effect and predict for OS [81–83], or describe imprecision in the estimation of time to progression [84] or include a model of growing metastatic tumors [85]. They can be used to investigate the value of individual dosage adjustment to control patient exposure or to adapt target concentrations individually by taking into account the variability of the concentration–effect

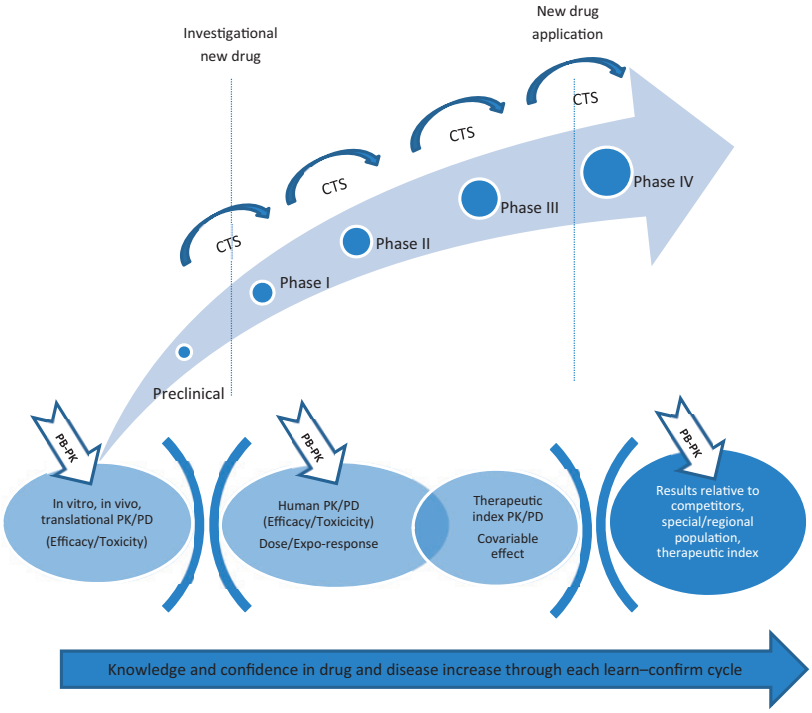


Figure 2. CTS during drug development.

Table 1. Benefits of PK/PD model during different phases of drug development in oncology

Objectives	Benefits of PK/PD modeling
Preclinical	
• Assessment of antitumor activity • Determination of effective concentration	• Selecting the optimal compound if several compounds • Predicting clinical potency estimates • Providing the guidance for the dose range to be tested in early clinical trials • Assessing the margin of safety on the basis of target efficacy concentrations • Predicting bioavailability • Assessing the potential for drug–drug interactions
Phase I	
• Determine most optimal dose and schedule • Assess toxicity profile • Study PK and relationships with toxicity dose	• Definition of a useful target for Phase I study • Selection of safe starting dose • Selection of treatment schedules
Phase IIa/IIb	
• Establish activity in a fairly large patients • Study PK/PD	• Develop a drug–disease model to understand the evolution of disease progression and dose–response to interventions • Test different study designs and assist in selection of the optimal design for the given conditions • Assess the impact of covariates, using a population PK/PD model • Assess the efficacy/toxicity profile, relative to comparators
Phase III	
• Confirmation of activity in randomized trials and comparison with standard treatment	• Assess the impact of applicable covariates (patient subpopulations—demographics, comorbidities and concomitant medication) • Validate the population PK/PD model • Establish or confirm dose exposure–response relationship in special population

relationship. For example, a relationship between exposure to monoclonal antibodies and their clinical effect (response or survival) has been reported for rituximab in non-Hodgkin's lymphoma [86], for cetuximab in metastatic colorectal cancer [87] and in head and neck cancer [88], as well as for bevacizumab in metastatic colorectal cancer [89]. They can enhance our use of protein kinase inhibitors (PKIs), as 518 kinases are known [90], and only 34 PKIs have received marketing authorization. In immunotherapy, CTS may provide a better definition of the recommended dose because, in contrast to conventional chemotherapy dosing, a maximum tolerated dose (MTD) approach does not always result in better clinical outcomes for novel targeted therapies because their efficacy is often robust at lower pharmacologically active doses at the MTD (e.g. patients with melanoma treated by pembrolizumab [91] or with metastatic renal cell carcinoma treated by nivolumab [92]). In 2016, Postel-Vinay *et al.* [93] reviewed several Phase 1 trials of immunotherapy and showed the inability of almost all trials to identify an MTD. During this phase, CTS can also use an adaptive design (accumulating data to modify certain future aspects of the study without undermining the validity and integrity of the trial) to simulate dose escalation during a Phase I cancer trial study design. Dose escalation can be guided by a mathematical model, combined with a Bayesian approach to provide a continuous flow of information [94]. Several methods exist, such as continual reassessment method (CRM), time-to-event continual reassessment method (TITE-CRM), escalation with overdose control (EWOC) and the time-to-event escalation with overdose control (TITE-EWOC) method [95–98]. Dose escalation can also be guided by a PK/PD model. For example, in 2014, Chalret du Rieu *et al.* [99] built a PK/PD model to determine the recommended dose predicting thrombocytopenia following administration of abexinostat in both lymphoma and solid tumor patients.

In this way, CTS can be used as a learning tool teaching us 'how to use the drug in representative patients so as to make acceptable benefit/risk likely', as well as to provide confirmation on 'how to use the drug in a large and representative patient population so that acceptable benefit/risk is achieved' [35, 100].

During confirmatory studies

The adaptive CTS design [101] can be aimed in particular at determining the timing of the interim analysis (group-sequential design). For example, sample size can be re-estimated using the information obtained during a confirmatory study to adjust the necessary sample size [102] (more flexible method than the fixed simple size approach) or during a combined Phase II/III study [103–105] in which the primary end point is the selection dose with a control group and different dose groups in two stage. These methods are effective and are increasingly used with the growth of interest in personalized medicine and targeted therapies in oncology [106, 107]. They can also improve subpopulation analysis and extend postmarketing authorization to help in choosing the best outcome [108–113]. They can also be used to optimize experimental designs [114] so as to better anticipate deviations from the protocol [115–120]. For instance, Bajard *et al.* [116] compared seven experimental designs for randomized clinical trials using *in silico* simulations to assist selection of the experimental design for a future clinical trial. Compliance with the protocol must allow evaluation of the effects of treatment with the expected or at least, with sufficient statistical power. In reality, deviations from the protocol (e.g. low recruitment [121], fewer included patients than expected [122–124], underestimated or overestimated effect of treatment, noncompliance [125], patient dropout or missing data [125–134]) may prevent a trial from achieving its objectives.

CTS projects in oncology

In this section, we describe the objectives and results of 10 CTS that we identified. In 2000, Veyrat-Follet *et al.* [135] simulated a Phase III trial on lung cancer patients to assess whether 125 mg/m^2 of docetaxel versus 100 mg/m^2 of docetaxel improved survival. The simulation results showed a slight improvement in median survival in the 125 mg/m^2 docetaxel group compared with the 100 mg/m^2 group (5.49 months versus 5.31 months). Despite encouraging results, the difference was significant in only 6 of 100 trials. Consequently, given the low power to detect a difference because of dose intensification, the investigators elected not to perform such a trial. In 2002, Jumbe *et al.* [112] performed a CTS to assess the feasibility of a fixed 200 μg dose of darbepoetin alfa administered every 2 weeks, compared with a weight-based dose of 3 μg/kg every 2 weeks in chemotherapy-induced anemia. The results in this article indicated that a fixed 200 μg dose of darbepoetin alfa administered every 2 weeks would be as effective as a weight-based dose of 3 μg/kg every 2 weeks in ameliorating anemia in patients with solid tumors receiving chemotherapy. In 2011, Paule *et al.* [136] performed CTS in capecitabine-induced hand-foot syndrome designed to reduce graded toxicity while maintaining efficacy. The results showed that individualized dose adaptation reduced the average duration of intolerable hand-foot syndrome by 10 days as compared with standard reductions, without compromising antitumor efficacy. The same year, Ternant *et al.* [137] performed CTS to quantify the benefits of the new dose regimen for rituximab to improve the PFS of patients with non-Hodgkin's lymphoma using a previously validated PK/PD model and to design clinical trials investigating optimization of rituximab dosage. The CTS results reported suggest that an improvement in PFS of patients with non-Hodgkin's lymphoma may be obtained by increased doses of rituximab. In 2012, Claret *et al.* [82] used CTS in an attempt to predict Phase III survival outcome using a drug–disease model framework and Phase II data. The results showed the usefulness of CTS for predicting drug activity and clinical efficacy results in Phase III based on Phase II data. However, the prediction model did not discriminate between failed and successful studies, suggesting that it is necessary to improve prediction of Phase III outcomes. In 2009, Ozawa *et al.* [138] performed CTS to evaluate the dose-reduction strategy of docetaxel in Japanese patients with liver dysfunction (standard dose of 60 mg/m^2 versus reduced dose of 40 or 50 mg/m^2). The simulation results showed that the median proportion of patients who experienced febrile neutropenia was decreased by about 20% in the reduced dose arm ($P < 0.05$) with no decrease in OS. This CTS has made it possible for authors to propose reducing the dose of docetaxel according to the extent of liver dysfunction in Japanese cancer patients. In 2014, Van Hasselt *et al.* [139] performed CTS to optimize a drug–drug interaction of vincristine with azole antifungals in the pediatric oncology population. They demonstrated the benefits of CTS for evaluation of PK clinical trial designs. In 2015, Lim *et al.* [140] used CTS to predict the efficacy of an oral paclitaxel (DHP107) formulation compared with intravenous paclitaxel. In terms of efficacy, DHP107 administration was predicted to show similar or greater treatment efficacy compared with intravenous paclitaxel. The results of their PK/PD modeling provided valuable

insights into paclitaxel that will enable accurate characterization of this compound and help improve future drug development processes. In 2015, Lim *et al.* [141] pursued two objectives: (1) to assess the accuracy with which individual patient-level exposure can be determined, and (2) whether a known food effect can be identified by CTS for a typical pharmacokinetic trial population. The authors demonstrated that CTS can be used to explore the ability of specific trial designs to evaluate the power to identify individual- and population-level exposures, covariate and variability effects. Finally, in 2015, Mazzocco *et al.* [80] performed CTS to demonstrate how a model representing the tumor size dynamics of low-grade adult gliomas before and after treatment with chemotherapy plus radiotherapy versus radiotherapy alone could be used as a tool to propose more effective therapeutic strategies. The results showed that the mean expected duration of the response was 4.3 years for low-grade gliomas treated with chemotherapy plus radiotherapy compared with 3.1 years in patients treated with radiotherapy alone ($P < 0.001$). The authors concluded that CTS could facilitate clinical research by helping to identify potentially more effective therapeutic strategies.

Principles underlying CTS

The pharmacometrics team

The first need is to put together a 'pharmacometrics team' incorporating statisticians/mathematicians, biologists, oncopharmacologists, clinicians and perhaps even a health economist. This pooling of specific professional skills will enable more effective sharing of knowledge in accordance with good practice guidelines issued by regulatory agencies [142], but also permit development of a relevant clinical/mathematical model. To perform a CTS, the team must first produce a detailed written simulation plan [115, 143] affording 'clarity', 'completeness' and 'parsimony', according to Holford's description [37, 38, 144], while also taking into account the specificity of the pathology. In oncology, for example, the plan should incorporate the notion of time-to-event [145].

Virtual patients

The models developed by the pharmacometrics teams are then applied to so-called 'virtual patients' (VPs). This term, in fact, is a misnomer, as these VP patients are defined only by a limited number of covariates (age, gender, weight, genotype or phenotype and biomarker) [146] that do not explain the complex interactions between the biology of an organism, the disease and treatment [147]. These variables can be generated by a model [125] or resampled from an existing database [148] (a concept developed only recently). For example, gender can easily be simulated from the binomial distribution, and age from a truncated log-normal distribution. However, when pathophysiological covariates are introduced, correlations between covariates cannot be ignored. Typically, size and weight are related to gender, and renal function to age. The problem becomes more complex when covariates vary over time, as is the case for the tumor expression of epidermal growth factor receptor in colorectal cancer [149], which requires a longitudinal model to describe these changes in more correlations. In the same way, it is of potential importance in the model to take into account the biological profile of the tumor, including oncogenic mutation (drivers) resistance-related factors [150–153]. The greater the number of covariates introduced, the more realistic the VP will be, but also the greater the risk of simulating highly improbable patients because of our inadequate understanding of the biological complexities of human beings. An alternative to the simulated VP [154] is the use of an epidemiological database from which actual patients can be resampled. This technique has led to the creation of so-called 'hybrid' patients, which are partly characterized by resampled data and partly by simulated data.

Big data, data sharing and data merging

The sample size is often tiny compared with the area in which we want to extrapolate its findings. The purpose of clinical trials is to validate empirical assumptions. When these assumptions cease to be verified, the process stops. This may be good news if the treatment is ineffective and dangerous. This can be bad news if the treatment is efficacious. However, it can appear bad simply because a number of nonresponders happened to be sampled. The main issue is to determine why a patient will respond or not to a treatment. All clinical experts agree that there are several profiles of responders and nonresponders. This is one of the reasons behind the development of personalized medicine. What is the probability, in a sample containing scores or hundreds of patients, that all the subgroups in a given sample are present in proportions comparable with those in the parent population? The probability is almost zero. Finally, today, there is no way to verify whether those populations were sampled in a balanced manner with respect to the parent population. Nevertheless, conclusions are drawn and decisions are taken based on the average of the study, on the assumption that the sample data are representative of reality. As is well known, these assumptions are never verified. A solution to this problem is the creation of a large database integrating multiple geographic sites. To obtain an accurate picture, the more the data available, the more representative they will be of reality. Databases are repositories of scientific literature and/or preclinical and clinical data (social, clinical, biological, imaging, genomic and quality of life data). Consequently, the stored data, harvested from hundreds of thousands of patients—big data (BD)—offer an excellent and unprecedented means to enhance the therapy of each individual patient [155–159]. However, from the methodological and informatics points of view, processing this vast amount of data is a massive challenge [160, 161].

