FACULTÉ DE MÉDECINE ET DE PHARMACIE
UNIVERSITÉ HASSAN II DE CASABLANCA

UNIVERSITÀ DEGLI STUDI DI TORINO

Doctoral thesis in co-supervision

To obtain PhD diploma

Specialty: "Complex Systems for Life Sciences"

And

"Génétique Et Pathologie Moléculaire"

between:

Department of Clinical and Biological Sciences - Doctoral School of the Università Degli Studi Di Torino

And:

Département de Génétique et Pathologie Moléculaire- Centre des études Doctorales - Faculté de Médecine et de Pharmacie - Université Hassan II de Casablanca

Presented and publicly defended by:

Laila AKHOUAYRI

Genomic, bioinformatic and statistical approaches to refine invasive breast carcinoma classification

Date and place of the defense, January 25, 2023 at: 14h00
At the Faculty of Medicine and Pharmacy of Casablanca
Jamila Hachim Amphitheater

In front of the Jury:

Pr. Mehdi KARKOURI	Faculté De Médecine Et Pharmacie De Casablanca, MOROCCO	**Supervisor**
Dr. Giovanna CHIORINO	Cancer Genomics Lab. Fondazione Edo Ed Elvo Tempia, Biella, ITALY	**Co-supervisor**
Pr. Michele DE BORTOLI	Università Degli Studi Di Torino, Torino, ITALY	**Examiner**
Pr. Hind DEHBI	Faculté De Médecine Et Pharmacie De Casablanca, MOROCCO	**Reviewer**
Pr. Nadia BENCHEKROUN	Faculté De Médecine Et Pharmacie De Casablanca, MOROCCO	**Reviewer**
Pr. Basma EL KHANNOUSSI	Institut National d'Oncologie Sidi Mohammed Ben Abdallah, Rabat, MOROCCO	**Reviewer**

Dissertation Acknowledgements:

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people. Throughout the course of this thesis and the writing of this dissertation, I have received a great deal of assistance. Therefore, I would like to pay off a debt of gratitude to anyone who intervened to facilitate and allowed me to conduct it in these difficult times.

Mainly, I would like to warmly thank my PhD supervisor, Pr. Mehdi KARKOURI, head of the pathology laboratory of CHU Casablanca, who helped me a lot throughout my PhD training. You were of invaluable help in the most delicate moments. Your expertise and tutelage were kindly supportive in formulating the research questions and methodology. Also, your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I also want to thank you for your support and for all of the opportunities I was given to further my research in Switzerland, France and Italy.

I would like to acknowledge the staff from the Pathology Department and Medicine College of Casablanca for their wonderful collaboration, particularly by singling out Pr. Hind DEHBI, Dr. Meriem REGRAGUI and Pr. Myriam RIYAD.

Likewise, I would like to thank my PhD supervisor Dr. Giovanna CHIORINO, head of Cancer Genomics Laboratory, and all the staff in Fondazione Edo ed Elvo Tempia per la lotta contro i tumori Onlus; Fondazione Policlinico Universitario A. Gemelli IRCCS in Rome and “Degli Infermi” Hospital, especially Dr Paola Ostano; Maurizia Mello-Grand; Ilaria Gregnanin; Francesca

Crivelli; Sara Laurora; Daniele Liscia; Francesco Leone; Angela Santoro; Antonino Mulè; Donatella Guarino; Claudia Maggiore; Angela Carlino; Stefano Magno; Maria Scatolini; Alba Di Leone and Riccardo Masetti. They were amply cordial, welcoming and benevolent, for sharing their expertise on a daily basis.

Thanks also to their confidence I was able to fulfill myself completely in my missions. Likewise, I thank them for sharing their professional knowledge and providing me with luminous advice and relevant remarks thus attesting to their benevolent availability.

I greatly appreciate the support received through the collaborative work undertaken within the Università Degli Studi di Torino during the last phase of my PhD course.

Thank you, Dr Giovanna, for trusting me the first day we met in the summer school organized in Al Akhawayn University and for having integrated me in your team.
Thanks to you I was able to make pleasant acquaintances and above all; I'm grateful to you for making this joint supervision possible and concretized. In addition, I would like to thank you for providing me with the tools that I needed, to choose the right direction and successfully complete my dissertation; this was really influential in shaping my experiment methods and critiquing my results.

Additionally, I would also like to thank my co-supervisor Pr. Michele DE BORTOLI, to whom I am very beholden for helping me overcome all the difficulties related to my stay and internship in Italy.
He also saved me a lot of difficulties and complications in terms of my registration.
His almost paternal support has always strengthened me, as well as that of Dr. Giovanna CHIORINO, who have constantly watched over the smooth running of this project on an international scale

Furthermore, I take this opportunity to thank Pr. Enzo MEDICO and Mrs. Stefania URSIDA, to whom I express gratitude for their treasured support.
I also thank my excellent thesis committee, who have provided helpful criticism of my work and have always been optimistic, encouraging and for taking time to provide occasional one-on-one mentoring.

Moreover, I would be remiss in not mentioning my family, especially my parents: Mina OUDAS and Es-Saïd AKHOUAYRI, for their wise counsel, sympathetic ear and financial support. They have endorsed me and had to put up with my stresses and moans for the past years of study. I love you both unconditionally and appreciate your efforts and love in bringing me up to be a better individual.

By the same token, special thanks to my lovely brothers, Younes AKHOUAYRI and Mohammad AKHOUAYRI, for all the help they have shown me through my career, lent moral and emotional support. Your belief in me has kept my spirits and motivation high during this process.
Also, a big thank you to uncle Omar AKHOUAYRI, who always listened to me and opened my eyes to new and promising paths. Likewise, I remain very grateful to uncle Hassan AKHOUAYRI, Soteria TSANGARI, and Latifa OUDAS for their listening, help and wonderful moments spent together.
Indeed, I have been blessed with an amazing family.

Furthermore, special thanks to Samir FOUDA, for his unparalleled help and understanding when I was going through difficult and critical times. I remain very grateful to him for making my life easier in such a crippling city.
May God bless you.

As stated above, this brief acknowledgment insufficiently recognizes the contributions of many people. Thank you to everyone, including those not listed here, who have helped, influenced, and supported me on my journey.

Finally, I could not have completed this dissertation without the support of some dear friends, lab mates, colleagues and research team.

I am indebted to all of you, mainly Dr. Anas BENJELLOUN; Dr. Eirini CHRYSANTHOU; Dr. Emir SEHOVIC; Dr. Giulio FERRERO; Dr. Raja KARAM; Dr. Soufiane HENNANI; Danilo LOMBARDI and Dr. Jamal EL HASNAOUI; for a cherished time spent together in the lab, and in social settings, who also provided me stimulating discussions as well as happy distractions to rest my mind outside of my research.

Please find here, all, the expression of a great esteem.

Laila Akhouayri

Table of contents

List of abbreviations:

AA: African-American

AD: Average distance

ADM: Average distance between means

Akt: serine–threonine kinase

APN: Average proportion of non-overlap

AUC-ROC: Area Under the Curve- Receiver operating characteristic

BC: Breast Cancer

BL: Basal-like

BLBC: Basal-like Breast Cancer

BLIA: basal-like immune-activated

BLIS: basal-like immune-suppressed

BRCA: BReast CAncer gene

cDNA: complementary DNA

COMT: catechol-O-methyltransferase

DDR: DNA damage response

DFI: Disease-Free Interval

DFS: Disease-Free Survival

DL: Deep learning

DNA: Deoxyribonucleic acid

DT: Decision Tree

EA: European-American

EGF: Epidermal growth factor

EGFR: epidermal growth factor receptor

EM: Estimation-Maximisation

EMT: Epithelio-mesenchymal transition

ER: Estrogen Receptor

ERK: extracellular-signal-regulated kinase

FISH: Fluorescence In Situ Hybridization

FLM: Fast Large Margin

FOM: Figure of merit

FUSCC: Fudan University Shanghai Cancer Center

GBT: Gradient Boosted Trees

GST: Gluthatione S-transferase

HER2: human epidermal growth factor receptor 2

HR: Hormonal Receptors

IBC: Invasive Breast Cancer

IHC: Immunohistochemistry

IM: immunomodulatory

kDa: One thousand daltons

KNN: K-nearest neighbours

LAR: luminl androgen receptor

LMIC: low- and middle-income countries

LR : Logistic Regression

Lum : Luminal

M : mesenchymal-like

ML: Machine Learning

MP: Multilayer Perceptron

MSL: mesenchymal stem-like

NB: Naive Bayes

NB-TNBC: Non-Basal-Like Triple Negative Breast Cancer

NNMP: Neural Network Multilayer Perceptron

NST: no special type

NTN-BLBC: Non-Triple Negative Basal-Like Breast Cancer

OOB: Out-Of-Bag

OS: Overall Survival

PAM: partition around medoids

PAM50: Prediction Analysis of Microarray 50.

PCA: Principal Component Analysis

pCR: Pathologic Complete response

PgR: Progesteron Receptor

PI3K: phosphatidylinositol 3-kinase

PRC: Precision-Recall curve

RF: Random Forest

RFS: relapse free survival

RNA: Ribonucleic acid

RSF: Random Survival Forests

SBR: Scarff-Bloom-Richardson

SVM: Support Vector Machine

TILs: tumor infiltrating lymphocytes

TN: Triple Negative

TNBC: Triple Negative Breast Cancer

VI: variable importance

WHO: World Health Organisation

CK: cytokeratin

FSH: Follicle-stimulating hormone

LH: Luteinizing hormone

Gn-RH: Gonadotropin-releasing hormone

GSTs: glutathione S-transferases

RFS : recurrence-free survival

ROS: reduced overall survival

LNR: Lymph Node Ratio

LIMMA: LInear Models for Microarray Analysis

CNV: copy number variations

FUSCC: Fudan University Shanghai Classification System

GEO: Gene Expression Omnibus

TCGA: the cancer genome atlas

List of figures

List of Tables

Abstract

Breast cancer is a pathology with diverse clinical, histopathologic, and molecular properties. Until recently, tumor morphology has been the gold standard for stratifying it into entities with well-established prognosis.

However, conventional morphologic classification has its own limitations, giving way to new methods that should improve patient stratification and prognostic prediction. Advances over the past decade have led to an important realization: these are different cancers. Some of them have a determined "molecular profile" that can be identified by genomic methods. Nonetheless, while there is still a long way to go to integrate a fully established picture, it will be necessary to do so in order to deploy diagnostics in clinical practice more easily.

This is where the biggest challenge is rooted, where most laboratories, hospitals and health centers do not necessarily have all the advanced technologies, allowing a better stratification of patients into their respective molecular subgroups, especially in low income countries like Morocco.

Therefore, the main goal of this work would be: to explore the heterogeneity of malignant breast cells to further elaborate their classification, and then, to elucidate the different genetic players and their underlying interactions governing each class.

Thus, our results will serve as a basis for the development and validation of new prognostic and theranostic biomarkers.

This will lead us to a much better equipped classification technique to predict the patients' membership to their respective molecular classes in a simpler, less sophisticated, less expensive and less time-consuming way to replace high throughput (transcriptomic or genomic) technologies which are potentially very gluttonous in terms of their conception.

In a similar logic and more notably, the development of a plausible treatment for triple-negative breast cancer subtypes is largely hampered by the great heterogeneity of their different phenotypes. Indeed, patients with triple-negative breast cancer are pathologically defined by the absence of ER, PgR and HER2 receptors expression.

In this project, using large transcriptomic datasets, and by applying Lehman's classification of six subtypes (Basal-like 1; Basal-like 2; Mesenchymal; Mesenchymal Stem-like; Immunomodulatory; Luminal Androgen Receptor), we were able to define a characteristic genetic signature of each subtype through several bioinformatics approaches such as clustering, and prediction algorithms such as neural networks.
This led us to discover 103 and 77 differentially and significantly over- and under- expressed genes, respectively.

Therefore, our results provide important new information that could help clinicians in the classification of triple negative breast cancer. Knowing that the chemotherapy treatment paradigm as a "one-size-fits-all" approach to managing the latter phenotype changes depending on molecular subtyping, a one-size-fits-all treatment approach is therefore questionable, making molecular subtyping crucial in determining the best treatment option for each patient.

Résumé

Le cancer du sein est une pathologie aux propriétés cliniques, histopathologiques et moléculaires variées. Jusqu'à présent, la morphologie tumorale a été l'étalon-or pour le stratifier en entités au pronostic bien établi. Cependant, la classification morphologique traditionnelle a ses propres limites, laissant place à de nouvelles méthodes qui devraient améliorer la stratification des patients et la prédiction de leur pronostic.

Les efforts des dix dernières années ont abouti à une conclusion importante: ces derniers englobent différents cancers. Certains d'entre eux ont un "portrait moléculaire" défini qui peut être identifié par des méthodes génomiques. Cependant, si le chemin vers l'intégration d'un portrait complètement établi est encore long, il faudra le faire afin de déployer plus aisément le diagnostic dans la pratique clinique.