BD is composed of two major types of data: structured and unstructured data [155, 162, 163]. Structured data are those whose set of possible values is determined and known in advance (age or sex). Unstructured data are impossible to categorize a priori (radiological data images, surgery reports or medical records). Given that 80–90% of the data are unstructured and only partly used [164], we have yet to take advantage of the immense potential these data sets can offer [165, 166]. In the field of oncology, access to comprehensive clinical data is often restricted [167]. However, several large consortiums, such as The Cancer Genome Atlas [168], the American Association for Cancer Research, Project Genomics, Evidence, Neoplasia, Information, Exchange (GENIE) [169] or the International Cancer Genome Consortium (ICGC) [170], have been created to generate large repositories of cancer samples and provide both molecular profiling and comprehensive clinical information. Data are provided in standardized format, allowing easy parsing of the information, and computational tools are available to help researchers explore the data easily. Table 2 provides a summary of existing repositories and tools. These repositories can be mined to refine our knowledge of the disease, and potentially define subgroups of patients based on molecular signatures

Table 2. Repositories and tools existing in oncology

Name	Description	URL	Reference
TCGA	High-throughput genetic characterization of > 30 human cancers	https://cancergenome.nih.gov/	[166]
Project GENIE	High-throughput genetic characterization of 59 types of humans cancers	http://www.aacr.org/research/research/pages/aacr-project-genie.aspx	[167]
ICGC	Comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes	http://icgc.org/	[168]
Cancer Cell Line Encyclopedia	Genetic characterization of a large panel of human cancer cell lines	http://www.broadinstitute.org/ccle	[170]
Gene Expression Omnibus	High-throughput functional gene expression data	https://www.ncbi.nlm.nih.gov/geo/	[174]
ArrayExpress	High-throughput functional gene expression data	https://www.ebi.ac.uk/arrayexpress/	[175]
NIH LINCS	High-throughput assays of cancer cell lines under various types of perturbations	http://www.lincsproject.org/LINCS/	[176]
TARGET	High-throughput genetic characterization of human childhood cancers	https://ocg.cancer.gov/programs/target	
Tools	**Description**	**URL**	
cBioPortal	Download, visualization and analysis of large-scale cancer genomics data sets	http://www.cbioportal.org/	
UCSC Xena	Analysis and visualization of public genomic/phenotypic data sets	http://xena.ucsc.edu/	

[171]. It will be crucial to take these molecular classifications into account when generating VP in our CTS, as different subtypes of cancer may show different responses to treatment [172]. Two important computational tools in preclinical studies were published recently [76, 173]. The first, named phenotypic network inference model, is a network-based inference algorithm for potential adverse drug event prediction during drug discovery. The authors tested their method on the MetaADEDB database they developed. This database (updated monthly), combining three databases (CTD, SPIDER and OFFSIDES), is an accessible, free-of-charge, comprehensive database of adverse drug events (ADE) and includes >3000 drugs, 13 000 adverse drug events and 520 000 drug–ADE associations. The authors concluded that their method provided a high potential prediction for adverse drug events. The second, named admetSAR (absorption, distribution, metabolism, excretion, toxicity structure−activity relationship) includes 22 qualitative classification models, and five quantity regression models developed using support vector machine algorithm. admetSAR is also a free-of-charge database compounded of >210 000 ADMET measurements of 96 000 unique compounds. The authors demonstrated that admetSAR provided more accurate prediction of the biodegradability of 27 novel compounds compared with two other models.

However, caution is required, as data sharing and data merging raise legal, political and technical issues [177], e.g. when merging databases involves associating, mixing and including data from disparate sources (interoperability [178–180], deidentification of data [181, 182] and quality information). To illustrate the concept of database merging techniques, a novel approach has recently been introduced [183]. Consequently, new skills are needed, especially programmers (data specialists) with expertise in transforming data into a format such as a relational database structure usable by researchers.

ML and software

ML may come to play a central role in CTS [184–186], a development that would make perfect sense with the arrival of BD. ML is a mode of artificial intelligence involving the development of algorithms to obtain a predictive analysis using data for a specific purpose. ML is divided into two types: supervised learning and unsupervised learning [184, 187, 188]. There are two other families of algorithms (semi-supervised learning and reinforcement learning) that we will not examine in detail, as they are little used for the moment in practice. Supervised learning is like building a correlation model between two or more variables known a priori—it 'feels' like a link between certain variables. There exist two major families:

1. Classification: This consists in allocating to their appropriate class new objects using known prior example (logistic regression, decision tree, support vector machine, naive Bayes classifier, random forest or random survival forest . . .)
2. Regression: This predicts the possible values of one or more variables from the from known values (linear regression, regression tree, splines, least squares regression, Cox regression . . .)

In contrast, unsupervised learning takes into account all the variables in a problem and extracts the strongest correlations. There exist two major families:

1. Clustering: This seeks to group data by similarity, without any prior information. It is also a form of classification, the difference being that the classes are not identified upstream, but emerge from the exploration of the data structure (*K*-means clustering, hierarchical clustering, artificial neural network/deep learning. . .)
2. Dimensionality reduction: The problem of transforming attributes in a high-dimensional space to a space of fewer dimensions [principal component analysis (PCA), factor analysis, kernel PCA and partial least squares]

Algorithms based on ML are useful only if the learning data set is complete, accurate and sufficiently large. Therefore, large, prospective, multicenter databases (including radiology, clinical, genetic and surgical) are needed to create more reliable algorithms that could greatly improve their usefulness in clinical and clinical research. One of the advantages of these techniques

is the reduction of the number of dimensions. But beware! If the number of dimensions is large, the reduction may not be sufficient to reduce the number of dimensions without loss of information. Moreover, these techniques are sensitive to outliers, so it is better to detect these errors and then to standardize the data before performing the operations. Furthermore, each of these computational methods has advantages and disadvantages, which we summarize on Table 3.

However, the ML method has proved effective for the analysis of microarray data in oncology [189, 190] and in radiology [156, 191]. Indeed, these algorithms detect groups of genes with similar expression and establish the genetic profile of the tumor. It is reasonable to assume that these methods are applicable to individuals characterized by a large number of clinical and biological parameters. The aim is to form groups such that the individuals contained within each group are as similar as possible according to their characteristics and so that the groups are as dissimilar as possible between themselves. Accordingly, it is conceivable to identify profiles of responders and/or drug-tolerant individuals [72, 192]. This approach is more powerful because it eliminates subjective human prejudgments and highlights hidden correlations that no human mind could have envisaged.

In terms of data processing, conventional statistical analysis methods must be abandoned and replaced by these novel learning techniques. Indeed, we must be careful to avoid overpowering the resulting multiplication in the statistical tests [193]. With the recent introduction of medical and laboratory information systems, we believe that as the volume of clinical data grows, both in the number of records and the number of variables stored, ML tools may become increasingly important.

Simulation statements require special software, storage space and adequate computing power calling on commercial software packages such as Matlab®, SAS®, EAST®, CERTANA's solutions [194] or open-source software such as Scilab, WinBugs and R®. The question of computational solutions has been addressed by Nyberg *et al.* [195] and Schadt *et al.* [160]. These latter presented the main categories of high-performance computing platforms with their advantages and their drawbacks.

Figure 3 depicts the set of points, which have been developed above. To summarize, the development of a model to demonstrate the initial hypothesis formulated by a clinician/pharmacist can be achieved through the consolidation of specific professional skills (pharmacometrics team). The use of data (BD, data sharing and data merging), drug models, trail design strategies and specific software and algorithms (ML and software) must enable virtual experiments to be performed. The validation of a hypothesis can be divided into two parts. The first entails internal validation: 'Does my simulation reproduce the data used to build my model?'; 'Are these results clinically relevant?' The second involves external validation of the initial model by comparing new data observed with those that were not used in its preparation. In some cases, validation is performed only after the fact, as, by definition, the simulation of clinical trials is designed to predict the results of a test that has not yet been performed. At the end of the cycle, all this can be summed up in a simple question: 'Has the hypothesis been confirmed?'

Good practices during CTS

It is recognized that, to be useful and credible, a CTS should be undertaken according to one type or other of good practice to homogenize the level of quality. For this, it is imperative that a simulation plan be developed, documenting precisely how the CTS will be conducted and how several clinically pertinent scenarios will be tested. Each scenario must be based on clearly identified assumptions. Simulation plan must include all the steps necessary for its realization, i.e. from conception to presentation of results, as performed by Holford et al. in 1999 [37], Smith and Marshall in 2010 [115], Giovagnoli in 2012 [196] and Marshall *et al.* [197] and Kelly *et al.* in 2016 [143]. It is necessary to take into account the level of complexity of a CTS, which may be simple or complex and may have different levels of importance (low, medium and high) as defined by the European Medicines Agency [198]. Moreover, all similarities and differences between the development of a simulated clinical trial and the development of a 'real' clinical trial must be taken into account. Rigor in the planning and execution of CTS will ensure that the design, analysis [199] and the decision-making process for the actual clinical trial are based on credible evidence that will need to be verified independently. This should allow readers not only to understand how the CTS was developed but also to enable them to reproduce it.

We provide several key points for expanding CTS:

1. Better sharing of knowledge and data between industry and/or institutions.
2. Priority should be given to sensitive areas such as pediatric diseases, which are rare, or where pathologies or recruitment of patients is restricted (e.g. oncology glioblastoma Grade IV, cholangiocarcinoma or hereditary cancers). The development of simulated clinical trials to permit the reduction of product tests on animals.
3. To prove the concept of these methods, we recommend comparing (and publishing) retrospectively the results obtained *in silico* with those obtained *in vivo*.
4. We recommend the double-blind implementation of CTS in parallel with *in vivo* clinical trials to best fit the models according to the available data.

However, with the development of CTS, simulation plans could in future form part of a review process with regulatory authorities, as is currently the case with any 'standard' clinical research protocol.

CTS challenges and prospects

Despite progress, the high rate of failure in R&D in industrial health and the steady rise in costs are wasting too many financial and human resources. This results in higher costs of new drugs and threatens the balance of health budgets in every country. Furthermore, the inclusion of patients in therapeutic trials whose clinical benefit may be described as 'marginal' reduces the number of patients that can be included for testing an innovative drug. In the United States, a number of oncology experts even hold that the moral red line between reasonable profits and profits has been exceeded [11, 200]. Consequently, changes are needed to achieve better health outcomes for all. The CTS approach aims to enhance learning during early studies and support proof of concept, early clinical decisions and design trials. However, this procedure has not yet been widely integrated into the oncology drug development process. It is essential for those involved in clinical research, whether industrial or academic, to adopt this methodology. To this end, it is essential that examples of CTS providing proof of concept are published, despite data privacy constraints to bring the CTS procedure to a wider audience and demonstrate both its benefits and limitations. The use of CTS in drug development is strongly advised by the regulatory agencies FDA and EMA. The FDA has been proactive in the development of regulatory science to

Table 3. Advantages and disadvantages of different ML algorithms

Algorithms	Advantages	Disadvantages
Logistic regression*	• Also easy to understand (odds ratio) • Computationally easy	• Variables must be linearly independent • Sensitive to outliers (continuous variables) • Do not treat missing values • Tendency for the model to overfit • Sometimes too simple to capture complex relationship between variables
Linear regression*	• Easy to understand—you clearly see what the biggest drivers of model are • Computationally easy	• Variables must be linearly independent • Sensitive to outliers • Do not treat missing values • Tendency for the model to overfit • Sometimes too simple to capture complex relationship between variables
Cox regression*	• Also easy to understand (hazard ratio) • Quick to converge • Robust and forgiving model • Commonly used	• The proportional hazards assumption • No absolute risks can be computed • Does not handle truly aggregated data • Really requires access to reliable cause-specific death rates
Decision tree	• Conceptually simple • Easy to understand and implement • Handle missing values well • Resistance to outliers • No normality assumptions	• Classes must be mutually exclusive • Results depend on the order of attribute selection • Risk of overly complex decision trees • Trees: unstable high variance • Lack of smoothness • Step function: values not always accurate
Random forest/ random survival forest	• Easy to use • Can be applied to decision trees to reduce instability • Resistance to outliers • No risk of overfitting • Fully parallelizable	• Extreme values often poorly estimated in case of regression • Regression cannot predict beyond range in the training data
Naïve Bayesian	• Missing values omitted • Easy to implement • Requires a small amount of training data to estimate the parameters • It predicts accurate results for most of the classification and prediction problems	• The precision of algorithm decreases if the amount of data is less • For obtaining good results, it requires a large number of records • Assumptions: class conditional independence, therefore loss of accuracy. Practically, dependencies exist among variables
K-nearest neighbors	• Classes need not be linearly separable • Zero cost of the learning process • Sometimes it is robust with regard to noisy training data • Well suited for multimodal classes	• Time to find the nearest neighbors in a large training data set can be excessive • It is sensitive to noisy or irrelevant attributes • Performance of algorithm depends on the number of dimensions used • Computationally expensive
Support vector machine	• Robust model • High accuracy • Work well even if data are not linearly separable in the base feature space	• Speed and size requirement both in training and testing is more • Sensitivity to choice of kernel parameters (test many values) • Risk of overfitting • High complexity and extensive memory requirements for classification in many cases
Artificial neural network/deep learning	• It is easy to use, with few parameters to adjust • A neural network learns and reprogramming is not needed • Easy to implement • Applicable to a wide range of problems in real life	• Requires high processing time if neural network is large • Risk of overfitting • Difficult to know how many neurons and layers are necessary • Learning can be slow • Difficult to understand predictions
K-means	• Easy to implement (Matlab, Python, SAS and R) • Complexity is linear to the number of individuals • Continuous improvement of classroom quality • Detection (with SAS) of outliers)	• It can converge to local optimum • Too sensitive to outliers, as an object with an extremely large value may substantially distort distribution of data • All items forced into a cluster • Need to specify k, the number of cluster, in advance • Not suitable to discover clusters with non-convex shapes

Note: *Can only be used in low-dimensional data (for high-dimensional data, analysis needs to be realized using penalized approach as ridge, lasso or elastic net).

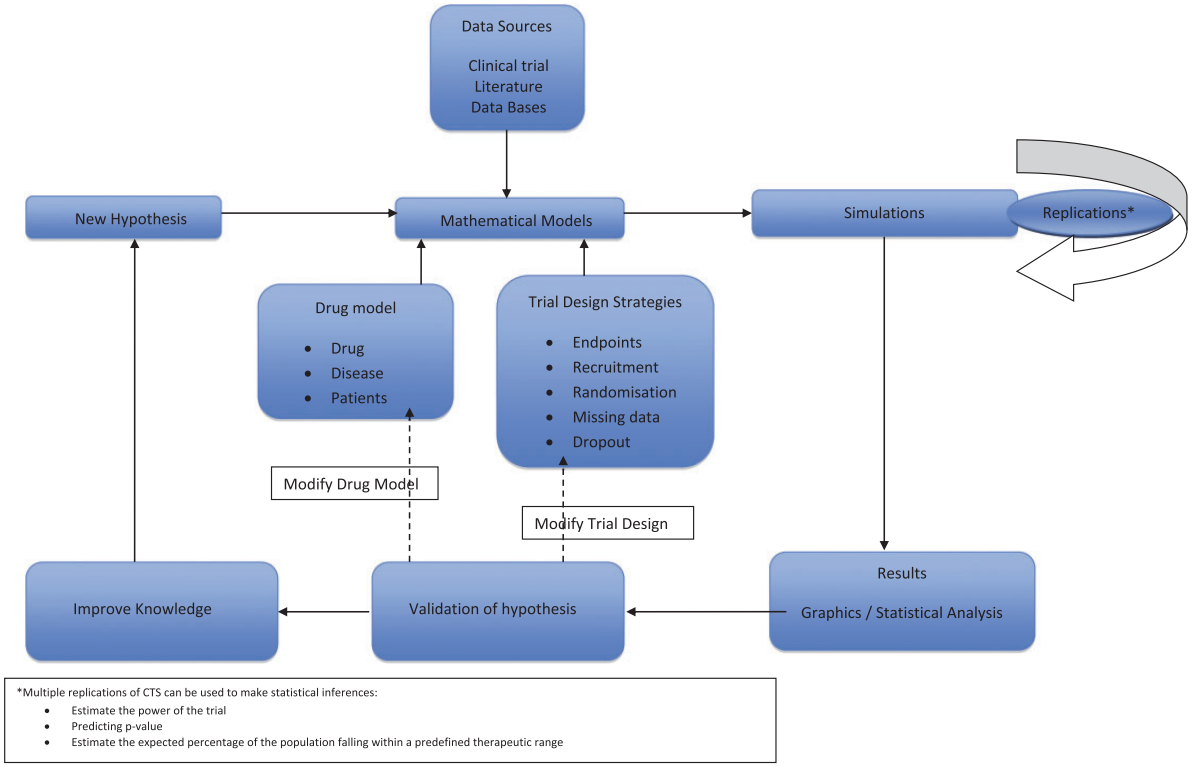


Figure 3. Schematic for a CTS model.

address an array of public health challenges [201]. Several FDA and International Council for Harmonization guidelines outlining the relevance of modeling in various areas of drug development have been published [36, 202, 203]. Similarly, in January 2013, the EMA created the Modeling and Simulation Working Group [204] to provide specialist scientific support to the Scientific Advice Working Party. However, while real progress has been made, a genuine common policy among the regulatory agencies designed to standardize certain rules, and more particularly those regarding data-sharing, is not yet forthcoming. The current rules are too dependent on individual countries or organizations. In a word, a global interactive process pooling all regulatory agencies is needed to assist drug developers and provide them with more coherent guidance but also, importantly, to protect patients. The modeling of drug activity coupled with the modeling of disease, optimal use of medical data already collected over a number of years and accelerating computing speed should place this goal within our reach. CTS, with their potential to improve the efficiency of drug development and optimize the design of clinical trials, must be extended in the future to provide quantitative support for drug development decisions [205] while never being used to replace 'real-life' clinical trials. CTS could in the future allow us to better understand certain mechanisms of tumor growth kinetics that we still do not control. For example, some patients with head and neck squamous cell carcinoma treated with anti programmed death/programmed death-ligand 1 (PD1/PD-L1) are hyperprogressors for reasons that are still unknown to us [206, 207]. However, as we have described, there remains a large number of technical and sociological hurdles to overcome before a standard procedure implementation of the 'digital patient' concept becomes a reality. This will require not only bioinformatics tools to facilitate interconnection and global integration of all individual data [208] (SHIVA clinical trial [60]) but also a universal legal framework to protect the privacy of every patient. This CTS approach is paving the way to P4 medicine: predictive, preventive, personalized and participatory [209]. Indeed, personalized medicine is one of the most promising approaches in cancer, as it aims to treat each patient individually in terms of the genetic and biological characteristics of their tumor but also taking into account the patient's environment, way of life, etc. The improvement of cancer treatments necessarily involves clinical trials. It is essential that new clinical trials adapt to this new medicine and patient inclusion must be done according to his/her genetic and biological profile [60, 61]. A paradigm shift is necessary to bring the benefits of CTS-based drug development to cancer patients, in which biomarkers and prognostic markers of OS are assessed to predict treatment outcome and disease progression.

Key Points

- *In silico*: By analogy with the expressions *in vivo* and *in vitro*, the term *in silico* was introduced to describe the numerical methods used to treat biological systems in the context of clinical research. The Latin term refers to 'silicon', the main material found in computer chips of all computers. The *in silico* field includes a broad set of numerical methods involving a mathematical approach to models and simulates biological phenomena using computers.
- Modeling and simulation are sometimes used synonymously because they both use an abstraction of some real system for prediction and control.
  - Modeling: It refers to the development of a mathematical representation of an entity, system or process.

In our case, this could be the action of a new drug on a tumor cell.
- Simulation: It refers to the procedure of solving on a computer the mathematical equations that resulted from model development.

- ML is the scientific discipline that focuses on how computers learn from clinical data to make predictions (e.g. toxicity, response or survey).

Acknowledgements

The authors sincerely thank Mr George Morgan for reviewing this article.

References

1. Chakravarthy R, Cotter K, DiMasi J, *et al*. *Public and private sector contributions to the research and development of the most transformational drugs of the last 25 years*. Boston: Tufts Center for the Study of Drug Development, 2015.
2. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov* 2004;**3**:417–29.
3. Pandey S, Pandey P, Tiwari G, *et al*. Bioanalysis in drug discovery and development. *Pharm Methods* 2010;**1**:14–24.
4. Howie L, Peppercorn J. The ethics of clinical trials for cancer therapy. *N C Med J* 2013;**75**:270–3.
5. Shamy MC, Stahnisch FW, Hill MD. Fallibility: a new perspective on the ethics of clinical trial enrollment. *Int J Stroke* 2015;**10**:2–6.
6. Rosenblatt M. The large pharmaceutical company perspective. *N Engl J Med* 2017;**376**:52–60.
7. Tang C, Sherman SI, Price M, *et al*. Clinical trial characteristics and barriers to participant accrual: the MD Anderson Cancer Center experience over 30 years, a historical foundation for trial improvement. *Clin Cancer Res* 2017;**23**:1414–21.
8. Venkatakrishnan K, Friberg LE, Ouellet D, *et al*. Optimizing oncology therapeutics through quantitative translational and clinical pharmacology: challenges and opportunities. *Clin Pharmacol Ther* 2015;**97**:37–54.
9. Rubin EH, Gilliland DG. Drug development and clinical trials–the path to an approved cancer drug. *Nat Rev Clin Oncol* 2012;**9**:215–22.
10. Fojo T, Mailankody S, Lo A. Unintended consequences of expensive cancer therapeutics-the pursuit of marginal indications and a me-too mentality that stifles innovation and creativity: the John Conley Lecture. *JAMA Otolaryngol Head Neck Surg* 2014;**140**:1225–36.
11. Light DW, Lexchin J. Why do cancer drugs get such an easy ride. *BMJ* 2015;**350**:h2068.
12. Kinch MS, Haynesworth A, Kinch SL, *et al*. An overview of FDA-approved new molecular entities: 1827–2013. *Drug Discov Today* 2014;**19**:1033–9.
13. Paul SM, Mytelka DS, Dunwiddie CT, *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;**9**:203–14.
14. deepMind. https://deepmind.com/blog/applying-machine-learning-radiotherapy-planning-head-neck-cancer/.
15. IBM. Watson. http://www-05.ibm.com/innovation/uk/watson/watson_in_healthcare.shtml.
16. Chen Y, Elenee Argentinis JD, Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clin Ther* 2016;**38**:688–701.
17. Google. TensorFlow is an open source software library for machine intelligence. https://www.tensorflow.org/2016.
18. Rampasek L, Goldenberg A. TensorFlow: biology's gateway to deep learning? *Cell Syst* 2016;**2**:12–4.
19. Systems D. Biointelligence. https://www.3ds.com/stories/biointelligence/ 2016.
20. Mendoza RL. The 21st century cures act: pharmacoeconomic boon or bane? *J Med Econ* 2017;**20**:315–17.
21. Upton F, DeGette D, Pitts J. HR 6–21st Century Cures Act. 2015.
22. Mateen FJ, Oh J, Tergas AI, *et al*. Titles versus titles and abstracts for initial screening of articles for systematic reviews. *Clin Epidemiol* 2013;**5**:89–95.
23. Jelliffe RW. Computer-controlled administration of cardiovascular drugs. *Prog Cardiovasc Dis* 1983;**26**:1–14.
24. Levy G, Gibaldi M, Jusko WJ. Multicompartment pharmacokinetic models and pharmacologic effects. *J Pharm Sci* 1969;**58**:422–4.
25. Peck CC, Sheiner LB, Martin CM, *et al*. Computer-assisted digoxin therapy. *N Engl J Med* 1973;**289**:441–6.
26. Rodman JH, Jelliffe RW, Kolb E, *et al*. Clinical studies with computer-assisted initial lidocaine therapy. *Arch Intern Med* 1984;**144**:703–9.
27. Sheiner LB, Rosenberg B, Marathe VV. Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. *J Pharmacokinet Biopharm* 1977;**5**:445–79.
28. Aarons L, Balant LP, Mentre F, *et al*. Population approaches in drug development. Report on an expert meeting to discuss population pharmacokinetic/pharmacodynamic software. *Eur J Clin Pharmacol* 1994;**46**:389–91.
29. Peck CC, Barr WH, Benet LZ, *et al*. Opportunities for integration of pharmacokinetics, pharmacodynamics, and toxicokinetics in rational drug development. *J Clin Pharmacol* 1994;**34**:111–9.
30. Steimer J-L, Vozeh S, Racine-Poon A, *et al*. The population approach: rationale, methods, and applications in clinical pharmacology and drug development. In: Welling PG, Balant LP (eds). *Pharmacokinetics of Drugs*. Springer, 1994, 405–51.
31. Aarons L, Balant LP, Mentre F, *et al*. Practical experience and issues in designing and performing population pharmacokinetic/pharmacodynamic studies. *Eur J Clin Pharmacol* 1996;**49**:251–4.
32. Aarons L, Balant L, Danhof M, *et al*. *The Population Approach: Measuring and Managing Variability in Response, Concentration and Dose*. Luxembourg: Office for Official Publications of the European Communities, 1997.
33. Peck CC. Drug development: improving the process. *Food Drug Law J* 1997;**52**:163–7.
34. Vozeh S, Steimer JL, Rowland M, *et al*. The use of population pharmacokinetics in drug development. *Clin Pharmacokinet* 1996;**30**:81–93.
35. Sheiner LB. Learning versus confirming in clinical drug development. *Clin Pharmacol Ther* 1997;**61**:275–91.
36. FDA. Guidance for industry: population pharmacokinetics, 1999, US Center for Drug Evaluation and Research, http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/WomensHealthResearch/UCM133184.pdf.
37. Holford N, Hale M, Ko H, *et al*. *Simulation in Drug Development: Good Practices*. http://holford.fmhs.auckland.ac.nz/docs/simulation-in-drug-development-good-practices.pdf 1999.
38. Holford NH, Kimko HC, Monteleone JP, *et al*. Simulation of clinical trials. *Annu Rev Pharmacol Toxicol* 2000;**40**:209–34.