C'est là qu'est enraciné le plus grand défi, où la plupart des laboratoires, hôpitaux et centre de santé n'ont pas tous nécessairement des technologies de pointe; permettant une meilleure stratification des patientes en leurs sous-groupes moléculaires respectifs, surtout dans les pays à faible revenu comme le Maroc.

Par conséquent, le but principal de ce travail serait: d'explorer l'hétérogénéité des cellules mammaires malignes pour étoffer davantage leur classification, ensuite, d'élucider les différents acteurs génétiques et leurs interactions sous-jacentes régissant chaque classe.
Ainsi, Nos résultats serviront de base au développement et à la validation de nouveaux biomarqueurs à la fois pronostiques et théranostiques.

Ceci nous conduira à une technique de classification beaucoup plus outillée à prédire l'appartenance des patientes à leurs classes moléculaires respectives, et ce, de manière plus simple, moins

sophistiquée, moins chère et moins chronophage pour remplacer les technologies haut débit (transcriptomiques ou génomiques) qui sont potentiellement très gloutonnes en termes de leur conception.

Dans la même suite logique et plus particulièrement, le développement d'un traitement plausible pour les sous-types de cancer du sein triple négatif est largement entravé par la grande hétérogénéité de leurs différents phénotypes. En effet, les patientes atteintes de cancer du sein triple négatif sont pathologiquement définies par l'absence d'expression des trois récepteurs ER, PgR et HER2.

Dans ce projet, en utilisant de grands ensembles de données transcriptomiques, et en appliquant la classification de Lehman en six sous-types (Basal-like 1; Basal-like 2; Mesenchymal; Mesenchymal Stem-like; Immunomodulatory; Luminal Androgen Receptor), nous avons pu définir une signature génétique caractéristique de chaque sous-type par le biais de plusieurs approches bio-informatiques de datamining comme le clustering, et d'algorithmes de prediction comme les réseaux de neurones. Ceci nous a conduit à découvrir 103 et 77 gènes différentiellement et significativement sur- et sous-exprimés respectivement, Par conséquent, nos résultats apportent de nouvelles informations importantes qui pourraient aider les cliniciens dans la classification du cancer du sein triple négatif.

Sachant que le paradigme de traitement par chimiothérapie comme approche " unique qui sied à tous les phénotypes" pour la gestion de ce dernier, change en fonction du sous-typage moléculaire, une approche thérapeutique unique est donc contestable, ce qui rend le sous-typage moléculaire crucial pour déterminer la meilleure option thérapeutique pour chaque patient.

ملخص

سرطان الثدي هو علم أمراض له خصائص سريرية ونسيجية وجزيئية متنوعة. حتى وقت قريب، كان مورفولوجيا الورم هو المعيار الذهبي لتقسيم سرطان الثدي إلى كيانات ذات تشخيص راسخ. ومع ذلك، فإن التصنيف المورفولوجي التقليدي له حدوده الخاصة، مما يفسح المجال للطرق الجديدة التي من شأنها تحسين التقسيم الطبقي للمريض والتنبؤ.

أدت التطورات التي حدثت خلال العقد الماضي إلى إدراك مهم: هذه سرطانات مختلفة. بعضها لديه "ملف جزيئي" محدد يمكن تحديده بالطرق الجينية. ومع ذلك، بينما لا يزال هناك طريق طويل يجب قطعه لدمج صورة كاملة التأسيس، سيكون من الضروري القيام بذلك من أجل نشر التشخيص في الممارسة السريرية بسهولة أكبر.

هذا هو المكان الذي يتجذر فيه التحدي الأكبر، حيث لا تحتوي معظم المعامل والمستشفيات والمراكز الصحية بالضرورة على جميع التقنيات المتقدمة؛ السماح بتقسيم أفضل للمرضى إلى مجموعات فرعية جزيئية خاصة بهم، خاصة في البلدان منخفضة الدخل مثل المغرب،

لذلك، سيكون الهدف الرئيسي من هذا العمل هو: استكشاف عدم تجانس خلايا الثدي الخبيثة لمزيد من التفصيل في تصنيفها، ومن ثم توضيح اللاعبين المختلفين والتفاعلات الأساسية التي تحكم كل فئة

وبالتالي، فإن نتائجنا ستكون بمثابة أساس لتطوير والتحقق من صحة المؤشرات الحيوية الإنذارية والتشريحية الجديدة.

سيقودنا هذا إلى تقنية تصنيف أفضل تجهيزًا للتنبؤ بعضوية المرضى في الفئات الجزيئية الخاصة بهم بطريقة أبسط وأقل تعقيدًا وأقل تكلفة وأقل استهلاكا للوقت لاستبدال التقنيات عالية الإنتاجية النسخية أو الجينومية

في منطق مماثل وبشكل أكثر وضوحًا، فإن تطوير علاج معقول لأنواع فرعية من سرطان الثدي ثلاثي السلبي يعوقه إلى حد كبير التباين الكبير في أنماطها الظاهرية المختلفة. في الواقع، يتم تعريف المرضى الذين يعانون من سرطان الثدي السلبي الثلاثي من الناحية المرضية من خلال التعبير السلبي الثلاثي عن مستقبلات الأستروجين والبروجسترون ومستقبلات عامل نمو البشرة البشرية 2

في هذا المشروع، باستخدام مجموعات بيانات كبيرة من النسخ، وتطبيق التصنيف إلى ستة أنواع فرعية (تشبه القاعدية 1 ؛ تشبه القاعدية 2 ؛ اللحمة المتوسطة ؛ تشبه الجذعية المتوسطة ؛ التعديل المناعي ؛ مستقبل الأندروجين اللمعي) ، تمكنا من تحديد التوقيع الجيني مميزلكل نوع فرعي من خلال العديد من مناهج بيانات المعلوماتية الحيوية مثل التجميع وخوارزميات التنبؤ مثل الشبكات العصبية. قادنا هذا إلى اكتشاف 103 جينات مفرطة التعبير و 77 جينة ناقصة التعبير بشكل تفاضلي وبشكل ملحوظ.لذلك ، توفر نتائجنا معلومات جديدة مهمة يمكن أن تساعد الأطباء في تصنيف سرطان الثدي الثلاثي السلبي.

مع العلم أن نموذج العلاج الكيميائي كنهج "مقاس واحد يناسب الجميع" لإدارة تغييرات العلاج الكيميائي استنادًا إلى التصنيف الفرعي الجزيئي، فإن نهجًا علاجيًا واحدًا يناسب الجميع أمر مشكوك فيه، مما يجعل التصنيف الفرعي للجزيء أمرًا ضروريًا لتحديد أفضل علاج خيار لكل مريض

FOREWORD

Breast carcinoma (BC) is a pathology with well-defined clinical, histopathological, and molecular characteristics. Tumor morphology has been the gold standard to classify breast tumors into entities with definite prognosis. However, the traditional morphological classification has limitations, which leaves room for new molecular methods supposed to improve the patients' stratification and prediction prognosis.

This study is a collaborative work between Italy and Morocco; therefore, the present manuscript is conventionally divided into two main and distinctive chapters, yet evoking complementary axes concerning the refinement of BC classification.

The first chapter evokes a theme elaborated mainly in the Cancer Genomics laboratory, Fondazione Edo ed Elvo Tempia, Biella - Italy, within the framework of the Complex Systems for Quantitative Medicine doctoral school of Torino university.

While the second chapter evokes another theme elaborated in the Pathology Department of Ibn Rochd University Hospital of Casablanca; under the tutelage of the Molecular Pathology and Genetics research unit of Hassan-II University Medicine College.

The main perspective discussed in the first chapter concerns Triple Negative Breast Cancers (TNBC) subtyping, according to the combination of several biostatistical and genomic techniques.

Therefore, the principal interest is finding a minimal genetic signature that will be able to help differentiate one triple negative subgroup from another in a distinctive way. Subsequently, in order

to guarantee a certain generalized reproducibility of our proposed outcome, we also recurred to an external validation based on the TNBC-TCGA and Italian datasets, containing TNBC records exclusively.

On the other hand, the main question addressed in the second chapter is whether the molecular classification actually accepted as a reference for BC subtypes determination, can be refined by statistical partitioning methods. This is especially useful in low-income countries such as Morocco, where laboratories do not necessarily have access to new molecular testing methods, which are proving to be very expensive like the gene expression profiling assays.

In addition, our work also aims to assess the prediction degree of these statistical approaches, along with their classification accuracy. Hence the interest of having used mainly a large Moroccan database, which was compared to two other different ones: METABRIC and TCGA-BRCA, the latters served as external validation for our Northern African comparative Cohort-study.

Both chapters follow the same basic structure, where they are subdivided into the introduction section, which presents the essential elements to the justification and easy understanding of the addressed issues. The material and methods section expands on the different methods and means used to study the evoked issues. The results section elucidates the outcome obtained. A final section discusses the results, according to other similar and comparative bibliographic sources. The latter is followed by a more general discussion to conclude the manuscript.

GENERAL OVERVIEW

A. CLINICAL OVERVIEW:

A.1. Breast histology:

a) The mammary gland:

The breast is an exocrine gland whose biological function is to produce milk to nourish the newborn. The mammary gland is made up of two distinct cellular compartments (Figure1):

-The epithelial compartment: made up mainly of fat, connective tissue, lobes, and ducts: this compartment is articulated around a network of lobes, each lobe consisting of several lobules, themselves made up of acini which produce milk. There are between fifteen and twenty lobes in the breast. Each lobe is made up of 20 to 40 lobules each having their own excretory duct or milk duct, into which the secondary ducts of the acini and lobules flow. The milk ducts converge towards the nipple, they widen to form the lactiferous sinuses, then narrow and emerge at the pores of the nipple.

The lobules and ducts are composed of two differentiated cell types, luminal and myoepithelial cells, as well as a low number of progenitor ones (stem cells and immature precursors). The luminal cells are in contact with the lumen of the lobules and channels, they express the cytokeratins CK8 / 18 and CK19. Myoepithelial cells surround luminal cells and are in contact with the basement membrane and surrounding stroma, they express CK5 / 6, CK14, CK17 and smooth muscle markers such as smooth muscle actin and p63(Hassiotou and Geddes 2013);(Aranda-Gutierrez and Diaz-Perez 2020).

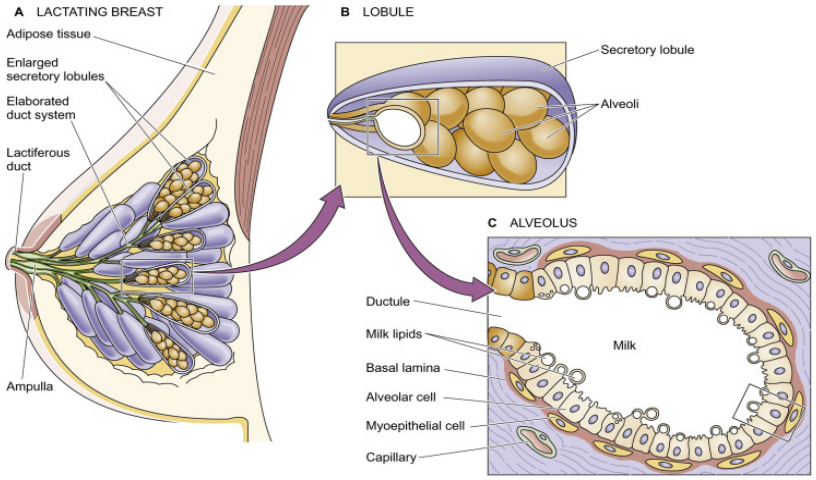

Figure1: Mammary gland structure. A:lactating breast; B: Lobule; C: Alveolus.

©ScienceDirect:Boron and Boulpaep (2012)

-The mesenchymal compartment: made up of blood and lymphatic vessels.

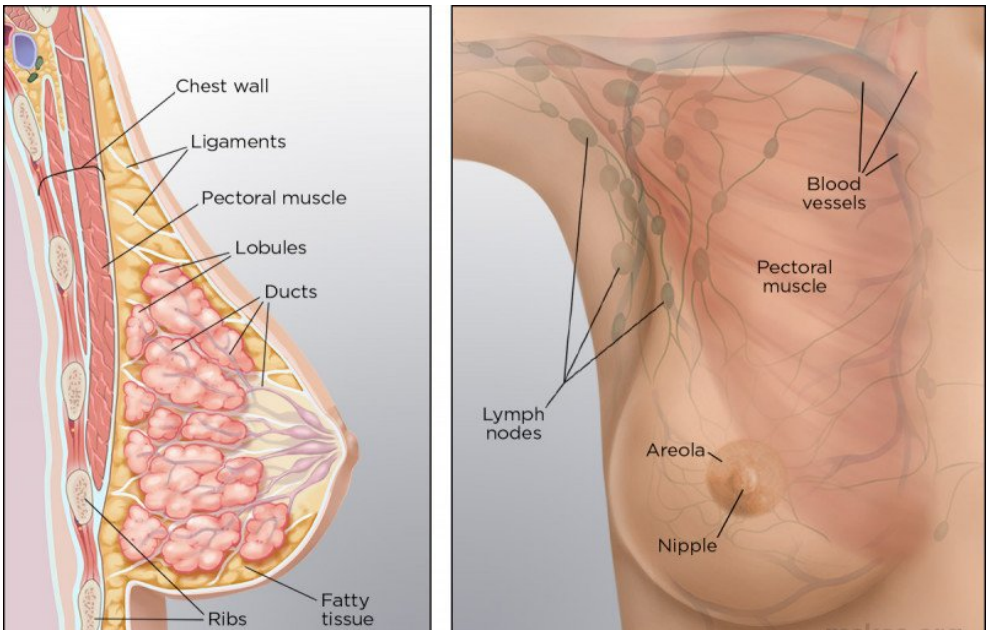

Figure2: Mammary gland anatomy (Right: frontal view, Left: lateral view)

©VirginiaOncologyAssociates

b) Structure:

The skin envelope: the skin covering the breast is not homogeneous, three areas are described.

-Peripheral zone : it is smooth and supple.

-Middle zone : it is the areola, it is pigmented, circular 35 to 50 mm in diameter. Its appearance is made granular by the presence of sebaceous glands : MORGAGNI tubers. These glands become larger during pregnancy and take the name of MONTGOMERY tubers.

-Central zone: this is the nipple, it occupies the center of the areola, its pigmentation is identical to that of the areola. (Figure2).

The cellulo-adipose envelope, which is formed by two fatty layers:

-The anterior pre glandular layer: does not exist at the level of the areola-nipple plate. It is partitioned by connective spans: the Cooper ligaments, which connect the skin to the gland.

-The posterior layer: is bounded by the fascia superficialis, it is separated from the aponeurosis of the pectoralis major by connective tissue. The gland-fat skin assembly slides over the major pectoralis

c) Lymphatics:

There are three ways of lymphatic drainage and their importance is capital in terms of the spread of BC (Figure2):

Axillary nodes, with 2 drainage channels:

-Main: towards the pectoral group, at the level of the axillary fossa.

-Accessory: towards the apical nodes

Parasternal nodes: they drain the medial part of the gland.

Supraclavicular nodes: they drain the upper part of the gland.

A.2. Breast Cancer

A.2.1. Breast Cancer related indicators

A) Biomarkers :

a) Estrogen receptor (ER):

Definition*:* Estrogens constitute a group of steroids. They are produced primarily by the development of ovary follicles by the placenta. Some estrogen is also produced in small amounts by other tissues such as the fatty tissue. These secondary sources of estrogen are especially important in postmenopausal women. They promote the development of female secondary sex characteristics, such as breasts, and are also involved in controlling the menstrual cycle, which is why most hormonal contraceptives like birth control pills contain them (Yager and Davidson 2006).

Estrogen receptor: They are intracellular proteins belonging to the nuclear receptor family and encoded by two distinct genes, possessing the two types of receptors: these are the alpha (ERα) and beta (ERβ). Which are distributed differently according to the organs. A third potential receptor,

belonging to another family (receptor coupled to G proteins), encoded by a third gene called GPR30 has been described by (Prossnitz, Arterburn, and Sklar 2007);(Xu et al. 2019). The effects of estrogen on their target cells / tissues through these receptors can be classified into two categories: genomic effects, i.e. on gene expression; and non-genomic effects which directly concern other molecular actors in cells, mainly proteins (Fujimoto and Kitamura 2004; “Estrogen Receptor and Breast Cancer” 2001).

Normal action of estrogen on the mammary gland: At puberty, the hypothalamic secretion of Gn-RH results in FSH and LH secretion by the adeno-pituitary gland. These determine changes in the ovaries that will be responsible for those affecting the genital tract (menstrual cycle). During the first menstrual cycles, under the influence of the ovarian estrogen secretion, the mammary glands develop: canalicular proliferation is accompanied by a significant development of the interlobar and interlobular connective tissue as well as an increase in adipose cells (Wu et al. 2017). Outside of pregnancy and breastfeeding, the mammary glands remain "at rest". Only a few tubulo-alveolar can develop in the second part of the cycle under the influence of progesterone. In the absence of pregnancy, these tubuloalveolar involute (Figure3). The normal human mammary gland undergoes a well-defined sequence of histological changes during the menstrual cycle in both epithelial structures and stroma. The extracellular matrix plays a central role in modulating a wide variety of cellular events, such as proliferation, differentiation and genes expression (Bernstein and Press 1998);(Pike et al. 1993).

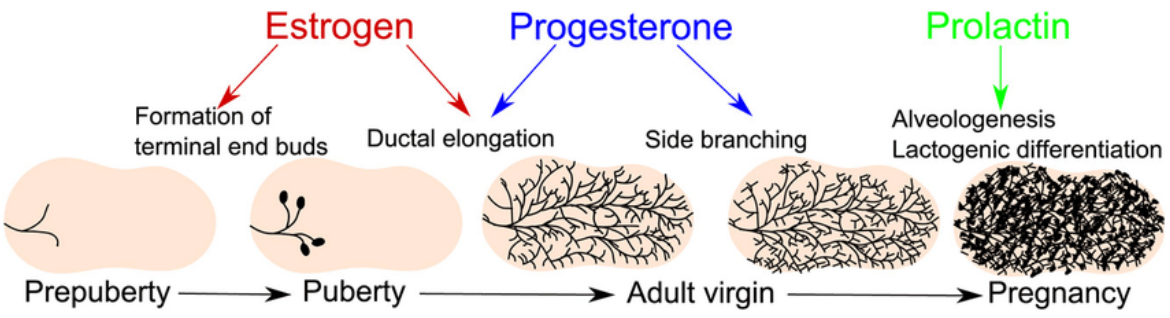

Figure 3: Hormonal regulation of the mammary gland development.

(© Rosario & al, 2007)

Estrogen and breast carcinogenesis: Several studies have demonstrated the great involvement of estrogens in the development and growth of BC. While the exact mechanisms remain to be fully elucidated, the alkylation of cellular molecules and the generation of active radicals that can damage DNA, including the potential for genotoxicity of estrogen and some of its metabolites (eg, catechol estrogens) are suspected as agents of tumorigenesis (Nandi, Guzman, and Yang 1995). Oxidized estrogen derivatives stimulate the growth of mammary tumors in vitro and in vivo (catechols) and, above all, can bind to DNA and proteins (genotoxicity of quinones). The production of catechols depends on the activities of the enzymes that produce them, which themselves depend on the tissue and the inducers (presence/absence). When the synthesis of catechols becomes excessive, the detoxification systems (COMT, sulfotransferases, UDP-glucuronosyltransferases) are overwhelmed and the derivatives, semiquinones and quinones, are produced. The second protective barrier is the conjugation of quinones using glutathione S-transferases (GSTs). When glutathione stores are depleted, quinones can exert their genotoxicity. The formation of quinones from semiquinones can lead to the formation of reactive oxygen derivatives such as the superoxide ion O2- as well as to “futile” oxidation-reduction cycles, even in the presence of small amounts of oxygen (Roy, Strobel, and Liehr 1991). Excessive production of these reactive derivatives can have deleterious consequences for cells because these molecules damage DNA, lipids and proteins (Figure4). While progesterone acts via PgRs on the lobuloalveolar differentiation of the breast. In normal human mammary epithelium, 98% of proliferative cells do not express PgR nor ER. On the other hand, PgRs and ERs are colocalized in luminal epithelial non-proliferative cells located near proliferative cells. This property is lost when cells become cancerous (Tian et al. 2018).

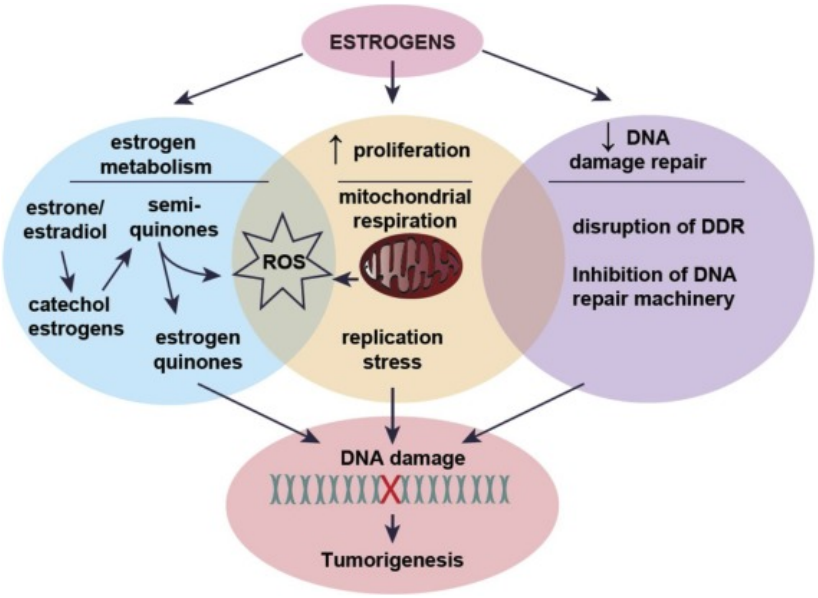

Figure4: estrogen action mechanism in breast cancer

b) Progesterone receptors (PgR):

Definition: Progesterone is a steroid hormone mainly secreted by the corpus luteum cells of ovaries and placenta. It is involved in pregnancy (by supporting gestation) and embryogenesis as well. It is produced in the second half of the menstrual cycle after ovulation, and its production decreases in the absence of fertilization. Progesterone acts on the endometrium, and allows menstruation to occur at the end of each menstrual cycle. In the absence of progesterone, the endometrium grows too much under the effect of estrogen, which can result in genital hemorrhages, an increase in the uterus volume, and promotes uterine tumor development. This is the main reason why, after menopause, when estrogen-based treatment (as part of hormonal treatment) is given to a patient, the combination of progesterone is mandatory.

Normal physiological function of Progesterone: The main physiological function of progesterone in the mammary gland is to prepare lactation in synergy with estradiol and prolactin. Pregnancy is a period of estrogenic, progesterone and prolactin inflation. The combination of these three hormones will result in maximum differentiation of the epithelial fraction. This would be the explanation of the protective effect against the risk of BC in the first full precocious term of pregnancy.

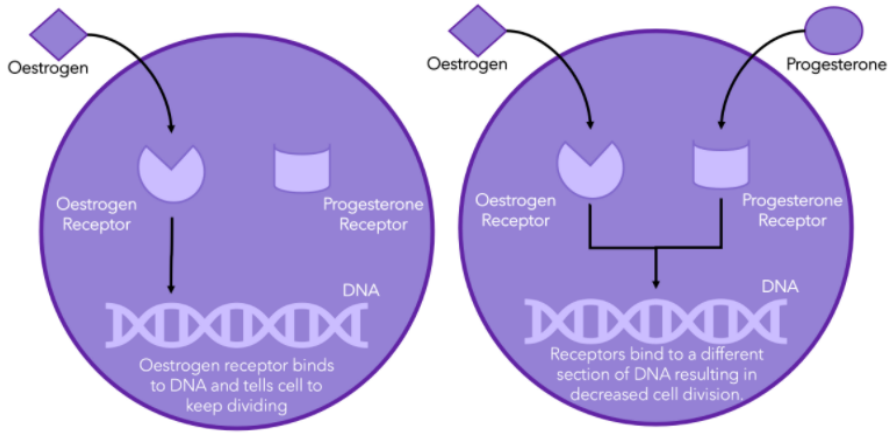

Figure 5: ER/PgR duality influence on breast cancer growth

(©OWis)

PgR is activated upon binding to progesterone. There are three isoforms of PgR: PgR-A (94 kDa); PgR-B (116 kDa), and PgR-C (60 kDa) isoforms. These receptors have different transcriptional activities. PgR-B for example is essential for normal breast development (Biserka Mulac-Jericevic et al. 2003), on the other hand, PgR-A is more important for uterine development (B. Mulac-Jericevic et al. 2000) and reproduction. PgR-C can enhance PgR activity in BC cells (Wei, Norris, and Baker 1997), or function as a dominant inhibitor of PgR-B in the uterus (Condon et al. 2006).

Progesterone and breast Carcinogenesis: Some authors have hypothesized that progesterone could have a synergistic effect with estradiol in its promoting role on the occurrence of BC (Figure5). Growth factors, including EGF or heregulin, promote transcriptional synergy with progestins on PgR-target genes (Qiu and Lange 2003);(Daniel et al. 2007);(Shen, Horwitz, and Lange 2001). Several genes appear to be regulated depending on PgR expression but not progesterone (Jacobsen et al. 2002);(Jacobsen et al. 2005). Other genes are downregulated in response to progesterone/PgR-dependent transcriptional repression by unknown mechanisms (Richer et al. 2002). Mainly the regulation of particular genes in response to progesterone is correlated to changes in cell biology. For example, many PgR-regulated genes have been associated with aspects of tumor progression towards aggressive tumor phenotype. Also, variation of the

PgR-A: PgR-B ratio is very frequent in breast tumors (Richer et al. 2002; Graham et al. 1995), and is associated with genetic alterations.

c) HER2 receptors :

HER2 is a membrane receptor with three domains: extracellular interacting with the ligand, lipophilic transmembrane, and intracellular tyrosine kinase activity. HER2, if activated by a growth factor, can induce two signaling pathways: the Ras / ERK one and the PI3 / AKT channel. This receptor is thus at the origin of the triggering of many biological reactions leading in particular to cell proliferation, angiogenesis and resistance to apoptosis, etc.(L.-M. Sun et al. 2019; Shah and Osipo 2016);(Gutierrez and Schiff 2011) (Figure6).