39. Bonate PL. Clinical trial simulation in drug development. *Pharm Res* 2000;**17**:252–6.
40. Gieschke R, Steimer JL. Pharmacometrics: modelling and simulation tools to improve decision making in clinical drug development. *Eur J Drug Metab Pharmacokinet* 2000;**25**:49–58.
41. Kimko H, Duffull SB. *Simulation for Designing Clinical Trials: A Pharmacokinetic-Pharmacodynamic Modeling Perspective*. CRC Press, 2002.
42. Holford N, Ma SC, Ploeger BA. Clinical trial simulation: a review. *Clin Pharmacol Ther* 2010;**88**:166–82.
43. Maxwell C, Domenet JG, Joyce CR. Instant experience in clinical trials: a novel aid to teaching by simulation. *J Clin Pharmacol New Drugs* 1971;**11**:323–31.
44. Madsen BW, Woodings TL, Ilett KF, *et al*. Clinical trial experience by simulation: a workshop report. *BMJ*1978;**2**:1333–5.
45. Lee KL, McNeer JF, Starmer CF, *et al*. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980;**61**:508–15.
46. Bland J. Computer simulation of a clinical trial as an aid to teaching the concept of statistical significance. *Stat Med* 1986;**5**:193–7.
47. Tiefenbrunn AJ, Graor RA, Robison AK, *et al*. Pharmacodynamics of tissue-type plasminogen activator characterized by computer-assisted simulation. *Circulation* 1986;**73**:1291–9.
48. Hoover SV, Perry RF, *Simulation: A Problem-Solving Approach*. Addison-Wesley Longman Publishing Co., Inc., 1989.
49. Kaufmann WJ, Smarr LL. *Supercomputing and the Transformation of Science*. WH Freeman & Co., 1992.
50. Johnson SC. The role of simulation in the management of research: what can the pharmaceutical industry learn from the aerospace industry? *Drug Inf J* 1998;**32**:961–9.
51. Sachs JR, Mayawala K, Gadamsetty S, *et al*. Optimal dosing for targeted therapies in oncology: drug development cases leading by example. *Clin Cancer Res* 2016;**22**:1318–24.
52. Karlsson MO, Sheiner LB. The importance of modeling interoccasion variability in population pharmacokinetic analyses. *J Pharmacokinet Biopharm* 1993;**21**:735–50.
53. Lauschke VM, Ingelman-Sundberg M. Requirements for comprehensive pharmacogenetic genotyping platforms. *Pharmacogenomics* 2016;**17**:917–24.
54. Mould DR, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development. *CPT Pharmacometrics Syst Pharmacol* 2012;**1**:e6.
55. Mould DR, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development-part 2: introduction to pharmacokinetic modeling methods. *CPT Pharmacometrics Syst Pharmacol* 2013;**2**:e38.
56. Upton RN, Mould DR. Basic concepts in population modeling, simulation, and model-based drug development: part 3-introduction to pharmacodynamic modeling methods. *CPT Pharmacometrics Syst Pharmacol* 2014;**3**:e88.
57. Biankin AV, Piantadosi S, Hollingsworth SJ. Patient-centric trials for therapeutic development in precision oncology. *Nature* 2015;**526**:361–70.
58. Goodsaid F. Challenges of biomarkers in drug discovery and development. *Expert Opin Drug Discov* 2012;**7**:457–61.
59. Le Tourneau C, Kamal M, Alt M, *et al*. The spectrum of clinical trials aiming at personalizing medicine. *Chin Clin Oncol* 2014;**3**:13.
60. Le Tourneau C, Delord JP, Goncalves A, *et al*. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol* 2015;**16**:1324–34.
61. Kaplan R, Maughan T, Crook A, *et al*. Evaluating many treatments and biomarkers in oncology: a new design. *J Clin Oncol* 2013;**31**:4562–8.
62. Kim ES, Herbst RS, Wistuba II, *et al*. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov* 2011;**1**:44–53.
63. Trippa L, Alexander BM. Bayesian baskets: a novel design for biomarker-based clinical trials. *J Clin Oncol* 2017;**35**:681–7.
64. Gosho M, Nagashima K, Sato Y. Study designs and statistical analyses for biomarker research. *Sensors* 2012;**12**:8966–86.
65. Zhang P, Brusic V. Mathematical modeling for novel cancer drug discovery and development. *Expert Opin Drug Discov* 2014;**9**:1133–50.
66. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature* 2012;**483**:531–3.
67. Barbolosi D, Iliadis A. Optimizing drug regimens in cancer chemotherapy: a simulation study using a PK-PD model. *Comput Biol Med* 2001;**31**:157–72.
68. Chen B, Dong JQ, Pan W-J, *et al*. Pharmacokinetics/pharmacodynamics model-supported early drug development. *Curr Pharm Biotechnol* 2012;**13**:1360–75.
69. Doudican NA, Kumar A, Singh NK, *et al*. Personalization of cancer treatment using predictive simulation. *J Transl Med* 2015;**13**:43.
70. Barbolosi D, Ciccolini J, Lacarelle B, *et al*. Computational oncology - mathematical modelling of drug regimens for precision medicine. *Nat Rev Clin Oncol* 2016;**13**:242–54.
71. Lefor AT. Computational oncology. *JPN J Clin Oncol* 2011;**41**:937–47.
72. Tao L, Zhang P, Qin C, *et al*. Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools. *Adv Drug Deliv Rev* 2015;**86**:83–100.
73. Roncaglioni A, Toropov AA, Toropova AP, *et al*. In silico methods to predict drug toxicity. *Curr Opin Pharmacol* 2013;**13**:802–6.
74. Agur Z. From the evolution of toxin resistance to virtual clinical trials: the role of mathematical models in oncology. *Fut Oncol* 2010;**6**:917–27.
75. Iliadis A, Barbolosi D. Optimizing drug regimens in cancer chemotherapy by an efficacy-toxicity mathematical model. *Comput Biomed Res* 2000;**33**:211–26.
76. Cheng F, Li W, Wang X, *et al*. Adverse drug events: database construction and in silico prediction. *J Chem Inf Model* 2013;**53**:744–52.
77. Bernard A, Kimko H, Mital D, *et al*. Mathematical modeling of tumor growth and tumor growth inhibition in oncology drug development. *Expert Opin Drug Metab Toxicol* 2012;**8**:1057–69.
78. Sanga S, Sinek JP, Frieboes HB, *et al*. Mathematical modeling of cancer progression and response to chemotherapy. *Expert Rev Anticancer Ther* 2006;**6**:1361–76.
79. Michelson S, Sehgal A, Friedrich C. In silico prediction of clinical efficacy. *Curr Opin Biotechnol* 2006;**17**:666–70.
80. Mazzocco P, Honnorat J, Ducray F, *et al*. Increasing the time interval between PCV chemotherapy cycles as a strategy to improve duration of response in low-grade gliomas: results from a model-based clinical trial simulation. *Comput Math Methods Med* 2015;**2015**:297903.
81. Bruno R, Mercier F, Claret L. Evaluation of tumor size response metrics to predict survival in oncology clinical trials. *Clin Pharmacol Ther* 2014;**95**:386–93.
82. Claret L, Lu JF, Bruno R, *et al*. Simulations using a drug-disease modeling framework and phase II data predict

phase III survival outcome in first-line non-small-cell lung cancer. *Clin Pharmacol Ther* 2012;**92**:631–4.
83. Bender BC, Schindler E, Friberg LE. Population pharmacokinetic-pharmacodynamic modelling in oncology: a tool for predicting clinical response. *Br J Clin Pharmacol* 2015;**79**:56–71.
84. Filleron T, Kouokam W, Gilhodes J, *et al*. Statistical controversies in clinical research: should schedules of tumor size assessments be changed? *Ann Oncol* 2016;**27**:1981–7.
85. Barbolosi D, Benabdallah A, Hubert F, *et al*. Mathematical and numerical analysis for a model of growing metastatic tumors. *Math Biosci* 2009;**218**:1–14.
86. Igarashi T, Kobayashi Y, Ogura M, *et al*. Factors affecting toxicity, response and progression-free survival in relapsed patients with indolent B-cell lymphoma and mantle cell lymphoma treated with rituximab: a Japanese phase II study. *Ann Oncol* 2002;**13**:928–43.
87. Azzopardi N, Lecomte T, Ternant D, *et al*. Cetuximab pharmacokinetics influences progression-free survival of metastatic colorectal cancer patients. *Clin Cancer Res* 2011;**17**:6329–37.
88. Pointreau Y, Azzopardi N, Ternant D, *et al*. Cetuximab pharmacokinetics influences overall survival in patients with head and neck cancer. *Ther Drug Monit* 2016;**38**:567–72.
89. Caulet M, Lecomte T, Bouche O, *et al*. Bevacizumab pharmacokinetics influence overall and progression-free survival in metastatic colorectal cancer patients. *Clin Pharmacokinet* 2016;**55**:1381–94.
90. Drenberg CD, Baker SD, Sparreboom A. Integrating clinical pharmacology concepts in individualized therapy with tyrosine kinase inhibitors. *Clin Pharmacol Ther* 2013;**93**:215–9.
91. de Greef R, Elassaiss-Schaap J, Chatterjee M, *et al*. Pembrolizumab: role of modeling and simulation in bringing a novel immunotherapy to patients with melanoma. *CPT Pharmacometrics Syst Pharmacol* 2017;**6**:5–7.
92. Motzer RJ, Rini BI, McDermott DF, *et al*. Nivolumab for metastatic renal cell carcinoma: results of a randomized phase II trial. *J Clin Oncol* 2015;**33**:1430–7.
93. Postel-Vinay S, Aspeslagh S, Lanoy E, *et al*. Challenges of phase 1 clinical trials evaluating immune checkpoint-targeted antibodies. *Ann Oncol* 2016;**27**:214–24.
94. Paoletti X, Ezzalfani M, Le Tourneau C. Statistical controversies in clinical research: requiem for the 3 + 3 design for phase I trials. *Ann Oncol* 2015;**26**:1808–12.
95. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* 1990;**46**:33–48.
96. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* 2000;**56**:1177–82.
97. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat Med* 1998;**17**:1103–20.
98. Chen Z, Cui Y, Owonikoko TK, *et al*. Escalation with overdose control using all toxicities and time to event toxicity data in cancer Phase I clinical trials. *Contemp Clin Trials* 2014;**37**:322–32.
99. Chalret du Rieu Q, Fouliard S, White-Koning M, *et al*. Pharmacokinetic/Pharmacodynamic modeling of abexinostat-induced thrombocytopenia across different patient populations: application for the determination of the maximum tolerated doses in both lymphoma and solid tumour patients. *Invest New Drugs* 2014;**32**:985–94.
100. Kimko H, Pinheiro J. Model-based clinical drug development in the past, present and future: a commentary. *Br J Clin Pharmacol* 2015;**79**:108–16.
101. Bretz F, Koenig F, Brannath W, *et al*. Adaptive designs for confirmatory clinical trials. *Stat Med* 2009;**28**:1181–217.
102. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;**30**:3267–84.
103. Jenniso C, Turnbull BW. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations. *Biom J* 2006;**48**:650–5. discussion 660-652.
104. Schmidli H, Bretz F, Racine A, *et al*. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biom J* 2006;**48**:635–43.
105. Bretz F, Schmidli H, Konig F, *et al*. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biom J* 2006;**48**:623–34.
106. Mehta C, Schafer H, Daniel H, *et al*. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Stat Med* 2014;**33**:4515–31.
107. Brannath W, Zuber E, Branson M, *et al*. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat Med* 2009;**28**:1445–63.
108. Altstein LL, Li G, Elashoff RM. A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial. *Stat Med* 2011;**30**:709–17.
109. Balis FM, Fox E, Widemann BC, *et al*. Clinical drug development for childhood cancers. *Clin Pharmacol Ther* 2009;**85**:127–9.
110. Bellanti F, Della Pasqua O. Modelling and simulation as research tools in paediatric drug development. *Eur J Clin Pharmacol* 2011;**67** (Suppl 1):75–86.
111. Dagher R, Cohen M, Williams G, *et al*. Approval summary: imatinib mesylate in the treatment of metastatic and/or unresectable malignant gastrointestinal stromal tumors. *Clin Cancer Res* 2002;**8**:3034–8.
112. Jumbe N, Yao B, Rovetti R, *et al*. Clinical trial simulation of a 200-microg fixed dose of darbepoetin alfa in chemotherapy-induced anemia. *Oncology* 2002;**16**:37–44.
113. Murgo AJ, Espinoza-Delgado I. Development of novel anticancer agents in older patients: pharmacokinetic, pharmacodynamic, and other considerations. *Cancer J* 2005;**11**:481–7.
114. Hmelo CE, Ramakrishnan S, Day RS, *et al*. Oncology thinking cap: scaffolded use of a simulation to learn clinical trial design. *Teach Learn Med* 2001;**13**:183–91.
115. Smith MK, Marshall A. Importance of protocols for simulation studies in clinical drug development. *Stat Methods Med Res* 2011;**20**:613–22.
116. Bajard A, Chabaud S, Cornu C, *et al*. An in silico approach helped to identify the best experimental design, population, and outcome for future randomized clinical trials. *J Clin Epidemiol* 2016;**69**:125–36.
117. Wolbers M, Helterbrand JD. Two-stage randomization designs in drug development. *Stat Med* 2008;**27**:4161–74.
118. Rashid I, Marcheselli L, Federico M. Estimating survival in newly diagnosed cancer patients: Use of computer simulations to evaluate performances of different approaches in a wide range of scenarios. *Stat Med* 2008;**27**:2145–58.
119. Cornu C, Kassai B, Fisch R, *et al*. Experimental designs for small randomised clinical trials: an algorithm for choice. *Orphanet J Rare Dis* 2013;**8**:48.