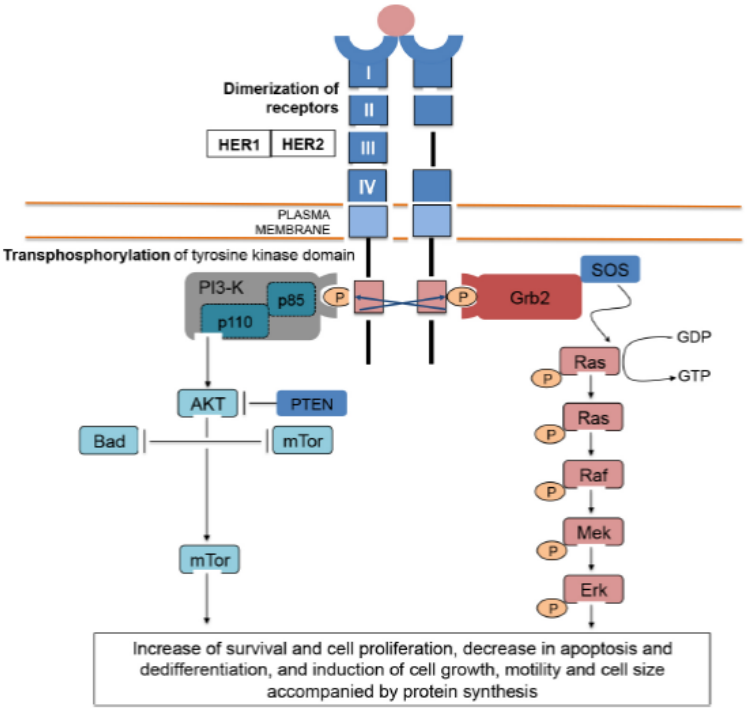

Figure6: Transmembrane receptors with tyrosine kinase activity (EGFR / HER2 / c-MET) and underlying signaling pathway promoting tumor growth
(Gutierrez and Schiff 2011; Ferreira and Pessoa 2017)

Normally in all healthy breast cells, HER2 receptors help in controlling the cell's growth, division, and itself reparation, but in about 20% to 25% of BC cases, HER2 undergoes different types of alterations which generates too many copies (known as HER2 gene amplification) or results in a constitutive activation of the receptor. Therefore, the presence of all these receptors produces increased amplification signals. This makes breast cells grow and divide in an uncontrolled way.

Overexpression of HER2 receptor is generally associated with a poor prognosis with high tumor aggressiveness, an important metastatic tendency and a reduced life expectancy.

The amplification of its gene is investigated in cancer cells. The pathologist can routinely use two techniques (Immunohistochemistry “IHC” and / or Fluorescence In Situ Hybridization “FISH”) to perform this analysis. BCs in which HER2 overexpression is found tends to behave more aggressively. However, since 2005, the use of a monoclonal antibody targeting HER2 (trastuzumab) has greatly improved the survival of women with this type of cancer. Since then, other similar drugs have been developed (Ishikawa et al. 2014).

In so-called pure HER2+ tumors, there is a strong overexpression of HER2 which is due to the gene amplification. This high level of expression leads to its spontaneous dimerization and the increased activation of signaling cascades located downstream and involved in cell proliferation and survival (Iqbal and Iqbal 2014). There is a semi-quantitative score to classify HER2 status in four categories: 0, 1+, 2+ or 3+. A search for amplification by in situ hybridization is required when the score is 2+.

d) KI-67 proliferation index

The Ki-67 antigen is a proliferation marker that is present on a nuclear protein encoded by the MKI-67 gene and expressed in all cell cycle phases, except the G2 phase (Gerdes et al. 1983).

It is located precisely in the proliferative cell's nucleus, and is expressed at the chromosome's periphery and acts as a surfactant which keeps the mitotic chromosomes separated. It also participates in maintaining cell proliferation and is involved in the first cell cycle phases in the ribosomal RNA synthesis by the RNA-polymerase-I enzyme.

KI-67 localization during interphase is associated with different functions compared to its localization during mitosis. Ki-67 is required for heterochromatin antigens normal distribution and for heterochromatin’s association in cells during the interphase (precisely it is located in the dense fibers of the nucleolus). On the other hand, Ki-67’s presence is mandatory for the perichromosomal layer formation, and chromosome condensation during mitosis (Figure7).

In this context, the latter usually acts to prevent aggregation of mitotic chromosomes (X. Sun and Kaufman 2018). Therefore, Ki-67 expression varies throughout the cell cycle and peaks during mitosis. Its function is not fully elucidated but its role in cell division and ribosomal RNA synthesis is clearly established (Matheson and Kaufman 2017).

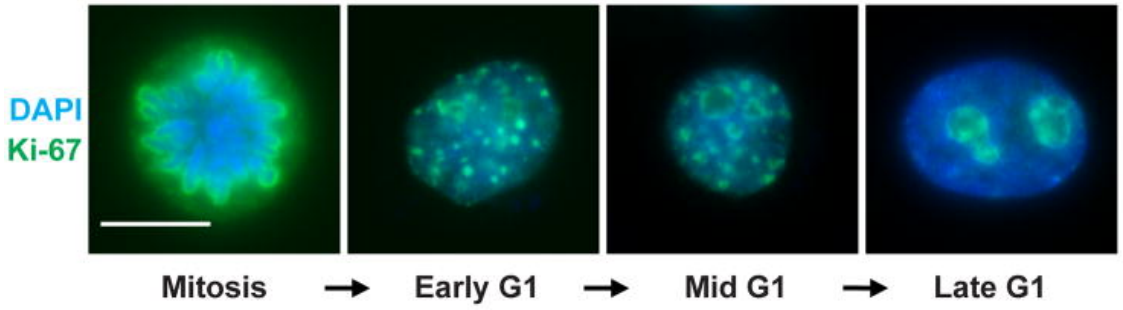

Figure7: Ki-67 localization throughout the cell cycle

(HeLa S3 cells were stained with anti-Ki-67 antibodies (green) and DAPI to visualize DNA (blue), illustrating different Ki-67 localizations across the cell cycle. In mitotic cells, Ki-67 coats the condensed chromosomes as the foundation of the perichromosomal layer. As cells exit mitosis and enter early G1 phase, small puncta of Ki-67 leave the decondensing chromosomes. These then coalesce at the periphery of the reformed nucleoli as G1 phase progresses. Scale bar, 10 µm) (Matheson and Kaufman 2017)

Its detection is done through the anti-Ki-67 antibody by immunohistochemistry and immunofluorescence and the result gives nuclear labeling. In practice, the Ki-67 labeling index represents the percentage of marked cells in the total number of invasive cancer cells and proves to be a useful test in order to assess cell proliferation in cancers; and predicts the sensitivity of a tumor to cytotoxic agents, in particular, in BC. Its evaluation is therefore mainly used as a prognostic factor to guide the adjuvant therapy decision. GeparTrio, being the largest study to have evaluated the clinical, prognostic and predictive ability of Ki-67 after neoadjuvant chemotherapy, distinguished 3 classes of patients, according to Ki-67 index staining levels (0- 15% versus 15.1-35% versus > 35.1%).

-The class with a low Ki-67 had a result comparable to that of the class with pathologic complete response (pCR).

-The class with a high Ki-67 had a significantly higher risk of recurrence and mortality than the class with a low or intermediate Ki-67.

Ki-67 is now mainly used to distinguish Luminal A from Luminal B among ER+/HER2- cancers and, therefore , to guide adjuvant chemotherapy decisions instead of hormone therapy alone. As

mentioned by the St. Gallen expert committee in 2015: “The distinction between Luminal A cancers, which are hormone-sensitive, weakly proliferative and with a good prognosis, and Luminal B cancers, less sensitive to hormone therapy, with higher proliferation and unfavorable prognosis, could be defined by IHC tests for ER, PgR and Ki-67.

To summarize the difficulty of Ki-67 clinical benefit assessment, C. Denkert et al. defined 3 different groups of cancers:

-Low-proliferation cancers that do not respond to chemotherapy but have a good prognosis (low Ki-67 is associated with a good prognosis)

-Highly proliferating and chemo-sensitive cancers, for which a high Ki-67 is associated with a better pCR rate and better survival (a high Ki-67 is associated with a good prognosis)

-Highly proliferating, chemo- or hormone-resistant cancers, for which an increase in Ki-67 is associated with a reduction in survival (a high Ki-67 is associated with a poor prognosis) "(Denkert et al. 2013; Alba et al. 2016).

e) Summary of standard biomarkers required in the IBC diagnosis

Table1 below presents the biomarkers summary required for IBC diagnosis according to the WHO (WHO classification of tumors series, 5th ed. vol. 2.2019)

Table 1: Summary of standard biomarkers required in the IBC diagnosis: purpose, reporting, and scoring criteria

(WHO classification of tumors series, 5th ed. vol. 2.2019)

Biomarker and purpose	Test type	Reporting categories	Scoring criteria (ASCO/CAP)
ER: Benefits from hormone therapies if positive **Other uses:** Categorization for overall treatment pathways Characterization as the luminal group if positive Poor prognostic marker if negative	IHC	Positive Negative	>1% of invasive cancer with nuclear staining of any intensity < 1% of invasive cancer with nuclear staining
PgR: Validated for primarily prognostic in ER-positive cancers **Other uses:** Poor prognostic marker if negative	IHC	Positive Negative	>1% of invasive cancer with nuclear staining of any intensity < 1 % of invasive cancer with nuclear staining

ERBB2 (HER2): Validated for HER2-targeted therapy if positive **Other uses:** Categorization for overall treatment pathways Characterization as the HER2-enriched subtype (if ER-negative) or luminal B if (ER-positive) Marker of aggressive biology	IHC FISH IHC	Positive Equivocal Negative	3+: Circumferential membrane staining that is complete, intense, and in >10% of tumor cells 2+: Weak to moderate complete membrane staining observed in >10% of tumor cells (MUST be confirmed with FISH) 1+/ 0+: Incomplete / no staining or barely perceptible in > 10% of tumor cells.

B) Histological factors:

a) Invasive breast cancer histopathology :

Invasive breast cancer (IBC) has a very broad spectrum of histological appearances that cannot be detailed in this report. Based on its architecture, IBC is divided into several subtypes, some have specific definitions such as lobular, tubular, papillary and mucinous tumors, constituting 20% of all BCs, while tumors lacking such specific features are designated as IBC of no special type (NST), thought to be non-specific histologically, and are the most common (80%).

Some specific histological types are considered to have a better prognosis than NST tumors, such as tubular carcinoma and cribriform carcinoma (Elston 2005).

b) Histological (Nottingham) Grading:

The most widely used histologic grading system of IBC is Nottingham combined with the histologic grade (Elston-Ellis modification of Scarff-Bloom-Richardson grading system), also known as the Nottingham grading system (WHO 2019). This grading evaluates 3 parameters: differentiation (glandular formation), nuclear pleomorphism and mitotic count as detailed in table2. Nottingham grading system is part of the minimum data set for BC pathology reporting as it proved to be an independent prognostic factor that provides outcome prediction in patients with IBC and with a very good interobserver agreement (Rakha et al. 2008).

Table 2: Nottingham grading system in breast tumors

(WHO classification of tumors series, 5th ed. vol. 2.2019)

Criteria	Note
Tubule formation scoring	
● The tumor mainly comprises tubes (>75%) ● Moderate (10 - 75%) ● No tube or very few (<10%)	1 2 3
Nuclear pleomorphism scoring	
● Small, regular, and uniform nuclei ● Size + Variability ● Marked variation	1 2 3
Number of mitoses	
● <10 mitoses ● 11< number of mitoses< 20 ● >21 mitoses	1 2 3
SBR final grading by adding the scores for gland formation, nuclear polymorphism, and mitotic count:	
Score	**Grade**
3 to 5 score	I: Least aggressive tumor
6 to 7 score	II: Intermediate aggression
8 to 9 score	III: Very aggressive tumor

c) Other histological factors:

Vascular emboli are mainly defined by the presence of tumor cells within vascular structures, outside the tumor, their presence is directly correlated with histological grade, breast invasion depth, age and tumor size, which makes it a defavorable prognosis marker, therefore, it should be reported when present. Vascular emboli are typically seen within vascular spaces throughout the tumor, but only extra-tumoral vascular invasion is routinely assessed in BC (David Nathanson et al. 2020).

Lymph node invasion is incorporated into the TNM staging and constitutes an important prognosis factor. Axillary lymph nodes are most likely to be affected. The other lymph nodes often affected are the nodes around the clavicle (sub- and supraclavicular nodes) and the mammary nodes near the breastbone.

Other factors include margin status, tumor infiltrating lymphocytes (TILs) and the in situ component.

d) TNM Staging

The TNM system for describing the anatomical disease extent is the most common staging system for tumors, and is based on three components assessment:

T– the extent of the primary tumor,

N– the absence or presence and extent of regional lymph node metastasis,

M– the absence or presence of distant metastasis. There are 2 classifications :

-*The clinical classification:* which is the pretreatment clinical classification, designated TNM.

-*The pathological classification:* the postsurgical histopathological classification, designated pTNM, used to guide adjuvant therapy and provides additional data to estimate prognosis and end results. This is based on evidence acquired before treatment, supplemented or modified by additional evidence acquired from surgery and from pathological examination (Table3).

Table 3: TNM system for staging Breast Cancer

(AJCC, 8th edition, 2017)

T – Primary Tumor TX Primary tumor cannot be assessed T0 No evidence of primary tumor Tis Carcinoma in situ Tis Ductal carcinoma in situ (DCIS) Tis Lobular carcinoma in situ(LCIS) Tis Paget disease of the nipple not associated with invasive carcinoma and/or (Paget) carcinoma in situ (DCIS and/or LCIS) in the underlying breast parenchyma. Carcinomas in the breast parenchyma associated with Paget disease are categorised based on the size and characteristics of the parenchymal disease, although the presence of Paget disease should still be noted. T1 Tumor 2 cm or less in greatest dimension T1mi Microinvasion 0.1 cm or less in greatest dimension T1a More than 0.1 cm but not more than 0.5 cm in greatest dimension T1b More than 0.5 cm but not more than 1 cm in greatest dimension T1c More than 1 cm but not more than 2 cm in greatest dimension T2 Tumor more than 2 cm but not more than 5 cm in greatest dimension T3 Tumor more than 5 cm in greatest dimension T4 Tumor of any size with direct extension to chest wall and/or to skin (ulceration or skin nodules) T4a Extension to chest wall (does not include pectoralis muscle invasion only) T4b Ulceration, ipsilateral satellite skin nodules, or skin oedema (including peau d'orange) T4c Both 4a and 4b
N – Regional Lymph Nodes NX Regional lymph nodes cannot be assessed (e.g., previously removed) N0 No regional lymph node metastasis N1 Metastasis in movable ipsilateral level I, II axillary lymph node(s) N2 Metastasis in ipsilateral level I, II axillary lymph node(s) that are clinically fixed or matted; or in clinically detected* ipsilateral internal mammary lymph node(s) in the absence of clinically evident axillary lymph node metastasis N2a Metastasis in axillary lymph node(s) fixed to one another (matted) or to other structures

N2b Metastasis only in clinically detected* internal mammary lymph node(s) and in the absence of clinically detected axillary lymph node metastasis N3 Metastasis in ipsilateral infraclavicular (level III axillary) lymph node(s) with or without level I, II axillary lymph node involvement; or in clinically detected* ipsilateral internal mammary lymph node(s) with clinically evident level I, II axillary lymph node metastasis; or metastasis in ipsilateral supraclavicular lymph node(s) with or without axillary or internal mammary lymph node involvement N3a Metastasis in infraclavicular lymph node(s) N3b Metastasis in internal mammary and axillary lymph nodes N3c Metastasis in supraclavicular lymph node(s)
M – Distant Metastasis M0 No distant metastasis M1 Distant metastasis
pTNM Pathological Classification pT – Primary Tumor The pT categories correspond to the T categories. pN – Regional Lymph Nodes The pathological classification requires the resection and examination of at least the low axillary lymph nodes (level I) (see page 152). Such a resection will ordinarily include 6 or more lymph nodes. If the lymph nodes are negative, but the number ordinarily examined is not met, classify as pN0. pNX Regional lymph nodes cannot be assessed (e.g., previously removed, or not removed for pathological study) pN0 No regional lymph node metastasis*

A.2.2. Breast Cancer classifications:

A) Breast Cancer molecular classification

The histological and clinical heterogeneity of BC, partly responsible for the therapeutic failures, reflects its complex molecular nature. A complete molecular characterization is essential. It is based on hierarchical cluster analyses of genes. Due to time and cost constraints, as high-throughput transcriptomic analysis is not available in the vast majority of settings, in most health care systems, the molecular classification of BC has been simplified and routinely based on immunohistochemical analysis of four main biomarkers (ER, PgR, HER2 and Ki-67). Their expression is now used as surrogate markers to establish the IBC molecular classification, an essential tool to assess differences in tumor behavior and prognosis. Nevertheless, examination of the global gene expression patterns has led to the identification of certain intrinsic molecular subtypes which have biological and clinical relevance, as well as certain genomic signatures predictive of the response to treatment (Colomer et al. 2018).

a) Transcriptomic Classification:

Since the 2000s, the study carried out by Sorlie & al. has shown that BCs could be classified into six molecular subgroups defined by their gene expression profile. They thus determined for the first time 6 molecular subtypes of BC: Luminal A, Luminal B, Luminal C, Normal Breast like, Pure HER2 + and Basal like *(Figure8)*.

This classification encompasses gene expression patterns derived from cDNA microarrays, and correlates tumor characteristics to clinical outcome, to guide the oncologist towards an adequate therapeutic strategy depending on the overexpressed receptors (Sørlie et al. 2001).

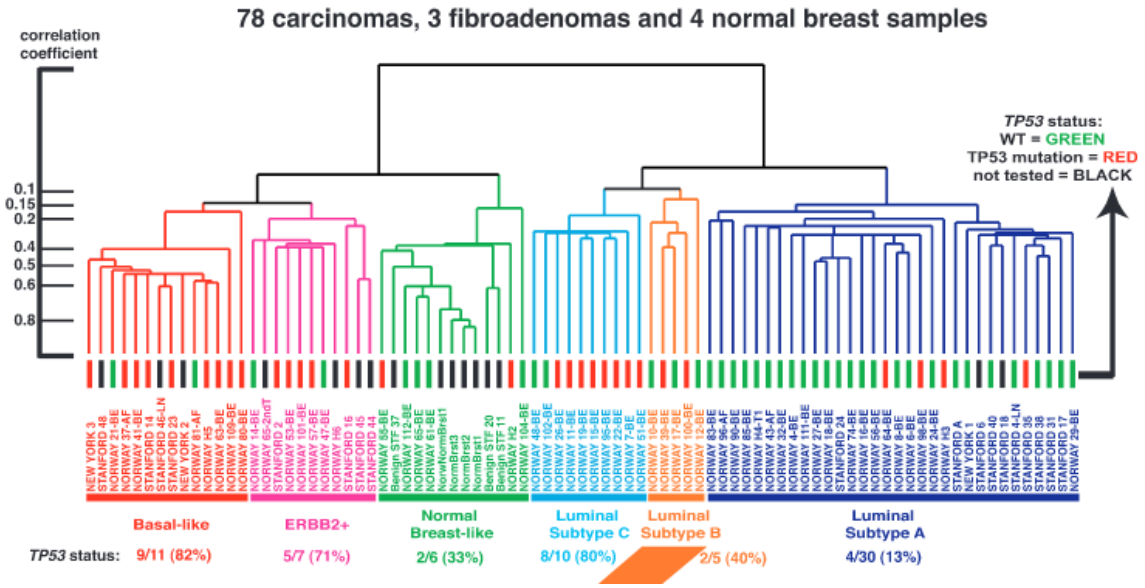

Figure8: Gene expression patterns of experimental samples representing 78 carcinoma clustered in six subtypes

(Sørlie et al. 2001; "Hormone Receptors in Breast Cancer" 1978)

ER and PgR overexpression allows the prediction of hormone therapy (Sørlie et al. 2001; "Hormone Receptors in Breast Cancer" 1978)). Good OS and survival without recurrence are linked to high rates of ER and PgR. The co-expression of these two hormone receptors is a good prognosis factor, especially as their overexpression increases; on the other hand, their total absence is associated with a poor prognosis. HER2 overexpression is correlated with a poor prognosis with recurrence-free survival and reduced OS. However, this overexpression is also predictive of a targeted anti-HER2 therapy response (Gajria and Chandarlapaty 2011). To try to obtain a better risk stratification with an optimization treatment approach (benefit / toxicity ratio), Perou-Sorlie and other authors analyzed the gene expression of mammary tumors by microarray and were able to identify five molecular groups: luminal A, luminal B, Triple Negatives, pure HER2 and the normal-like group. The latter was ruled out because it is artifactual and corresponds to tumors contaminated with healthy breast tissue. However, microarray analyses are not always possible given their cost and the technical difficulties in performing them. To resolve this problem, several authors have demonstrated that immunohistochemistry can serve as a surrogate for the microarray

to define the intrinsic classification of molecular subtypes. Thus,(Carey et al. 2006) and other authors ((Cheang et al. 2015; Fakhri et al. 2018); (Allison et al. 2020); (Weigel and Dowsett 2010) reproduced by immunohistochemistry the protein expression of mammary tumors, based on the ER, PgR, HER2 and Ki-67 (Fitzgibbons et al. 2014); (Nielsen et al. 2020); (Cserni et al. 2006); (Zaha 2014).

b) Intrinsic subtype classification

An analysis of several gene clusters that vary between breast tumors, revealed the presence of different major BC subtypes (luminal A, luminal B, HER2-enriched, basal-like and a normal breast-like group). Other much rarer subtypes such as claudin-low have also been identified, this latter subgroup predominantly contains triple-negative tumors and has the worst prognosis. And because high-throughput transcriptomic analysis is expensive and by no means widespread, a classification based on the above-mentioned immunohistochemical biomarkers was further developed, classifying tumors into the five subtypes aforementioned and summarized in *table4*.

Table 4: BC molecular classification based on ER, PgR, HER2 and Ki-67 immunohistochemical staining status

IBC subgroups	ER and PgR status	HER2 status	Ki-67 status
Luminal A	ER+ and/or PgR+	Negative	Low
Luminal B (HER2+)	ER+ and/or PgR+	Positive	High
Luminal B (HER2-)	ER+ and/or PgR+	Negative	High
HER2	ER- and PgR-	Positive	Any
Triple Negative	ER- and PgR-	Negative	Any

This classification has direct consequences on therapy. The groups with HR+ and/or HER2+ can benefit from hormone therapy and/or anti-HER2 treatment. In contrast, the HR-/HER2-group, also called "Triple Negative Breast Carcinoma” or “TNBC" because ER-/PgR-/HER2- tumors currently does not benefit from any targeted therapy (except very recently from immunotherapy). Triple negative tumors represent 15-20% of breast tumors and are therefore highly aggressive. The group of luminal B tumors is distinguished from the luminal A group by a higher proliferation index. Several studies have subsequently reinforced this classification into molecular subgroups as well as

the correlation between subgroups and relapse-free survival time (Yersal & Barutca, 2014). The different detailed molecular subgroups are as follows:

Luminal phenotype: constitutes approximately 75% of BC and comprises:

- *Luminal A subtype:* the most common, represents 50 to 60% of all cancers. It has a favorable prognosis and is characterized by a low histologic grade, low p53 mutation rate, HER2 under-expression, high level of hormone receptor expression and epithelial lumen cytokeratin 8 & 18 overexpression.

-*Luminal B subtype*: accounts for 15 to 20% of BCs. It has the same characteristic regarding hormonal receptors (ER + and/or PgR +) but recognizes a more aggressive phenotype and a less favorable prognosis than luminal A, since its tumors are usually of high histological grade. The major difference between these two subtypes is the greater expression of proliferation genes (*Myb*, *CCNE1*, etc.) than in luminal A (Dai et al. 2016). However, this aspect may prove to be difficult to use in the clinical routine, because the proliferation gene expression forms a continuum between luminal A and B and it is difficult to choose where to place the threshold. Currently, the Ki-67 index is proposed as a proliferation marker that would differentiate luminal A tumors from B, with a cut-off of 20% although there is an international dissensus between the 14% and 20% expression thresholds. According to a large study conducted by Butreo and al, in terms of Disease-Free Interval (DFI) and Disease-Free Survival (DFS), patients bearing tumors with Ki-67 <14% did not differ from those with Ki-67 values between 14% and 20%. On the other hand, patients with Ki-67 >20% tumors showed the poorest prognosis. Several other studies have shown that 20% Ki-67 cut-off is the best to stratify high-risk patients in luminal BCs, and suggest to integrate it as a prognostic factor (Inic et al. 2014), (Ahn et al. 2015).

Pure HER2 phenotype: As their name suggests, these tumors highly overexpress HER2, and account for 10-20% of BCs. This profile defines a group of tumors with a much greater proliferative power than those of the previously mentioned subgroups, characterized by a high histological grade, poor prognosis and TP53 mutations in more than two thirds of cases. The latter is also associated with SBR II and III grades.