120. Nony P, Kurbatova P, Bajard A, *et al*. A methodological framework for drug development in rare diseases. *Orphanet J Rare Dis* 2014;**9**:164.
121. Jiang Y, Simon S, Mayo MS, *et al*. Modeling and validating Bayesian accrual models on clinical data and simulations using adaptive priors. *Stat Med* 2015;**34**:613–29.
122. Anisimov VV. Predictive event modelling in multicenter clinical trials with waiting time to response. *Pharm Stat* 2011;**10**:517–22.
123. Anisimov VV, Fedorov VV. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Stat Med* 2007;**26**:4958–75.
124. Mijoule G, Savy S, Savy N. Models for patients' recruitment in clinical trials and sensitivity analysis. *Stat Med* 2012;**31**:1655–74.
125. Westfall PH, Tsai K, Ogenstad S, *et al*. Clinical trials simulation: a statistical approach. *J Biopharm Stat* 2008;**18**:611–30.
126. Groves RM. Nonresponse in sample surveys. In: *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., 2005, 133–83.
127. Huisman M, Van der Zouwen J. Item nonresponse in scale data from surveys: Types, determinants, and measures. In: ME Huisman (ed). *Item Nonresponse: Occurrence, Causes, and Imputation of Missing Answers to Test Items*, Vol. **63**. Leiden: DSWO press, 1999, 63–90.
128. Rubin DB. Inference and missing data. *Biometrika* 1976;**63**:581–92.
129. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2014.
130. Schafer JL. *Imputation of Missing Covariates Under a Multivariate Linear Mixed Model*. Tech, 1997.
131. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *J Comput Graph Stat* 2002;**11**:437–57.
132. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media, 2009.
133. Little RJ. Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc* 1995;**90**:1112–21.
134. Marshall A, Altman DG, Royston P, *et al*. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol* 2010;**10**
135. Veyrat-Follet C, Bruno R, Olivares R, *et al*. Clinical trial simulation of docetaxel in patients with cancer as a tool for dosage optimization. *Clin Pharmacol Ther* 2000;**68**:677–87.
136. Paule I, Tod M, Henin E, *et al*. Dose adaptation of capecitabine based on individual prediction of limiting toxicity grade: evaluation by clinical trial simulation. *Cancer Chemother Pharmacol* 2012;**69**:447–55.
137. Ternant D, Cartron G, Henin E, *et al*. Model-based design of rituximab dosage optimization in follicular non-Hodgkin's lymphoma. *Br J Clin Pharmacol* 2012;**73**:597–605.
138. Ozawa K, Minami H, Sato H. Clinical trial simulations for dosage optimization of docetaxel in patients with liver dysfunction, based on a log-binominal regression for febrile neutropenia. *Yakugaku Zasshi* 2009;**129**:749–57.
139. van Hasselt JG, van Eijkelenburg NK, Beijnen JH, *et al*. Design of a drug-drug interaction study of vincristine with azole antifungals in pediatric cancer patients using clinical trial simulation. *Pediatr Blood Cancer* 2014;**61**:2223–9.
140. Lim HS, Bae KS, Jung JA, *et al*. Predicting the efficacy of an oral paclitaxel formulation (DHP107) through modeling and simulation. *Clin Ther* 2015;**37**:402–17.
141. Li CH, Sherer EA, Lewis LD, *et al*. Clinical trial simulation to evaluate population pharmacokinetics and food effect: capturing abiraterone and nilotinib exposures. *J Clin Pharmacol* 2015;**55**:556–62.
142. FDA. FDA pharmacometrics 2020 strategic goals. 2012.
143. O'Kelly M, Anisimov V, Campbell C, *et al*. Proposed best practice for projects that involve modelling and simulation. *Pharm Stat* 2016;**16**:107–13.
144. Onar-Thomas A, Xiong Z. A simulation-based comparison of the traditional method, Rolling-6 design and a frequentist version of the continual reassessment method with special attention to trial duration in pediatric phase I oncology trials. *Contemp Clin Trials* 2010;**31**:259–70.
145. Holford N. A time to event tutorial for pharmacometricians. *CPT Pharmacometrics Syst Pharmacol* 2013;**2**:e43.
146. Alkema W, Rullmann T, van Elsas A. Target validation *in silico*: does the virtual patient cure the pharma pipeline? *Expert Opin Ther Targets* 2006;**10**:635–8.
147. Bangs A. Predictive biosimulation and virtual patients in pharmaceutical R and D. *Stud Health Technol Inform* 2005;**111**:37–42.
148. Teutonico D, Musuamba F, Maas HJ, *et al*. Generating virtual patients by multivariate and discrete re-sampling techniques. *Pharm Res* 2015;**32**:3228–37.
149. Siravegna G, Mussolin B, Buscarino M, *et al*. Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nat Med* 2015;**21**:795–801.
150. Sameen S, Barbuti R, Milazzo P, *et al*. Mathematical modeling of drug resistance due to KRAS mutation in colorectal cancer. *J Theor Biol* 2016;**389**:263–73.
151. Sun X, Bao J, Shao Y. Mathematical modeling of therapy-induced cancer drug resistance: connecting cancer mechanisms to population survival rates. *Sci Rep* 2016;**6**:22498.
152. O'Donnell JS, Long GV, Scolyer RA, *et al*. Resistance to PD1/PDL1 checkpoint inhibition. *Cancer Treat Rev* 2017;**52**:71–81.
153. Martinelli E, Morgillo F, Troiani T, *et al*. Cancer resistance to therapies against the EGFR-RAS-RAF pathway: the role of MEK. *Cancer Treat Rev* 2017;**53**:61–9.
154. Zazzu V, Regierer B, Kuhn A, *et al*. IT future of medicine: from molecular analysis to clinical diagnosis and improved treatment. *N Biotechnol* 2013;**30**:362–5.
155. Andreu-Perez J, Poon CC, Merrifield RD, *et al*. Big data for health. *IEEE J Biomed Health Inform* 2015;**19**:1193–208.
156. El Naqa I. Perspectives on making big data analytics work for oncology. *Methods* 2016;**111**:32–44.
157. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;**309**:1351–2.
158. Gu J, Taylor CR. Practicing pathology in the era of big data and personalized medicine. *Appl Immunohistochem Mol Morphol* 2014;**22**:1–9.
159. Vicini P, Fields O, Lai E, *et al*. Precision medicine in the age of big data: the present and future role of large-scale unbiased sequencing in drug discovery and development. *Clin Pharmacol Ther* 2016;**99**:198–207.
160. Schadt EE, Linderman MD, Sorenson J, *et al*. Computational solutions to large-scale data management and analysis. *Nat Rev Genet* 2010;**11**:647–57.
161. Noor AM, Holmberg L, Gillett C, *et al*. Big Data: the challenge for small research groups in the era of cancer genomics. *Br J Cancer* 2015;**113**:1405–12.
162. Dumbill E. Getting up to speed with big data. *Big Data Now*, 2012.
163. Atienza AA, Serrano KJ, Riley WT, *et al*. Advancing cancer prevention and behavior theory in the era of big data. *J Cancer Prev* 2016;**21**:201–6.

164. Grimes S. *Unstructured Data and the 80 Percent Rule*. Carabridge Bridgepoints, 2008.
165. Cook JA, Collins GS. The rise of big clinical databases. *Br J Surg* 2015;**102**:e93–e101.
166. Luo L, Li L, Hu J, *et al*. A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. *BMC Med Inform Decis Mak* 2016;**16**:114.
167. Quackenbush J. Perspective: learning to share. *Nature* 2014;**509**:S68.
168. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol* 2015;**19**:A68–77.
169. Project GENIE Goes Public. *Cancer Discov* 2017;**7**:118.
170. International Cancer Genome Consortium, Hudson TJ, Anderson W, *et al*. International network of cancer genome projects. *Nature* 2010;**464**:993–8.
171. Chang JT, Lee YM, Huang RS. The impact of the Cancer Genome Atlas on lung cancer. *Transl Res* 2015;**166**:568–85.
172. Barretina J, Caponigro G, Stransky N, *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7.
173. Cheng F, Li W, Zhou Y, *et al*. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model* 2012;**52**:3099–105.
174. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2013;**41**:D991–5.
175. Kolesnikov N, Hastings E, Keays M, *et al*. Array express update–simplifying data submissions. *Nucleic Acids Res* 2015;**43**:D1113–6.
176. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 2016;**32**:2338–45.
177. Lo B. Sharing clinical trial data: maximizing benefits, minimizing risk. *JAMA* 2015;**313**:793–4.
178. Dolin RH, Rogers B, Jaffe C. Health level seven interoperability strategy: big data, incrementally structured. *Methods Inf Med* 2015;**54**:75–82.
179. McKeever S, Johnson D. The role of markup for enabling interoperability in health informatics. *Front Physiol* 2015;**6**:152.
180. Masseroli M, Kaitoua A, Pinoli P, *et al*. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods* 2016;**111**:3–11.
181. White SE. De-identification and the sharing of big data. *J AHIMA* 2013;**84**:44–7.
182. Kaplan B. Selling health data: de-identification, privacy, and speech. *Camb Q Healthc Ethics* 2015;**24**:256–71.
183. Villani C. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.
184. Deo RC. Machine learning in medicine. *Circulation* 2015;**132**:1920–30.
185. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016;**375**:1216–9.
186. DiMasi JA, Hermann JC, Twyman K, *et al*. A tool for predicting regulatory approval after phase II testing of new oncology compounds. *Clin Pharmacol Ther* 2015;**98**:506–13.
187. Tarca AL, Carey VJ, Chen XW, *et al*. Machine learning and its applications to biology. *PLoS Comput Biol* 2007;**3**:e116.
188. Larranaga P, Calvo B, Santana R, *et al*. Machine learning in bioinformatics. *Brief Bioinform* 2006;**7**:86–112.
189. Vidyasagar M. Identifying predictive features in drug response using machine learning: opportunities and challenges. In: Insel P. A. (ed). *Annu Rev Pharmacol Toxicol* 2015; Vol**55**:15–34.
190. Angermueller C, Parnamaa T, Parts L, *et al*. Deep learning for computational biology. *Mol Syst Biol* 2016;**12**:878.
191. Bibault JE, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: State of the art and future prospects. *Cancer Lett* 2016;**382**:110–7.
192. Maltarollo VG, Gertrudes JC, Oliveira PR, *et al*. Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin Drug Metab Toxicol* 2015;**11**:259–71.
193. Head ML, Holman L, Lanfear R, *et al*. The extent and consequences of p-hacking in science. *PLoS Biol* 2015;**13**:e1002106.
194. CERTARA. Optimize Your Drug Development Decisions With Certara. https://www.certara.com/.
195. Nyberg J, Bazzoli C, Ogungbenro K, *et al*. Methods and software tools for design evaluation in population pharmacokinetics-pharmacodynamics studies. *Br J Clin Pharmacol* 2015;**79**:6–17.
196. Giovagnoli A, Zagoraiou M. Simulation of clinical trials: a review with emphasis on the design issues. *Statistica* 2012;**72**:63.
197. Workgroup EM, Marshall SF, Burghaus R, *et al*. Good practices in model-informed drug discovery and development: practice, application, and documentation. *CPT Pharmacometrics Syst Pharmacol* 2016;**5**:93–122.
198. Hemming R. M&S Good Practices and Next Steps. In: *EMA-EFPIA Modelling and Simulation Workshop*, London 2011.
199. Burton A, Altman DG, Royston P, *et al*. The design of simulation studies in medical statistics. *Stat Med* 2006;**25**:4279–92.
200. Kantarjian H, Steensma D, Rius Sanjuan J, *et al*. High cancer drug prices in the United States: reasons and proposed solutions. *J Oncol Pract* 2014;**10**:e208–11.
201. Huang SM, Abernethy DR, Wang Y, *et al*. The utility of modeling and simulation in drug development and regulatory review. *J Pharm Sci* 2013;**102**:2912–23.
202. FDA. Challenge and opportunity on the critical path to new medical products, 2004.
203. FDA. *Advancing Regulatory Science at FDA*. Rockville, MD: US Department of Health and Human Services, 2011.
204. EMA. Modelling and Simulation Working Group. http://www.ema.europa.eu/ema/index.jsp?curl=pages/contacts/PDCO/people_listing_000123.jsp&mid=WC0b01ac058063f485.
205. Bhattaram VA, Booth BP, Ramchandani RP, *et al*. Impact of pharmacometrics on drug approval and labeling decisions: a survey of 42 new drug applications. *AAPS J* 2005; **7**:E503–12.
206. Champiat S, Dercle L, Ammari S, *et al*. Hyperprogressive disease is a new pattern of progression in cancer patients treated by anti-PD-1/PD-L1. *Clin Cancer Res* 2017;**23**:1920–28.
207. Saâda-Bouzid E, Defaucheux C, Karabajakian A, *et al*. Hyperprogression during anti-PD-1/PD-L1 therapy in patients with recurrent and/or metastatic head and neck squamous cell carcinoma. *Ann Oncol* 2017, in press.
208. Servant N, Romejon J, Gestraud P, *et al*. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. *Front Genet* 2014;**5**:152.
209. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* 2011;**8**:184–7.

III.1.3. Discussion

La production des données de santé dans le monde est en pleine expansion avec environ 80 milliards d'objets connectés, 50.000 applications de santé et 2,3 milliards de giga octets de données d'ici 2020. Ces données de grande dimension (big data) incluant les données omiques et leur agrégation grâce à l'amélioration des moyens informatiques ainsi que la mise en commun de bases internationales permettent aujourd'hui d'offrir un potentiel d'exploitation important. Ainsi, avec la puissance de calcul des ordinateurs qui ne cesse de croître et le nombre de données de plus en plus important, l'IA et les ETS pourraient devenir des techniques essentielles d'aide au développement de nouveaux médicaments. Cela permettrait de réduire le coût, la durée et le nombre d'essais thérapeutiques. Comme nous avons pu le constater, à travers notre revue systématique de la littérature, les ETS sont principalement basés sur des modèles intégrant un grand nombre de variables biologiques et cliniques. Toute cette complexité physiopathologique sera alors traduite par un modèle mathématique. C'est à ce moment-là que des patients simulés sont générés à partir de données recueillies lors d'essais thérapeutiques. Ces patients sont alors inclus dans le modèle afin de mieux prédire l'effet du médicament sur l'organisme. Le terme de patient simulé n'est pas tout à fait exact car ils ne sont définis que par un nombre «limité» de covariables cliniques (âge, sexe, Indice de masse corporelle…) ou biologiques (clairance du médicament, déficit enzymatique…) [145] qui n'expliquent pas toutes les interactions complexes qui existent entre la biologie de l'organisme, la maladie et le traitement [146]. Par exemple, la variable « âge » d'un patient peut être simulée à partir d'une loi normale et la variable « sexe » à partir d'une loi binomiale. Mais, lorsque de nombreuses covariables physiopathologiques sont introduites, les corrélations entre elles sont importantes. Généralement, le poids et la taille sont liés au sexe et la fonction rénale à l'âge. Le problème devient encore plus complexe lorsque ces covariables varient au cours du temps, comme par exemple l'expression tumorale du récepteur du facteur de croissance épidermique dans le cancer colorectal [147], qui nécessite un modèle longitudinal pour décrire ces changements avec davantage de corrélations. Comme nous pouvons le constater, la complexité des modèles dépend non seulement de la quantité de données disponibles [148] mais aussi du niveau de connaissance des phénomènes biologiques. Une compréhension insuffisante de certains de ces phénomènes entraîne qu'ils sont tout simplement ignorés des modèles alors qu'ils pourraient avoir une incidence sur les résultats ce qui rend actuellement la simulation de patients simulés ou la création de jumeaux numériques de patients réels difficilement réalisables. De plus, il est important que les modèles mathématiques soient couplés aux profils génétiques des patients identifiés à l'aide de méthodes d'IA, ceci afin de prendre en compte les facteurs liés à la résistance et/ou à la réponse vis-à-vis d'une molécule [149-152].

A ce jour de nombreux médicaments anti-cancéreux sont développées sur des cibles identiques telles que les anti-REGF, les anti-angiogéniques visant le VEGFR2, et plus récemment les inhibiteurs de CDK 4-6 et les inhibiteurs de PARP. On se trouve face à des «me-too» [153], médicaments de structure très proche, souvent des inhibiteurs de tyrosine-kinase. L'IA devrait pouvoir offrir des solutions d'optimisation à ce niveau. Ainsi grâce à de l'information dense et multiples issues du développement des premiers «me-too» sur une classe, on devrait pouvoir prévoir la pharmacodynamique (toxicité et efficacité) d'un «me-too» suivant (en rétrospectif et en prospectif). En rentrant ainsi dans le paradigme des ETS, le développement des médicaments nouveaux dans une classe donnée se ferait avec moins de patients, sur une durée réduite et un moindre coût [154]. Cette solution offerte par les ETS répond à un véritable besoin actuel compte tenu des cibles nouvelles à explorer et valider cliniquement et le nombre relativement réduit de patients disponibles pour des essais thérapeutiques développés dans un cadre traditionnel. Pour toutes ces raisons, il est essentiel que les acteurs de la recherche clinique, qu'ils soient industriels ou académiques, examinent de près cette possibilité méthodologie. À cette fin, il est indispensable que des exemples d'ETS fournissant une preuve de concept soient réalisés et publiés, malgré certaines contraintes de confidentialité des données, afin d'en examiner les avantages et les inconvénients. La mise en œuvre des ETS nécessitera non seulement des outils bioinformatiques facilitant l'interconnexion et l'intégration globale de toutes les données individuelles, mais également un cadre juridique global protégeant la vie privée de chaque patient.

[1]médicaments dont la structure est très similaire aux médicaments déjà connus, avec seulement des différences mineures, tout en n'améliorant pas la qualité de vie des patients

III.2. Conclusion générale et développement ultérieur

III.2.1. Conclusion générale

Ce projet de thèse se situait dans la thématique de l'apport des méthodes d'IA sur l'analyse et la modélisation de données omiques en oncologie. Ces méthodes recueillent de plus en plus d'intérêt dans la communauté scientifique car elles apportent des réponses adaptées à de nombreux problèmes qui deviennent de plus en plus complexes. Comme on a pu le constater tout au long de ce travail, même avec un nombre limité de patients et un grand nombre de variables, les méthodes d'apprentissage supervisé et non supervisé ont non seulement montré un certain niveau de performance statistique mais aussi que les résultats obtenus étaient pertinents à la fois d'un point de vue clinique et biologique.

Le premier axe de recherche de cette thèse a tout d'abord été de prédire à l'aide d'une méthode d'apprentissage supervisé l'impact du traitement par inhibiteur de point de contrôle immunitaire (IPCI) sur le plan de l'efficacité thérapeutique mais aussi de la survenue d'effets indésirables sur 94 patients traités en monothérapie par IPCI. Cette analyse, réalisée sur 163 SNPs, a permis de mettre en évidence que premièrement, le taux de réponse objective était associé à sept SNPs liés au microenvironnement immunitaire de la tumeur et que deuxièmement, la toxicité était associée à cinq SNPs liés aux cibles thérapeutiques. L'ensemble des résultats obtenus est cohérent puisque parmi les SNPs sélectionnés nous avons retrouvé des SNPs qui ont déjà été identifiés dans la littérature scientifique. En effet, il a été montré par exemple que le SNP rs13900 (*CCL2*) a un impact sur l'efficacité de l'immunothérapie [155, 156] ou que le SNP rs413815 (*PD-L1*) est associé aux cibles thérapeutiques anti PD-L1 [157, 158]. A l'aide d'une régression pénalisée Elastic net, les deux modèles (efficacité et toxicité) ont montré une performance prédictive (AUC) pouvant être jugée comme très satisfaisante. Deux scores génomiques prédictifs s'appuyant sur ces SNPs ont été développés et ont permis d'identifier trois niveaux de risque d'échec au traitement (faible, modéré, élevé) ainsi que deux niveaux de risque de survenue d'effets indésirables (faible, élevé). En raison du caractère rétrospectif de l'étude et du nombre limité de patients, il serait intéressant de valider les SNPs sélectionnés sur une cohorte plus importante et dans le contexte d'un essai clinique prospectif, afin de s'assurer de la reproductibilité et de la fiabilité des résultats à plus grande échelle. C'est dans cette optique que notre score génomique prédictif sera testé dans le cadre de l'essai d'Unicancer TOPNIVO.

Le second axe de cette thèse, a été de comparer cinq méthodes d'apprentissage non supervisé appliquées à des données de métabolomique chez 52 patientes atteintes d'un cancer du sein et traitées par chimiothérapie adjuvante. Les résultats ont montré qu'il était possible à partir de 449 métabolites, d'identifier trois groupes de patientes qui se sont révélés correspondre à des profils de pronostic bon, intermédiaire ou mauvais, définis selon les facteurs cliniques et biologiques reconnus.

Avec un indice de silhouette de 0,91 et de 0,85 respectivement, les méthodes K-sparse et SIMLR ont été les méthodes les plus performantes en termes de clustering. Une analyse clinique de ces trois clusters a permis de confirmer les performances des méthodes K-sparse et SIMLR. En effet, ces dernières étaient parmi les méthodes les plus discriminantes avec six variables cliniques présentant des différences statistiquement significatives. Parmi ces variables, nous avons mis en évidence une différence en termes de grade histologique, de phénotype tumoral, d'index de prolifération Ki67 et de taille tumorale. Les résultats ont aussi montré une hétérogénéité au sein des tumeurs triple-négatives où un tiers de ces tumeurs ont été classées dans le groupe bon pronostic, ce qui est concordant avec les données de la littérature [138, 139]. Notre approche a permis de mettre en évidence qu'il existe une dysrégulation de trois voies métaboliques (à savoir la glycolyse et la production de lactate, la glutaminolyse et les acides aminés) qui jouent un rôle central dans la croissance tumorale [159, 160]. Les cinq méthodes ont montré que les patientes du groupe de mauvais pronostic présentaient des taux plus élevés de guanine, de L-isoleucine, de L-méthionine et de spermine. Ces résultats suggèrent que ces métabolites pourraient être des biomarqueurs candidats prédictifs d'un moins bon pronostic [161]. Les résultats obtenus pendant ce travail constituent seulement une petite partie de ce que la métabolomique peut nous apporter. Avec la génomique et la protéomique, la métabolomique complète la trilogie des études omiques. La métabolomique nous apportera de nouvelles connaissances et sera certainement utilisée pour de nombreuses applications en santé.

Enfin, le troisième axe de cette thèse a été d'évaluer l'apport des ETS dans le développement des essais thérapeutiques simulés à travers une revue systématique de la littérature. Les résultats ont mis en évidence toutes les difficultés rencontrées lors de l'élaboration d'un ETS. C'est certainement une des raisons expliquant qu'à la date de réalisation de cette revue systématique, seulement 10 ETS en oncologie avaient été publiés. A la vue de nos résultats, il est apparu clairement que les méthodes d'IA appliquées sur des données omiques constituaient une piste intéressante pour l'analyse de ces données. Ce constat est confirmé par nos résultats lors de l'analyse des études « IMMUNOCARTA » et « EMMEEA ». De toute évidence, ces essais *in silico* devront toujours être validés dans le cadre d'un essai thérapeutique prospectif et devront faire partie intégrante des futurs programmes de développement des médicaments.

A la vue de ces éléments et afin de permettre le développement de nouveaux outils pronostiques, il semblerait que la combinaison de différentes approches, tant d'un point de vue mathématique que biologique, semble être une perspective prometteuse dans la recherche de biomarqueurs [162-165], nécessitant une collaboration étroite entre cliniciens, biologistes et biostatisticiens [166].

III.2.2. Développement ultérieur

Avec plus de 300 essais cliniques enregistrés sur le site ClinicalTrials.gov sous le titre «intelligence artificielle», «apprentissage automatique» ou «apprentissage en profondeur» et pas moins de 30 algorithmes d'IA, approuvés par la Food and Drug Administration américaine [167], la validation des techniques d'IA devient une nécessité. La validation externe des modèles reste à ce jour le meilleur moyen pour juger de la pertinence d'un nouvel algorithme d'IA. Un manque de transparence et d'informations détaillées sur les modèles peut entrainer des difficultés à juger de la validité et de la reproductibilité des résultats obtenus, pouvant donner lieu à des fausses interprétations et ainsi surévaluer l'efficacité d'une méthode d'IA. Prenons l'exemple de la signature du cancer du sein publié en 2000 par Perou et Sorlie [43] qui a donné lieu à la première classification moléculaire. Depuis cette date, même si de nombreuses autres signatures génétiques pronostiques ont fait l'objet d'études approfondies, seules quelques-unes (OncotypeDX®, Prosigna®, MammaPrint®, Endopredict® Genomic grade index® and BC Index®[168]) sont utilisées actuellement en routine clinique [169, 170]. De plus, bien que cette étape de validation soit essentielle, il est nécessaire de démontrer que cela puisse apporter un gain pour le patient, ceci ne pouvant se faire que dans le cadre d'un essai clinique prospectif, où l'apport de l'IA pourrait être alors cliniquement évalué. C'est pour relever ces défis que les groupes CONSORT-IA et SPIRIT-IA préparent des extensions internationales des déclarations CONSORT [171] et SPIRIT [172]. Ces nouvelles recommandations, utilisant le cadre méthodologique du réseau EQUATOR [173], porteront spécifiquement sur les essais cliniques dans lesquels une composante d'IA sera introduite [174].

Dans le prolongement des travaux menés durant cette thèse et dans un objectif de validation de nos résultats, plusieurs extensions sont envisagées :

Tout d'abord concernant le projet IMMUNOCARTA : 1-Une analyse comparative entre les résultats obtenus avec la méthode Elastic net et ceux obtenus en utilisant les méthodes sparse group Lasso, sparse group PLS et sparse group Elastic net devrait être très prochainement réalisée et fera l'objet d'une publication. 2- Dans le prolongement de ce premier travail, nous appliquerons ces méthodes sur une série de 250 patients porteurs d'un carcinome épidermoïde de la tête et du cou, en rechute et/ou métastatiques, et réfractaires aux sels de platine dans le but de confirmer nos résultats obtenus (Essai Unicancer TOPNIVO). L'objectif de ce travail est double, premièrement valider nos résultats biologiques obtenus sur le panel de 163 SNPS et deuxièmement, mieux évaluer les performances des méthodes utilisées.

Ensuite, pour ce qui concerne le projet EMMEEA, une analyse de validation de nos premiers résultats devrait débuter début 2020. Cette validation s'effectuera sur une cohorte indépendante de 39 patientes traitées par chimiothérapie néoadjuvante. Cette analyse présentera plusieurs avantages :

1- Suivi des patientes plus important (suivi médian : 70 mois); 2-Données de métabolomique avant traitement et après traitement; 3- Données de TEP associées permettant de combiner une analyse métabolomique et radiomique. En effet, avec l'introduction de la radiomique, l'information portée par l'image pourrait être complémentaire de celles provenant des sources cliniques ou biologiques permettant ainsi d'enrichir le nombre des caractéristiques de la tumeur [175, 176]. La première est que les caractéristiques tumorales cliniques, à l'échelle tissulaire, cellulaire et/ou génomique auraient un retentissement phénotypique en imagerie médicale. Cela revient à considérer que des caractéristiques de l'image sont fortement corrélées à des caractéristiques cliniques et biologiques. Le deuxième rationnel est que l'information portée par l'image serait complémentaire de celles provenant d'autres sources d'informations médicales permettant ainsi d'enrichir le nombre des caractéristiques de la tumeur.