Triple negative phenotype: This phenotype is included in the basal-like molecular subgroup that encompasses triple negative (TN) and other rarer groups. It is included in the global basal-like tumors. The latters are generally ER- and PgR-negative, and lack HER2 overexpression. The most common histological type of BLBC is NST invasive carcinoma, but it also involves some other histological types including invasive lobular, medullary, metaplastic, myoepithelial, apocrine, neuroendocrine, adenoid cystic, and secretory breast carcinoma. The Basal-like tumors subtype is characterized by a particularly high histological grade, high mitotic index and high rate of metastases, especially to the brain and lungs. They are often associated with *BRCA1* dysfunction (Montagna et al. 2013). The origin of the "basal" term comes from 1980s scientific works, which designated cells present in certain stratified epitheliomas, and which mainly express the high molecular weight cytokeratin, characteristic of the basal like type, CK14+, CK17+ and CK5+, as well as the epidermal growth factor receptor EGFR. In over 80% of cases, TP53 is mutated. These cells are in the "basal" position just in contact with the basal membrane of myoepithelial cells. This form represents 15 to 20% of BCs.

B) Integrative cluster classification:

This classification made it possible to divide breast tumors into 10 integrative, well-differentiated clusters based on genomics and transcriptomic data. Each subgroup has different copy-number aberrations, different clinical outcomes and responses to treatment. Six clusters (1, 2, 3, 6, 7 and 8) mainly include ER-positive tumors and PAM50 luminal A and luminal B subtypes, but with distinct genomic alterations. Cluster 10 consists mainly of ER-negative tumors, with instability and the worst prognosis of all *(Figure9)*. Cluster 4 mainly comprises tumors with fairly extensive intra-tumor lymphocyte infiltration (Russnes et al. 2017), (WHO classification of tumors series, 5th ed. vol. 2.2019).

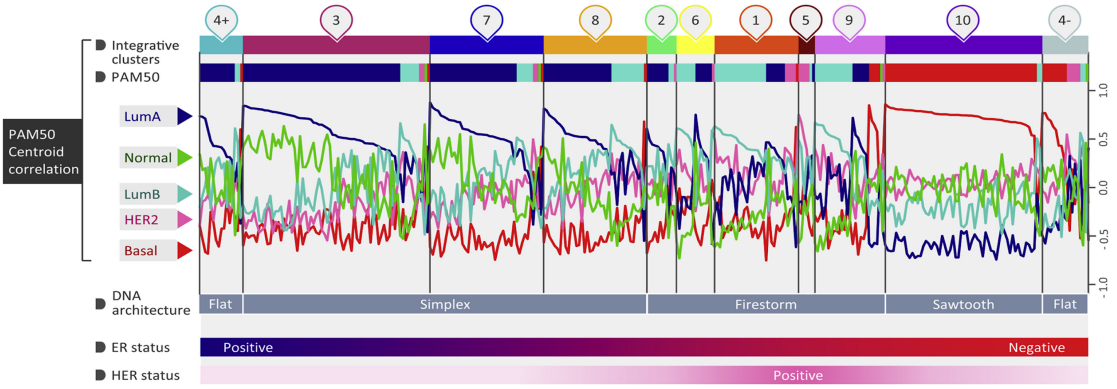

Figure 9: integrative clusters and PAM50 subtypes comparison

An indication of the type of DNA architectural pattern frequently found within each integrative cluster subtypes is given below as is an illustration of which subgroup is dominated by estrogen receptor (ER) and/or HER2 positivity (ER+ is blue, ER– is red, HER2+ is purple). Basal, basal-like; HER2, HER2-enriched; LumA, luminal A; LumB, luminal B; Normal, normal-like.(Russnes et al. 2017)

C) Breast cancer classification problematic:

As already mentioned, in practice, the therapeutic indications for BC are based on histological, clinical and molecular prognostic factors. The latters do not fully explain all the evolutionary heterogeneity of the disease. They may sometimes result in unsuitable, toxic, unnecessary or ineffective therapies.

Given the increasing availability of new anti-tumor molecules, it is crucial to improve BC classification prognostic in order to refine therapeutic indications and improve patient survival, through a more detailed and objective molecular characterization of the disease.

Indeed, the numerous and complex BC molecular alterations; the polygenic and multifactorial genetic changes, confer to each tumor its own phenotype and evolutionary potential. These profiles have allowed the emergence of multiple classifications (histopathological or molecular) and allowed to precisely identify subgroups according to their intrinsic tumor properties. Those properties are called molecular “signatures” or “expression profiles”, they allow a better definition of an individual prognosis and a better therapeutic indication discrimination.

The precise characterization of molecular alterations in BCs leads to new therapeutic targets discovery, and to successfully implement targeted and individualized therapies. On the other hand, parallel to the description of the new BC “molecular taxonomy”, the use of expression genes signatures

for prognostic purposes is developing in a major way. The hypothesis emitted was that multigene signatures would predict the relapse better than the clinicopathologic criteria used in daily practice.

A.2.3. Triple Negative Breast Cancer

A) Generalities

TNBC does not express any of the three receptors commonly found on BC cells: ER, PgR and HER2. This suggests that the cancer's growth is not fueled by estrogen and progesterone hormones, nor by HER2, making its treatment different from the other molecular subgroups.

TNBC are high-grade ductal carcinomas with a high Ki-67 mitotic index and numerous nuclear atypia. The tumor cells also usually display a solid growth pattern, frequently pushing borders, and geographical necrosis.

They are often related to the basal subtype and show similarities to cancers developed on germline BRCA mutation. Their grade is usually high and with a more aggressive profile: strong p53 expression. Which confers to this particular molecular subgroup a higher risk of metastatic recurrence in the first three years after diagnosis.

B) TNBC subtypes classifications:

The TNBC molecular subgroup shows some differences compared to other subgroups, such as:

-Premature relapse / Poor prognosis / High visceral metastasis;

-Short relapse free survival (RFS) and short median time to death;

-Larger tumor size with earlier lymph node involvement;

-The loss/gain of function including genes associated with repair of DNA damage and PI3K signaling, TP53, RB1, BRCA1;

TNBC is a very heterogeneous malignancy at the morphological level, for which a furthermore deep subclassification is necessary. This molecular subtype is a subject of extreme importance both in the field of basic scientific research and in clinical practice, for many reasons: poor prognosis compared to non-TNBC tumors; the absence of a specific and efficient targeted therapy.

In a meta-analytical study, Chiu & al. were able to demonstrate that the heterogeneous types analysis of datasets with cross-platform technologies resulted in a fine partitioning into very distinct clusters compared with partitioning established on a single dataset. TNBCs are therefore very

heterogeneous, depending on their original cells, mutational signatures, survival, genetic variations, tumor histology, and clinical phenotype. This heterogeneity is correlated with more pronounced tumor size, SBR grade, mitotic score and metastasis. Other more recent studies on gene expression profiles have shown that among TNBCs there are also several so-called “molecular” subtypes with different sensitivity to treatments and different prognosis. They express different molecules that could serve as therapeutic targets specific to each subtype. Thus, this inherent heterogeneity can be used to prioritize patients for precision medicine.

a) TNBC subtypes by (Lehmann et al. 2011):

High throughput transcriptomic studies have recently revealed several subgroups of TNBC, with different gene profiles, biology and sensitivity to treatments. The team from Vanderbilt University (United States), supervised by J. Pietenpol and B. Lehmann analyzed the transcriptomic profiles of 386 CSTNs, strictly selected for the absence of ER, PgR and HER2 gene expressions, as well as ERBB2 / HER2 amplification. They described the existence of 6 TNBC subtypes. Thy are detectable by a specific software on an Affymetrix expression profile, and have been called: basal-like 1 and 2 (BL1, BL2), mesenchymal-like (M), mesenchymal stem-like (MSL), immunomodulatory (IM) and luminal androgen receptor (LAR).

The BL1 subtype is characterized by genomic signatures such as DNA damage response (DDR) and cell cycle, while BL2 shares the same cell cycle genes with BL1, but is not enriched with DDR genes. BL2 is characterized by overexpressed genes of the myoepithelial growth factor receptor family signaling pathway.

The M and MSL subtypes are both enriched with genes regulating cell motility, invasion and mesenchymal differentiation, but the MSL subtype is the only enriched one with EMT regulatory genes and cancer stem cells. In addition, the MSL subtype shares with the IM subtype many genes involved in the regulation of the immune response. However, the IM subtype is enriched with genes responsible for interactions between the host and the cancer, i.e., immune antigens and genes involved in immune signal transduction pathways. Finally, the LAR subtype is characterized by the overexpression of genes coding for luminal differentiation (Lehmann et al. 2011).

Table 5: Lehman's TNBC subtypes classification:

Basal-like type		Mesenchymal		Luminal Androgen Receptor (LAR)	ImmunoModulatory (IM)
BL1	**BL2**	**M**	**MSL**		
Enhanced cell cycle/division	Growth factors signaling alteration	Growth factors signaling		High Androgen receptor expression	Immune cell markers enriched
Altered transcription of genes involved in DNA damage response pathways	Myoepithelial markers	Enhanced Epithelial to mesenchymal transition pathways (EMT)		High luminal gene (estrogen-regulated genes (PgR, GATA5)) expression	Tumors with> 50% lymphocyte infiltrate

b) The alternative classification by (Y.-R. Liu et al. 2016; Burstein et al. 2015)

Two very recent genomic studies have opened the possibility of other TNBC molecular classifications. After profiling 198 TNBC DNA and RNA records. (Burstein et al. 2015) and (Liu & al, 2016) distinguished 4 molecular subtypes, identified by specific amplifications: basal-like immune-activated (BLIA), basal-like immune-suppressed (BLIS), mesenchymal (MES) and LAR.

Table 6: Burstein's TNBC alternative classification:

Basal like immunosuppressed	Basal like immune activated	MES	LAR
Equivalent of BL1 and BL2 in Lehman's classification B, T and NK cells under-expressed Have the worst prognosis	Equivalent of IM subtype in Lehman's classification	Equivalent of Lehman's Mesenchymal subtype Expression growth factors IGF-1 genes	Equivalent of Lehman's LAR subtype

c) FUSCC classification by (Y.-R. Liu et al. 2016)

Table7: Liu's FUSCC TNBC classification

Immunomodulatory (IM)	luminal androgen receptor (LAR)	Mesenchymal-like (MES)	basal-like and immune suppressed (BLIS)
Cytokine-cytokine receptor interaction↑ T/ B cell receptor signaling pathway ↑ Chemokine and NF-kappa B signaling pathway ↑	Steroid hormone biosynthesis ↑ Androgen and estrogen metabolism ↑	ECM-receptor interaction ↑ Focal adhesion ↑ TGF-beta signaling pathway ↑ Adipocytokine signaling pathway ↑	Mitotic cell cycle↑ DNA replication↑ DNA repair↑ Immune response↓ Innate immune response ↓ T cell receptor signaling ↓

Data from (Y.-R. Liu et al. 2016)

Abbreviations: FUSCC Fudan University Shanghai Cancer Center, IM : immunomodulatory, LAR : luminal androgen receptor, MES : mesenchymal-like, BLIS : basal-like and immune suppressed, ECM : extracellular matrix, TGF : transforming growth factor

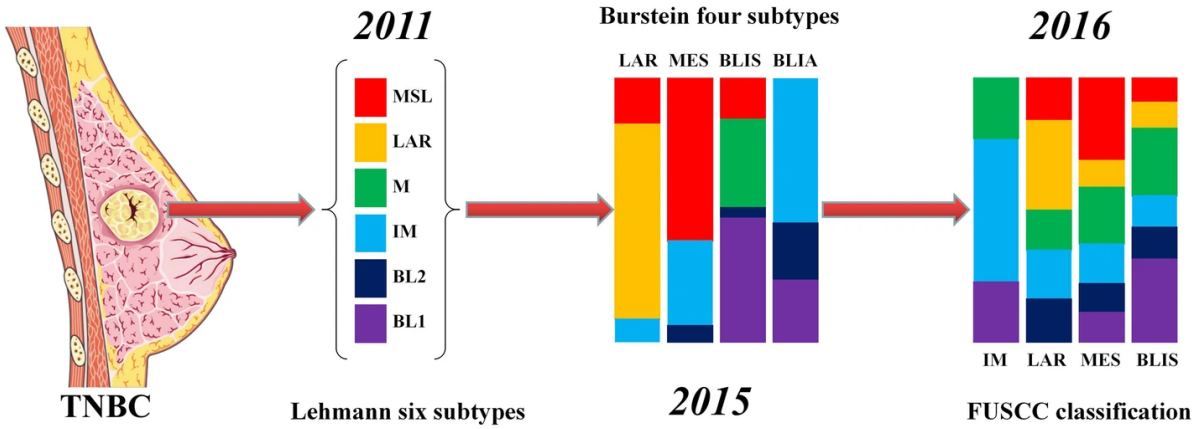

Figure 10: Progress in classification of TNBC subtypes, and interaction analysis of the Burstein four subtypes/FUSCC classification and Lehmann six subtypes, rectangle size varies in proportion to the number of samples (Yin et al. 2020)

C) TNBC vs. BLBC:

The names BLBC and TNBC are often considered synonymous, but they are not completely so *(Figure11)*. Clinical, transcriptomic and immunohistochemistry data have shown that although there is a large overlap between TNBC and BLBC tumors, it is not integral: more than 20% of basal-like tumors are not TNBCs. BLBC and TNBC cancers are therefore distinct subtypes. In addition, 60 to 70% of BC in patients presenting a familial *BRCA1* gene mutation have a morphological appearance of the BLBC or TNBC type (Foulkes et al., 2010; Kreike et al., 2007; Oakman et al., 2010; Stover et al., 2016).