De même, nous mènerons une seconde analyse dont l'objectif sera cette fois-ci de comparer les performances de six méthodes d'apprentissage supervisé (ANN, SVM, RF, KNN, GBM et SK-Sparse) afin de prédire le statut HER2 chez 52 patientes traitées par chimiothérapie adjuvante à l'aide de données de métabolomique. Nous comparerons les résultats ainsi obtenus avec ceux obtenus avec un jeu de données simulées comme recommandé dans la littérature [177]. De plus cette comparaison nous permettra de valider les conclusions d'Alawaa *et al.* encourageant l'utilisation de méthodes d'apprentissage profond (ici ANN) pour l'analyse de données de métabolomique [17].

Les ETS peinent à être intégrés dans le processus de développement des médicaments en oncologie et sont encore mal considérés lors d'une autorisation de mise sur le marché d'un médicament [178], malgré les recommandations de la FDA [179-182] et de l'EMEA [183]. Nous avons montré, à travers notre revue systématique de la littérature tous les enjeux et bénéfices qu'ils pourraient offrir à la recherche clinique, telle qu'une réduction du nombre d'essais thérapeutiques négatifs ou simplement la diminution de leurs coûts. C'est pour cela que dans le prolongement de cette thèse, nous avons décidé d'essayer de mettre en application le concept d'ETS à l'aide des données de patients disponibles au Centre Antoine Lacassagne.

Les essais thérapeutiques simulés et l'IA sont en train de révolutionner le domaine de la santé en permettant une meilleure modélisation des mécanismes du cancer, des diagnostics plus rapides et plus efficaces et une optimisation de la prise en charge du patient. Notre travail constitue un élément de cette évolution qui devrait permettre d'améliorer, sans la remplacer, la capacité d'analyse et de prise de décision des oncologues.

IV. Annexes

Nombre de publications par catégorie et par position au 22/10/2019

Période : 2010 - 2019							
Position	**Total**	**A**	**B**	**C**	**D**	**E**	**NC**
1	1	1	0	0	0	0	0
2	13	1	3	5	0	4	0
3	22	0	6	7	3	3	3
k	18	4	4	4	3	3	0
ADA	4	1	1	0	2	0	0
DA	2	1	0	0	0	1	0
Total	**60**	**8**	**14**	**16**	**8**	**11**	**3**

Répartition par catégorie et par année au 22/10/2019

Période : 2010 - 2019								
Année	**A**	**B**	**C**	**D**	**E**	**NC**	**Total**	**Score**
2010	0	1	1	0	0	0	2	14
2011	0	1	0	0	1	0	2	12
2012	1	2	1	1	1	1	7	57
2013	0	2	1	0	2	2	7	52
2014	0	1	2	1	2	0	6	29
2015	1	0	1	1	2	0	5	26
2016	1	1	0	2	1	0	5	67
2017	2	2	1	0	1	0	6	56
2018	2	2	3	3	1	0	11	138
2019	1	2	6	0	0	0	9	94
Total	**8**	**14**	**16**	**8**	**11**	**3**	**60**	**545**

V. Bibliographie

1. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery, Lancet Oncol 2019;20:e262-e273.
2. Shaikhina T, Lowe D, Daga S et al. Machine learning for predictive modelling based on small data in biomedical engineering, IFAC-PapersOnLine 2015;48:469-474.
3. Bellman R. Dynamic programming and Lagrange multipliers, Proceedings of the National Academy of Sciences of the United States of America 1956;42:767.
4. Turing AM. Computing machinery and intelligence. Parsing the Turing Test. Springer, 2009, 23-65.
5. Moor J. The Dartmouth College artificial intelligence conference: The next fifty years, Ai Magazine 2006;27:87-87.
6. Minsky ML. Computation. Prentice-Hall Englewood Cliffs, 1967.
7. Moore GE. Cramming more components onto integrated circuits. McGraw-Hill New York, NY, USA:, 1965.
8. Moore GE. Progress in digital integrated electronics. In: Electron Devices Meeting. 1975, p. 11-13.
9. LeCun Y, Bottou L, Bengio Y et al. Gradient-based learning applied to document recognition, Proceedings of the IEEE 1998;86:2278-2324.
10. Deng J, Dong W, Socher R et al. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. 2009, p. 248-255. Ieee.
11. Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks, Nature 2017;542:115.
12. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine, N Engl J Med 2019;380:1347-1358.
13. Kourou K, Exarchos TP, Exarchos KP et al. Machine learning applications in cancer prognosis and prediction, Comput Struct Biotechnol J 2015;13:8-17.
14. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis, Cancer Inform 2007;2:59-77.
15. Schmidhuber J. Deep learning in neural networks: An overview, Neural networks 2015;61:85-117.
16. Eraslan G, Avsec Ž, Gagneur J et al. Deep learning: new computational modelling techniques for genomics, Nature Reviews Genetics 2019:1.
17. Alakwaa FM, Chaudhary K, Garmire LX. Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data, J Proteome Res 2018;17:337-347.
18. Sutton RS, Barto AG. Introduction to reinforcement learning. MIT press Cambridge, 1998.
19. Silver D, Hubert T, Schrittwieser J et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, Science 2018;362:1140-1144.
20. Abbeel P, Coates A, Quigley M et al. An application of reinforcement learning to aerobatic helicopter flight. In: Advances in neural information processing systems. 2007, p. 1-8.
21. Liu Z, Yao C, Yu H et al. Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things, Future Generation Computer Systems 2019;97:1-9.
22. Frey BJ, Dueck D. Clustering by passing messages between data points, Science 2007;315:972-976.
23. Hotelling H. Analysis of a complex of statistical variables into principal components, Journal of educational psychology 1933;24:417.
24. Pearson K. LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 1901;2:559-572.
25. Jolliffe I. Principal component analysis. Springer, 2011.

26. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Biocomputing 2000. World Scientific, 1999, 455-466.
27. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data, Bioinformatics 2001;17:763-774.
28. Adam B, Jerzy Z, Jerzy K et al. A principal component analysis of patients, disease and treatment variables: a new prognostic tool in breast cancer after mastectomy, Reports of Practical Oncology & Radiotherapy 2000;5:83-89.
29. Lenz M, Müller F-J, Zenke M et al. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data, Scientific Reports 2016;6:25696.
30. Lynch CM, van Berkel VH, Frieboes HB. Application of unsupervised analysis techniques to lung cancer patient data, PLoS One 2017;12:e0184370.
31. Nyamundanda G, Poudel P, Patil Y et al. A novel statistical method to diagnose, quantify and correct batch effects in genomic studies, Scientific Reports 2017;7:10849.
32. Ding C, He X, Zha H et al. Adaptive dimension reduction for clustering high dimensional data. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings., 2002, p. 147-154. IEEE.
33. Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review, Acm Sigkdd Explorations Newsletter 2004;6:90-105.
34. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis, The computer journal 1998;41:578-588.
35. Bridges Jr CC. Hierarchical cluster analysis, Psychological reports 1966;18:851-854.
36. Dubes RC, Jain AK. Algorithms for clustering data. Prentice hall Englewood Cliffs, 1988.
37. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009.
38. Lance GN, Williams WT. A general theory of classificatory sorting strategies: II. Clustering systems, The computer journal 1967;10:271-277.
39. Lance GN, Williams WT. A general theory of classificatory sorting strategies: 1. Hierarchical systems, The computer journal 1967;9:373-380.
40. Lambert J, Williams W. Multivariate methods in plant ecology: VI. Comparison of information-analysis and association-analysis, The Journal of Ecology 1966:635-664.
41. Ward Jr JH. Hierarchical grouping to optimize an objective function, Journal of the American statistical association 1963;58:236-244.
42. Eisen MB, Spellman PT, Brown PO et al. Cluster analysis and display of genome-wide expression patterns, Proc Natl Acad Sci U S A 1998;95:14863-14868.
43. Perou CM, Sorlie T, Eisen MB et al. Molecular portraits of human breast tumours, Nature 2000;406:747-752.
44. Sorlie T, Perou CM, Tibshirani R et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, Proc Natl Acad Sci U S A 2001;98:10869-10874.
45. Sorlie T, Tibshirani R, Parker J et al. Repeated observation of breast tumor subtypes in independent gene expression data sets, Proc Natl Acad Sci U S A 2003;100:8418-8423.
46. Lloyd S. Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM, IEEE Trans. Inform. Theor.(1957/1982) Google Scholar 1957.
47. Steinhaus H. Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV: 801-804 1956.
48. Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. 2007, p. 1027-1035. Society for Industrial and Applied Mathematics.
49. Wacquet G. Classification spectrale semi-supervisée: Application à la supervision de l'écosystème marin. 2011.

50. Shi M, Xu G. Spectral clustering using Nyström approximation for the accurate identification of cancer molecular subtypes, Scientific Reports 2017;7:4896.
51. Chin AJ, Mirzal A, Haron H. Spectral clustering on gene expression profile to identify cancer types or subtypes, Jurnal Teknologi 2015;76.
52. Jiang L, Xiao Y, Ding Y et al. Discovering cancer subtypes via an accurate fusion strategy on multiple profile data, Frontiers in genetics 2019;10:20.
53. Mercer J. Xvi. functions of positive and negative type, and their connection the theory of integral equations, Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character 1909;209:415-446.
54. Canu S, Mary X, Rakotomamonjy A. Functional learning through kernels, arXiv preprint arXiv:0910.1013 2009.
55. Cortes C, Vapnik V. Support-vector networks, Machine learning 1995;20:273-297.
56. Schölkopf B, Tsuda K, Vert J-P. Support vector machine applications in computational biology. MIT press, 2004.
57. Gönen M, Alpaydın E. Multiple kernel learning algorithms, Journal of machine learning research 2011;12:2211-2268.
58. Bach FR, Lanckriet GR, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the twenty-first international conference on Machine learning. 2004, p. 6. ACM.
59. Wang B, Zhu J, Pierson E et al. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning, Nat Methods 2017;14:414-416.
60. Tibshirani R. Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 1996;58:267-288.
61. Witten DM, Tibshirani R. A framework for feature selection in clustering, Journal of the American statistical association 2010;105:713-726.
62. Gilet C, Deprez M, Caillau J-B et al. Clustering with sparse feature selection using alternating minimization and an exact projection-gradient splitting method.
63. Condat L. Fast projection onto the simplex and the $$\pmb {l} _\mathbf {1} $$ ball, Mathematical Programming 2016;158:575-585.
64. Fu K. IEEE Transactions on Pattern Analysis and Machine Intelligence, Encyclopedia of Statistical Sciences 2004.
65. Caliński T, Harabasz J. A dendrite method for cluster analysis, Communications in Statistics-theory and Methods 1974;3:1-27.
66. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 1987;20:53-65.
67. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2001;63:411-423.
68. Hubert L, Arabie P. Comparing partitions, Journal of classification 1985;2:193-218.
69. Strehl A, Ghosh J. Cluster ensembles---a knowledge reuse framework for combining multiple partitions, Journal of machine learning research 2002;3:583-617.
70. Kawakami E, Tabata J, Yanaihara N et al. Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers, Clin Cancer Res 2019;25:3006-3015.
71. Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data, Artificial intelligence in medicine 2018;90:1-14.
72. Hastie T, Tibshirani R, Friedman J et al. The elements of statistical learning: data mining, inference and prediction, The Mathematical Intelligencer 2005;27:83-85.
73. Tenenhaus M. La régression PLS: théorie et pratique. Editions technip, 1998.
74. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 1970;12:55-67.
75. Zou H, Hastie T. Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2005;67:301-320.

76. Das J, Gayvert KM, Bunea F et al. ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers, BMC Genomics 2015;16:263.
77. Gelman A, Park DK. Splitting a predictor at the upper quarter or third and the lower quarter or third, The American Statistician 2009;63:1-8.
78. Nakas CT, Dalrymple-Alford JC, Anderson TJ et al. Generalization of Youden index for multiple-class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening, Statistics in medicine 2013;32:995-1003.
79. Nelder JA, Mead R. A simplex method for function minimization, The computer journal 1965;7:308-313.
80. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods, Bioinformatics 2005;21:3301-3307.
81. Stone M. Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society: Series B (Methodological) 1974;36:111-133.
82. Lachenbruch PA, Mickey MR. Estimation of error rates in discriminant analysis, Technometrics 1968;10:1-11.
83. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models, Global ecology and Biogeography 2008;17:145-151.
84. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 1982;143:29-36.
85. Gal J, Milano G, Ferrero JM et al. Optimizing drug development in oncology by clinical trial simulation: Why and how?, Brief Bioinform 2018;19:1203-1217.
86. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine, Nat Rev Clin Oncol 2011;8:184-187.
87. Genomes Project C, Abecasis GR, Altshuler D et al. A map of human genome variation from population-scale sequencing, Nature 2010;467:1061-1073.
88. International HapMap C, Frazer KA, Ballinger DG et al. A second generation human haplotype map of over 3.1 million SNPs, Nature 2007;449:851-861.
89. International HapMap C. A haplotype map of the human genome, Nature 2005;437:1299-1320.
90. Olivier M. A haplotype map of the human genome, Physiol Genomics 2003;13:3-9.
91. Hardy GH. Mendelian proportions in a mixed population, Classic papers in genetics. Prentice-Hall, Inc.: Englewood Cliffs, NJ 1908:60-62.
92. Weinberg W. ber den Nachweis der Vererbung beim Menschen, Jahres. Wiertt. Ver. Vaterl. Natkd. 1908;64:369-382.
93. Hill WG, Mackay TF. DS Falconer and Introduction to quantitative genetics, Genetics 2004;167:1529-1536.
94. Hill WG, Robertson A. Linkage disequilibrium in finite populations, Theor Appl Genet 1968;38:226-231.
95. Beaton D, Filbey F, Abdi H. Integrating partial least squares correlation and correspondence analysis for nominal data. New perspectives in partial least squares and related methods. Springer, 2013, 81-94.
96. Fiehn O. Metabolomics—the link between genotypes and phenotypes. Functional genomics. Springer, 2002, 155-171.
97. Trivedi MS, Holger D, Bui AT et al. Short-term sleep deprivation leads to decreased systemic redox metabolites and altered epigenetic status, PLoS One 2017;12:e0181978.
98. Pan Z, Raftery D. Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics, Analytical and bioanalytical chemistry 2007;387:525-527.
99. Riekeberg E, Powers R. New frontiers in metabolomics: from measurement to insight, F1000Research 2017;6.
100. Hejazi L, Ebrahimi D, Hibbert DB et al. Compatibility of electron ionization and soft ionization methods in gas chromatography/orthogonal time-of-flight mass spectrometry, Rapid Communications in Mass Spectrometry 2009;23:2181-2189.