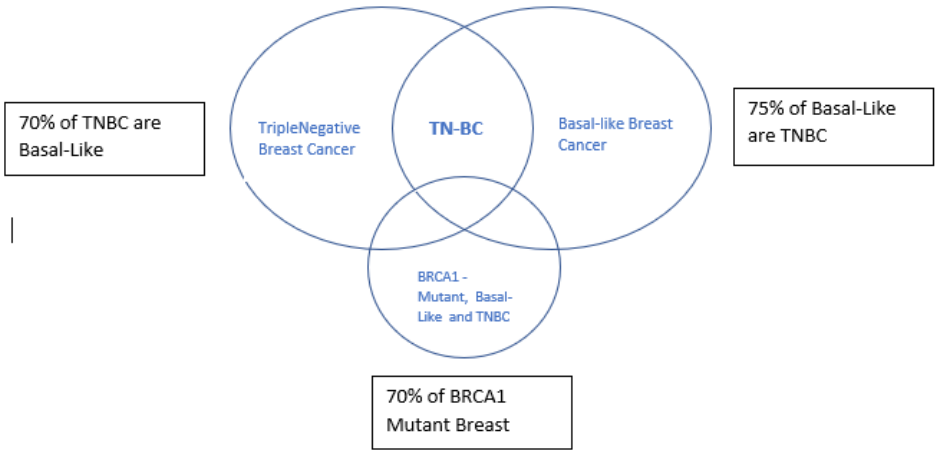

Figure 11: overlap between breast cancers TNBC, BLBC and the mutated BRCA pathway

BLBCs, unlike TNBCs, particularly express biomarkers often found in patients with hereditary BC linked to *BRCA1* mutation. The most frequent mutations of the BLBC subgroup concern P53 and PI3K, other mutations may also be considered but are less frequently found (Figure12).

It should be noted that BLBC is called so because its gene expression profile is similar to the basal-myoepithelial layer of healthy tissue. The latter overlaps with TNBC because it is usually negative for both hormonal and HER2 receptors. Although these two subtypes show a strong clinical and biological correlation, both definitions are far from being synonymous. On average, we can conclude that 75% of TNBCs are BLBC. In a study by Prat et al, they performed a PAM50 systematic analysis of BLBC and TNBC. 21% of TNBC weren't profiled as BLBC, while on the other hand, 31% of BLBC weren't profiled as TNBC (Prat et al.2013). Therefore, equating TNBC with BLBC cannot be fully trusted. This incomplete overlap requires much more analysis to refine these subgroups identification accuracy. But although they are not the same entity, TNBC replaces BLBC in terms of diagnosis and treatment because its immunohistochemical study is more feasible, accessible and rapid than gene expression signature examination (Yao et al. 2017).

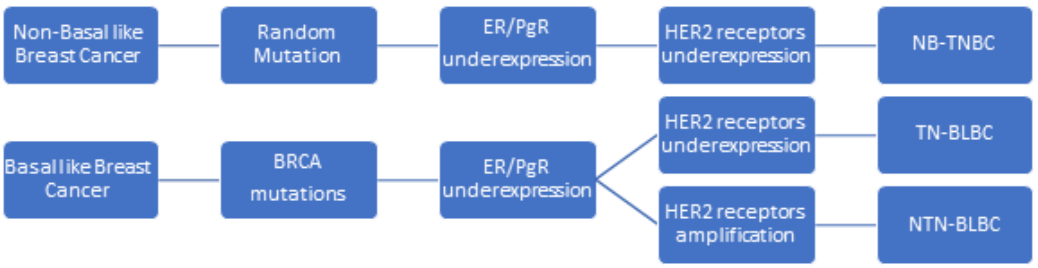

Figure12: Points of convergence / divergence between BLBC and TNBC

(NTN-BLBC: Non-Triple Negative Basal-Like Breast Cancer; TN-BLBC: Triple Negative Basal-Like Breast Cancer; NB-TNBC: Non-Basal-Like Triple Negative Breast Cancer.) (Yao et al. 2017)

D) Prognosis

The prognosis worsens with increasing SBR grade in specific morphological subgroups. TNBCs have an unfavorable prognosis with an OS which is usually less than 30%, despite the increased sensitivity to the current cytotoxic therapies. TNBC tumors owe their aggressive biological behavior to the predominance of Grade III, high proliferation index, central nervous system, and visceral metastases.

Rare are the TNBC tumors with favorable prognosis: these include adenoid cystic carcinomas even with a high proliferation index, which usually have a good prognosis.

Many efforts have been made by the scientific community to classify TNBCs in distinct subgroups, since this class of lesions is very heterogeneous. However, no standardized classification has yet

been reached. This is a clinical unmet need that may allow better personalization of treatments, therefore survival improvement.

B. STATISTICAL OVERVIEW

1. Machine learning:

Machine Learning (ML) is a field of artificial intelligence, which explores how algorithms can learn by studying examples. From a data frame with several observations (samples, individuals, etc.), the algorithm tries to discover a link in this data that allows it to generalize predictions. In other words, input (X) and output (Y) data are presented to an algorithm which parameters the mathematical equation that defines the link between X and Y until it has understood the latter (learning phase). Once it is achieved, the prediction phase starts, which is the ultimate goal of a ML algorithm. In real ML problems, there is usually more than one data as an input (Badillo et al. 2020). Here is an example:

Image recognition: It consists in making an algorithm that takes an image as input data and aims to guess what the image represents as an output. This therefore requires an algorithm that can take several data as input (thousands of pixels) and above all to capture very complex relationships between input and output data. This is where neural networks come in.

A neuron is a mathematical function relating inputs X with outputs Y. It is important to clarify that what we are talking about here is an artificial neuron, that is roughly mimicking the functioning of a real neuron. True biological neurons are cells of the nervous system which are connected to each other, each neuron has an extension called an axon through which the neuron can send a signal to other neurons.

The way the neuron works is as follows, whether or not it receives an electrical signal from other neurons connected to it, based on which it does a fairly binary thing. Either it does not send anything in its axon or it sends an electrical signal, and in this case, we say that it discharges. The idea of the artificial neuron, which dates back several decades, is to mimic this behavior by a mathematical function whose principle is as follows (Figure13).

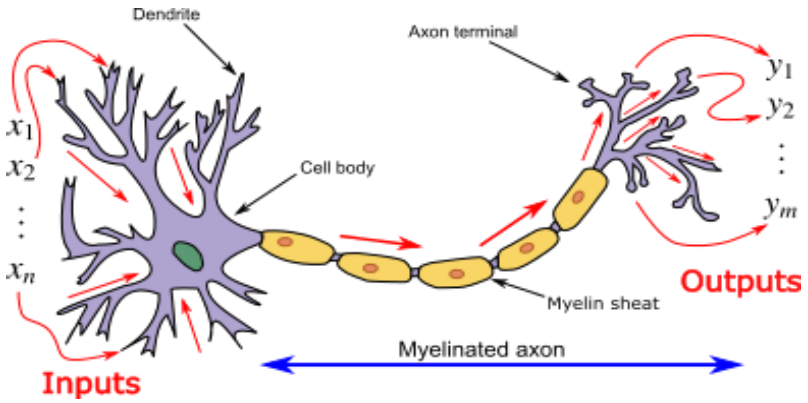

Figure 13: Neuron and myelinated axon, with signal flow from inputs at dendrites to outputs at terminal axon

Let's imagine an artificial neuron with 3 inputs called (X1, X2 and X3); we add these 3 inputs by assigning a coefficient to each, that we call "weight". If the sum obtained is greater than a certain threshold, the neuron will send 1 on output and otherwise it will send 0.

An artificial neuron is a mathematical function which takes X as input and which outputs a Y and this function has "knobs" that can be turned to set "weights" and "thresholds". The problem is that a neuron on its own is not enough to make very complicated relations, but what is interesting is that several neurons can be stacked to accomplish very complicated functions, which is why they are called neural networks. By stacking neurons, we can make functions as complicated as we want with lots of inputs and outputs and especially with lots of weights and thresholds to set.

Their advantages are that they are very versatile and can be adapted to several types of inputs and outputs. We take a neural network, we present it with an input and output database, and it defines the thresholds, until it weighs correctly between the input and output data. We recall that once the learning phase is done, the network is trained, which makes it capable of predicting the output if it is presented with new unseen input data. This is the prediction phase. Besides, it seems that what goes on in artificial neural networks (also called Multilayer Perceptron) looks a lot like what is going on in our brains; when we learn new things, the strength of the connections between our neurons changes, which is called "synaptic efficiency", and this is done in a way that we can compare to the way we "play" with the weights in the artificial neural network.

a) Clustering techniques in Machine Learning:

Clustering is a descriptive technique of data analysis in ML, which is generally used when we have a large volume of data within which we seek to distinguish homogeneous subgroups (clusters), which have two important features:

-The subgroups are not predefined by the analyst.

-Subgroups group together objects with similar characteristics (internal homogeneity) and separate objects with different characteristics (external heterogeneity), which can be measured by interclass inertia.

In the medical sector, classification makes it possible to determine groups of patients likely to be subjected to different treatment protocols, each subgroup comprising all the patients behaving similarly with respect to a specific measured variable.

The advantage of using clustering techniques is firstly because most predictive algorithms do not cope well with too many variables, due to the correlations that exist between them and which disrupts their predictive power. However, it is difficult to define a heterogeneous population by a small limited number of variables. The advantage of classification is therefore to create fairly homogeneous subgroups that can be described by a small number of variables specific to each subgroup.

Therefore, the aim of automatic classification would be to minimize the intra-class inertia, to a number of K fixed classes.

Broadly speaking, clustering can be divided into two distinct categories:

-*Hard Clustering:* In hard clustering, each data point either belongs to a cluster completely or not. For example, each patient is put into one cluster out of the other clusters.

-*Soft Clustering:* In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, each patient is assigned a probability to be in each cluster.

There is currently a plethora of clustering algorithms, so deciding which clustering method to use can be a very crucial task for the statistician. Here, we will briefly develop the most famous clustering methods used in this research project:

a.1) Estimation-Maximization (EM) Clustering:

General principle: The EM algorithm is an iterative data partitioning algorithm originally invented in 1977 which allows to find the maximum likelihood parameters of a probabilistic model when thz latter depends on unobservable latent variables (Dempster A.P et al.1977). It is therefore an iterative method which attempts to maximize the likelihood of the target probability in two steps.

Objective: The objective of the EM classification will consist in estimating the means and standard deviations of each class, so as to maximize the likelihood of the observed data (distribution). In other words, the EM algorithm will seek to approximate the observed distributions of the values on the basis of combinations of different distributions in the different classes.

Advantages:

-Variables types: EM clustering allows categorical variables to be processed in addition to other data types, unlike other algorithms.

-Classification probabilities instead of classifications: The results of the EM classification are different from those produced by other classification methods: the latter will assign the observations to the classes while seeking to maximize the distances between the classes. The EM algorithm does not calculate the real assignments of cases to classes, but classification probabilities. In other words, each observation belongs to a class with a certain probability. Lastly, the assignment of cases to classes is addressed, based on the (highest) classification probabilities.

-K-Cross Validation: in practice, the analyst generally does not know in advance the number of classes that he/she will be able to identify in the dataset. This is the reason why the program integrates a cross validation algorithm by k-sets (k-fold) in order to automatically determine the number of classes in the data. Adapted to allow structure detection. The general idea of this method is to divide the overall sample into a number “k” of sets, or to draw subsamples at random. The same analysis is then applied successively to all the observations of the k-1 sets (training samples), and the results of the analyses are then applied to the sample k (sample or set that was not used to estimate the parameters, build the decision tree, determine the classes, etc; this is more precisely the test sample) in order to calculate an index of predictive validity. The replications results are then

aggregated (the average is determined) in order to produce a synthetic measure of the respective model robustness, or the model's validity in predicting new unseen observations.

Disadvantages:

This method has the disadvantage of being so greedy in computing time and memory resources. The EM algorithm does not rely on distances. On the other hand, the program will calculate probabilities for each observation belonging to each of the classes according to the chosen distribution (by default, the normal distribution); the objective of the EM classification algorithm then consists in finding the classification solutions which will maximize the overall probability of the data, given the final classification solution with the desired number of classes.

a.2) K-means clustering:

It is an iterative clustering algorithm that aims to find local maxima in each iteration. It makes it possible to analyze a set of data characterized by a set of descriptors, in order to group "similar" data into groups (or clusters). It consists in five different steps:

1. Specify the desired number of clusters K
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat steps 4 and 5 until no improvements are possible: Similarly, 4th and 5th steps will be repeated until it reaches global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm.