101. Annesley TM. Ion suppression in mass spectrometry, Clin Chem 2003;49:1041-1044.
102. Antignac JP, Brosseaud A, Gaudin-Hirret I et al. Analytical strategies for the direct mass spectrometric analysis of steroid and corticosteroid phase II metabolites, Steroids 2005;70:205-216.
103. Dunn WB, Ellis DI. Metabolomics: current analytical platforms and methodologies, TrAC Trends in Analytical Chemistry 2005;24:285-294.
104. Wishart DS. Advances in metabolite identification, Bioanalysis 2011;3:1769-1782.
105. Sugimoto M, Kawakami M, Robert M et al. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis, Current bioinformatics 2012;7:96-108.
106. Bino RJ, Hall RD, Fiehn O et al. Potential of metabolomics as a functional genomics tool, Trends in plant science 2004;9:418-425.
107. Fiehn O, Sumner LW, Rhee SY et al. Minimum reporting standards for plant biology context information in metabolomic studies, Metabolomics 2007;3:195-201.
108. Goodacre R, Broadhurst D, Smilde AK et al. Proposed minimum reporting standards for data analysis in metabolomics, Metabolomics 2007;3:231-241.
109. Considine EC, Salek RM. A Tool to Encourage Minimum Reporting Guideline Uptake for Data Analysis in Metabolomics, Metabolites 2019;9:43.
110. Spicer RA, Salek R, Steinbeck C. Compliance with minimum information guidelines in public metabolomics repositories, Scientific data 2017;4:170137.
111. Gorrochategui E, Jaumot J, Lacorte S et al. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow, TrAC Trends in Analytical Chemistry 2016;82:425-442.
112. Wu Y, Li L. Sample normalization methods in quantitative metabolomics, J Chromatogr A 2016;1430:80-95.
113. Manzoni C, Kia DA, Vandrovcova J et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences, Brief Bioinform 2018;19:286-302.
114. Rosato A, Tenori L, Cascante M et al. From correlation to causation: analysis of metabolomics data using systems biology approaches, Metabolomics 2018;14:37.
115. Xia J, Wishart DS. Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis, Curr Protoc Bioinformatics 2016;55:14 10 11-14 10 91.
116. Krämer A, Green J, Pollard Jr J et al. Causal analysis approaches in ingenuity pathway analysis, Bioinformatics 2013;30:523-530.
117. Refae S, Gal J, Ebran N et al. Germinal Immunogenetics predict treatment outcome for PD-1/PD-L1 checkpoint inhibitors, Invest New Drugs 2019.
118. Gal J, Bailleux C, Chardin D et al. Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer, Briefings Bioinformatics 2019.
119. Paule I, Tod M, Henin E et al. Dose adaptation of capecitabine based on individual prediction of limiting toxicity grade: evaluation by clinical trial simulation, Cancer Chemother Pharmacol 2012;69:447-455.
120. Refae S, Gal J, Brest P et al. Germinal immunogenetics as a predictive factor for immunotherapy, Critical reviews in oncology/hematology 2019.
121. Refae S, Gal J, Ebran N et al. Predicting checkpoint inhibitor treatment outcome in head and neck cancer patients: a potential role for host immunogenetics, Head and Neck 2019;Under review.
122. Refae S, Gal J, Brest P et al. Hyperprogression under Immune Checkpoint Inhibitor: a potential role for germinal immunogenetics, Scientific Reports 2019.
123. Bazhin A, von Ahn K, Fritz J et al. Interferon-α up-regulates the expression of PD-L1 molecules on immune cells through STAT3 and p38 signaling, Frontiers in immunology 2018;9:2129.
124. Clarke R, Ressom HW, Wang A et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nat Rev Cancer 2008;8:37-49.
125. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data, BMC Bioinformatics 2013;14:106.

126. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2008;70:849-911.
127. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice, Statistics in medicine 2011;30:377-399.
128. Lee S, Kerns S, Ostrer H et al. Machine Learning on a Genome-wide Association Study to Predict Late Genitourinary Toxicity After Prostate Radiation Therapy, Int J Radiat Oncol Biol Phys 2018;101:128-135.
129. Lima L, Oliveira D, Ferreira JA et al. The role of functional polymorphisms in immune response genes as biomarkers of bacille Calmette-Guerin (BCG) immunotherapy outcome in bladder cancer: establishment of a predictive profile in a Southern Europe population, BJU Int 2015;116:753-763.
130. de Maturana EL, Ye Y, Calle ML et al. Application of multi-SNP approaches Bayesian LASSO and AUC-RF to detect main effects of inflammatory-gene variants associated with bladder cancer risk, PLoS One 2013;8:e83745.
131. Simon N, Friedman J, Hastie T et al. A sparse-group lasso, Journal of Computational and Graphical Statistics 2013;22:231-245.
132. Liquet B, de Micheaux PL, Hejblum BP et al. Group and sparse group partial least square approaches applied in genomics context, Bioinformatics 2015;32:35-42.
133. Samarov DV, Allen D, Hwang J et al. A Coordinate-Descent-Based Approach to Solving the Sparse Group Elastic Net, Technometrics 2017;59:437-445.
134. Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. In: Advances in neural information processing systems. 2007, p. 41-48.
135. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2008;70:53-71.
136. Obozinski G, Taskar B, Jordan M. Joint covariate selection for grouped classification. Technical Report 743, Dept. of Statistics, University of California Berkeley, 2007.
137. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2006;68:49-67.
138. Bianchini G, Balko JM, Mayer IA et al. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease, Nat Rev Clin Oncol 2016;13:674-690.
139. Mills MN, Yang GQ, Oliver DE et al. Histologic heterogeneity of triple negative breast cancer: A National Cancer Centre Database analysis, Eur J Cancer 2018;98:48-58.
140. Keun HC, Ebbels TM, Bollard ME et al. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles, Chemical research in toxicology 2004;17:579-587.
141. Elia I, Rossi M, Stegen S et al. Breast cancer cells rely on environmental pyruvate to shape the metastatic niche, Nature 2019;568:117-121.
142. Fernandez MF, Reina-Perez I, Astorga JM et al. Breast Cancer and Its Relationship with the Microbiota, Int J Environ Res Public Health 2018;15.
143. Eckhart AD, Beebe K, Milburn M. Metabolomics as a key integrator for “omic” advancement of personalized medicine and future therapies, Clinical and translational science 2012;5:285-288.
144. Guyon I, Von Luxburg U, Williamson RC. Clustering: Science or art. In: NIPS 2009 workshop on clustering theory. 2009, p. 1-11.
145. Alkema W, Rullmann T, van Elsas A. Target validation in silico: does the virtual patient cure the pharma pipeline?, Expert Opin Ther Targets 2006;10:635-638.
146. Bangs A. Predictive biosimulation and virtual patients in pharmaceutical R and D, Stud Health Technol Inform 2005;111:37-42.
147. Siravegna G, Mussolin B, Buscarino M et al. Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients, Nat Med 2015;21:827.
148. Benzekry S, Tracz A, Mastri M et al. Modeling Spontaneous Metastasis following Surgery: An In Vivo-In Silico Approach, Cancer Res 2016;76:535-547.
149. Sameen S, Barbuti R, Milazzo P et al. Mathematical modeling of drug resistance due to KRAS mutation in colorectal cancer, J Theor Biol 2016;389:263-273.

150. Sun X, Bao J, Shao Y. Mathematical Modeling of Therapy-induced Cancer Drug Resistance: Connecting Cancer Mechanisms to Population Survival Rates, Sci Rep 2016;6:22498.
151. O'Donnell JS, Long GV, Scolyer RA et al. Resistance to PD1/PDL1 checkpoint inhibition, Cancer Treat Rev 2017;52:71-81.
152. Martinelli E, Morgillo F, Troiani T et al. Cancer resistance to therapies against the EGFR-RAS-RAF pathway: The role of MEK, Cancer Treat Rev 2017;53:61-69.
153. Fojo T, Mailankody S, Lo A. Unintended consequences of expensive cancer therapeutics-the pursuit of marginal indications and a me-too mentality that stifles innovation and creativity: the John Conley Lecture, JAMA Otolaryngol Head Neck Surg 2014;140:1225-1236.
154. Viceconti M, Henney A, Morley-Fletcher E. In silico clinical trials: how computer simulation will transform the biomedical industry. Research and technological development roadmap. Brussels: Avicenna Consortium, 2016.
155. Yao M, Brummer G, Acevedo D et al. Cytokine Regulation of Metastasis and Tumorigenicity, Adv Cancer Res 2016;132:265-367.
156. Fridlender ZG, Buchlis G, Kapoor V et al. CCL2 blockade augments cancer immunotherapy, Cancer Res 2010;70:109-118.
157. Nomizo T, Ozasa H, Tsuji T et al. Clinical Impact of Single Nucleotide Polymorphism in PD-L1 on Response to Nivolumab for Advanced Non-Small-Cell Lung Cancer Patients, Sci Rep 2017;7:45124.
158. Munn DH. The host protecting the tumor from the host - targeting PDL1 expressed by host cells, J Clin Invest 2018;128:570-572.
159. Haukaas TH, Euceda LR, Giskeodegard GF et al. Metabolic Portraits of Breast Cancer by HR MAS MR Spectroscopy of Intact Tissue Samples, Metabolites 2017;7.
160. DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism, Sci Adv 2016;2:e1600200.
161. Melone MAB, Valentino A, Margarucci S et al. The carnitine system and cancer metabolic plasticity, Cell Death Dis 2018;9:228.
162. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark, Nucleic Acids Res 2019;47:1044.
163. Mitra S, Saha S. A multiobjective multi-view cluster ensemble technique: Application in patient subclassification, PLoS One 2019;14:e0216904.
164. Wu C, Zhou F, Ren J et al. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection, High Throughput 2019;8.
165. Ramazzotti D, Lal A, Wang B et al. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival, Nat Commun 2018;9:4453.
166. Savage N. Collaboration is the key to cancer research, Nature 2018;556:S1-S1.
167. Administration UFaD. Artificial Intelligence and Machine Learning in Software as a Medical Device. Artificial Intelligence and Machine Learning in Software as a Medical Device.
168. Wesolowski R, Ramaswamy B. Gene expression profiling: changing face of breast cancer classification and management, Gene Expr 2011;15:105-115.
169. Chibon F. Cancer gene expression signatures–the rise and fall?, European journal of cancer 2013;49:2000-2009.
170. Michiels S, Ternes N, Rotolo F. Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice, Ann Oncol 2016;27:2160-2167.
171. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials, BMC medicine 2010;8:18.
172. Chan A-W, Tetzlaff JM, Altman DG et al. SPIRIT 2013: new guidance for content of clinical trial protocols, Lancet 2013;381.
173. Network E. Reporting guidelines under development (EQUATOR Network, accessed 4 August 2019). http://www.equator-network.org/library/reporting-guidelines-underdevelopment/.
174. Liu X, Rivera SC, Faes L et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nature Publishing Group 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA, 2019.

175. Miranda Magalhaes Santos JM, Clemente Oliveira B, Araujo-Filho JAB et al. State-of-the-art in radiomics of hepatocellular carcinoma: a review of basic principles, applications, and limitations, Abdom Radiol (NY) 2019.
176. Yip SS, Aerts HJ. Applications and limitations of radiomics, Phys Med Biol 2016;61:R150-166.
177. Boulesteix AL, Binder H, Abrahamowicz M et al. On the necessity and design of studies comparing statistical methods, Biometrical Journal 2018;60:216-218.
178. Huang SM, Abernethy DR, Wang Y et al. The utility of modeling and simulation in drug development and regulatory review, J Pharm Sci 2013;102:2912-2923.
179. Williams PJ, Ette EI. The role of population pharmacokinetics in drug development in light of the Food and Drug Administration's 'Guidance for Industry: population pharmacokinetics', Clin Pharmacokinet 2000;39:385-395.
180. Guidance for industry on Population Pharmacokinetics; availability. Food and Drug Administration, HHS. Notice, Fed Regist 1999;64:6663-6664.
181. Food U, Administration D. Challenge and opportunity on the critical path to new medical products. 2004, March, p11 2008;24.
182. Food U, Administration D. Advancing regulatory science at FDA, US Dept. of Health and Human Services: Rockville, MD 2011.
183. EMA E. Modelling and Simulation Working Group; Available from URL: http://www.ema.europa.eu/ema/index.jsp?curl=pages/contacts/PDCO/people_listing_000123.jsp&mid=WC0b01ac058063f485.
http://www.ema.europa.eu/ema/index.jsp?curl=pages/contacts/PDCO/people_listing_000123.jsp&mid=WC0b01ac058063f485.