The similarity between two data can be inferred from the "distance" between their descriptors; thus, two very similar data are two data whose descriptors are very similar. This definition makes it possible to formulate the data partitioning problem as the search for K "prototype data", around which the other data can be grouped.

These prototype data are called centroids; in practice, the algorithm associates each data with its closest centroid, in order to create clusters. On the other hand, descriptors means define their centroid position.

Like any algorithm, K-means has advantages and disadvantages: it is simple, fast and easy to understand; however, it requires that the k number of clusters should be predefined in advance.

a.3) Hierarchical clustering:

Hierarchical clustering is an algorithm that builds hierarchy of clusters.

In the case of agglomerative (or bottom-up) clustering, we start by considering that each point is a cluster on its own. Then, we find the two closest clusters, and we aggregate them into a single cluster. This step is repeated until all the points belong to a single cluster, made up of the agglomeration of all the initial clusters.

The opposite approach, divisive (or top-down) clustering, consists of initializing with a single cluster containing all the points, then iteratively separating each cluster into several, until each point belongs to its own cluster.

Therefore, the hierarchical clustering can be summarized as follows, it requires:

-a distance (Euclidean distance: if the variables are numerical), or (Gower distance: if the variables are of mixed types “Nominal; categorical; ordinal…”). It helps in calculating the dissimilarity matrix: Dissimilarity matrix is a mathematical expression of how different, or distant, the points in a data set are from each other, which is a core idea of clustering.

Hierarchical clustering properties:

-can be quite unstable.

-may depend a lot on the distance and the aggregation method.

-sensitive to changes in the scale of one of the variables.

-difficult to know at what height to cut to determine the clusters.

-but on the other hand, it is deterministic.

a.4) PAM clustering:

It is similar to K-means, but is considered more robust because it admits the use of other dissimilarities besides Euclidean distance.

Objective: We define k objects representative of the classes, called medoids, located at the classes center. The medoid is the object for which the average dissimilarity with respect to other objects of the class is the weakest.

Principle:

Step 1:

A sample is taken randomly from the set.

The PAM algorithm is used to determine the k medoids.

The other objects are then assigned to the nearest medoids.

The partition is characterized by the value of the corresponding objective function.

2nd step:

We repeat step 1 several times and we retain the medoids and the corresponding partition.

The structure of the obtained classes: Three cases are possible:

The first case: two classes are always disjoint as is the case with partitioning methods. The number of classes is generally defined a priori but some methods can be freed from this constraint.

The second case: two classes are disjoint or one contains the other. These are the ascending hierarchical methods called "agglomerative" or descending and called "divisive", and are generally based on a notion of distance.

The third case: two classes can have several objects in common, called "encroaching" classes and we speak of fuzzy analysis or fuzzy clustering. This method assigns each object a probability of belonging to a given class.

b) The optimal number of classes:

The definition of natural classes is very delicate because the results do not always appear obvious and may differ depending on the chosen algorithm. Certain methods such as that of mobile centres and its variants require this number to be fixed a priori, which obviously greatly affects the

classification's quality when this number does not correspond to the actual distribution of individuals. Other methods on the other hand, such as classification by aggregation of similarities, let the algorithm automatically determine the optimum number of classes.

b.1) Optimal k-number of clusters determination:

Choosing k-number of clusters is not necessarily intuitive. Especially when the dataset is large and we don't have a priori partition or assumptions about it. A large k-number can lead to overly fragmented data partitioning. This will prevent the discovery of interesting patterns within the data. On the other hand, a too small number of clusters, will lead to having, potentially, too generalistic clusters containing a lot of data. In this case, we won't have any "fine" patterns to discover.

For the same dataset, there is not a single clustering possibility. The difficulty will therefore lie in choosing a k-number of clusters which will make it possible to highlight interesting patterns between the data. Unfortunately, there is no automated process for finding the right number of clusters. It therefore remains subjective and depends on the method used to measure the similarities for the partitioning. One of the most common and simple ways is to inspect the dendrogram produced by the hierarchical clustering method, but this too remains subjective as well.

Another common way to choose the number of clusters is to run k-Means with different values of k and calculate the different clusters variance. The variance is the distances sum between each centroid in a cluster and the different observations included in the same cluster. Thus, we seek to find a number of clusters k such that the selected clusters minimize the distance between their centers (centroids) and the observations in the same cluster. We are talking about minimizing the intra-class distance.

The principle is essentially based on the fact that the more homogeneous the data inside the classes, the smaller their distances from the point representing the class. Therefore, a low value for intra-class inertia describes homogeneity of data within classes. The more heterogeneous the classes are, the greater the distances between the points representing the profiles of the classes. Therefore, a high value of the interclass inertia reflects heterogeneity between the classes. This index has the defect of increasing when we increase the number of classes.

We will therefore briefly discuss the most commonly used methods to determine the optimal number of clusters. These methods consist in optimizing this criterion and are summarized below:

Quality indices:

Several validation criteria have been defined and proposed in the literature. Indeed, in addition to the comparative study by Milligan and Cooper (1985) which examined 30 indices on simulated data, new indices have been proposed in (Dunn, 1974), (Rousseeuw, 1987), (Kauf-man et Rousseeuw, 1990), (Tibshirani et al., 2001), (Lebart et al., 2000), (Halkidi et al., 2000) and (Halkidi and Vazirgiannis, 2001). But in this research work we will rely on the most recent method found in the "NBClust" package (Charrad et al., 2012) (Charrad et al., 2014), the objective of which is to gather all the indices implemented in R bookstores in the same library, and to add the indices that are not yet implemented. This provides the user with an exhaustive list of validation indices allowing him to estimate the correct number of classes in a dataset.

We will therefore briefly discuss the most commonly used methods to determine the optimal number of clusters. These methods consist in optimizing a criterion, such as the sums of squares within the clusters. These methods are summarized in 30 quality indices.

Validation Measures:

Various measures exist and are mainly aimed at validating the clustering analysis results, and determining which clustering algorithm works best for a particular experiment have been proposed (Kerr and Churchill, 2001; Yeung et al., 2001; Datta and Datta , 2003). This validation can be based only on the internal properties of the data or on an external reference, and on expression data alone or in conjunction with relevant biological information (Gibbons and Roth, 2002; Gat-Viks et al., 2003; Bolshakova et al., 2003; Bolshakova et al., 2003; Bolshakova et al., 2005; Datta and Datta, 2006).

Another related problem is determining the most appropriate number of clusters for the data. Ideally, the resulting clusters should not only have good statistical properties (compact, well separated, connected and stable), but also give biologically relevant results (Brock et al., March 2008).

-Internal validation:

The internal validation can be measured by the compactness, connectedness, and separation of the cluster partitions. Connectedness relates to what extent observations are placed in the same cluster as their nearest neighbors in the data space, and is here measured by the connectivity (Handl et al., 2005). Compactness assesses cluster homogeneity, usually by looking at the intra-cluster variance, while separation quantifies the degree of separation between clusters (usually by measuring the distance between cluster centroids). Since compactness and separation demonstrate opposing trends (compactness increases with the number of clusters but separation decreases), popular methods combine the two measures into a single score. The Dunn Index (Dunn, 1974) and Silhouette Width (Rousseeuw, 1987) are both examples of non-linear combinations of the compactness and separation; and with the connectivity comprise the three most important internal measures available (Handl et al.2005).

-Stability measures:

The stability measures compare the results from clustering based on the full data to clustering based on removing each column, one at a time. These measures work especially well if the data are highly correlated. The included measures are:

1. *Average proportion of non-overlap (APN)*: APN measures the average proportion of individuals who switch classes when performing a classification by removing a variable from the database.
2. *Average distance (AD):* AD measures the average distance between observations placed in the same cluster under both cases (full data set and removal of one variable).
3. *Average distance between means (ADM):* ADM measures the average distance between cluster centers for observations placed in the same cluster under both cases.
4. *Figure of merit (FOM)*: FOM measures the average intra-cluster variance of the deleted variable, where the clustering is based on the remaining (undeleted) variables. (Datta and Datta, 2003; Yeung et al., 2001)

b.2) Interpretation of the obtained clusters

To this day, there is no objective and universal scale for comparing the different classifications. It is, however, possible to compare any predictive ranking model by measuring the rate of well-classified observations.

But in general, a good classification:

-Detects the structures present in the data.

-Allows to determine the optimal number of classes.

-Provides well-differentiated classes (External heterogeneity).

-Provides stable classes against slight data changes (Internal homogeneity).

-Effectively process large volumes of data.

c) Prediction techniques:

Predictive statistical and data mining techniques are often used, whether in the clinics, to calculate a disease occurrence probability for example. Among the techniques, we find two main operations: classification (or discrimination) and prediction (or regression). They both aim to estimate a variable's value (called the variable to be explained) according to other variables values, indicated as explanatory variables.

Classification is therefore the operation which makes it possible to place each individual of the studied population in a class, among several predefined classes. The assignment to a class from explanatory characteristics is usually done by a formula, an algorithm or a set of rules which constitutes a model.

The prediction, on the other hand, has the goal of producing models that generalize, and are able to make good predictions on new data. It also tries to develop a model which is sufficiently complex to capture the data's nature (and thus avoid under-learning), but simple enough to avoid over-learning. (Figure14)

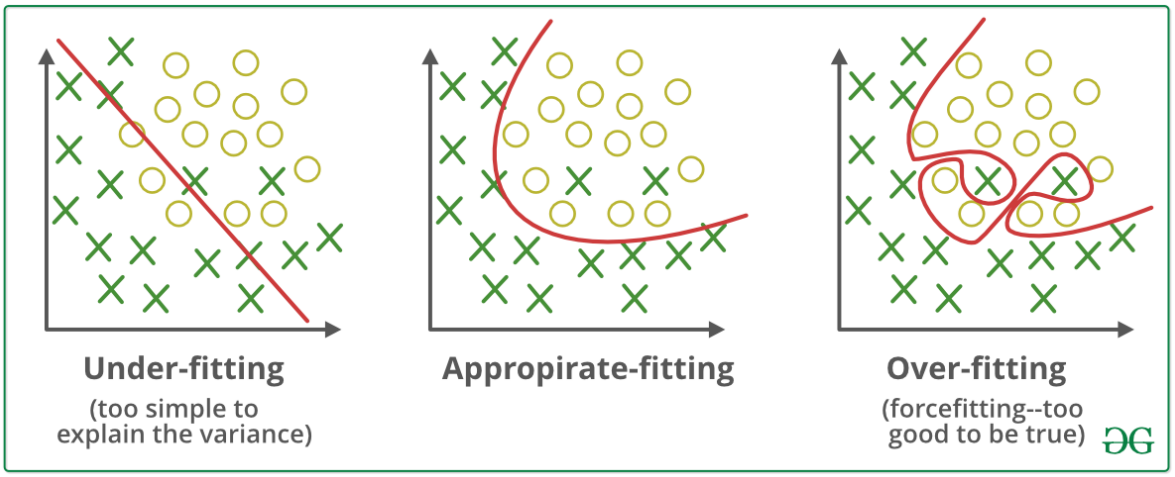

Figure14: Underfitting and Overfitting in Machine Learning

(©GeeksForGeeks)

There are two types of prediction techniques:

-*Transductive techniques:* they only include a single step (possibly reiterated), during which each individual is directly predicted; there is no development of a model, a fortiori, there is no determination of parameters.

-*Inductive techniques*: they consist of a learning phase (inductive phase) which makes it possible to develop a model that summarizes relationships between variables and which can be applied to new data to deduce a prediction.

The latter takes place in three or four stages, which can be summarized as follows:

-A learning step: carried out with a sample of individuals whose predictions we know and who are randomly drawn from the population to be modeled.

-A test step: to verify the model obtained by training on another sample of individuals whose prediction is known and who are drawn randomly from the same population as the training sample. This step makes it possible to select the best of the models developed in the training step by avoiding the bias that would cause the test on the same sample as the training.

-A validation step: on a third sample whose ranking is known to measure the performance of the best model selected in the two previous steps. This step aims to predict the quality of the results that will be obtained during the application. It takes place deliberately on a sample which did not participate in the learning process, or can be carried out on another following sample, in order to check the stability of the predictive power.

-An application step: to apply the model obtained to the entire population.

The objective is to prove that the model generates good estimates of the analyzed variable. To do this, it is necessary to work with at least one training dataset and one validation dataset. Quite simply, the training data is used to calibrate the model while the validation set is used to show that the model is reliable and relevant. To be as objective as possible, the learning and validation datasets should come from an independent population (to be sure not to bias the result). More generally, what is done is to separate a basic initial set of samples into a training set and a validation set. Note that if there is the possibility of acquiring two different datasets, one for calibration and another for model validation, it is recommended to do so. It is often recommended to use between 60% and 80% of the initial dataset as a training set and the remaining 20 to 40% as a validation set. However, these percentages are not fixed.

In this section, we will briefly review the majority of the main predictive ML models that we have used in this research.

All predictive ML models are classified into two categories: supervised or unsupervised.

d) Predictive analytics models in Machine Learning: Supervised Learning

Supervised learning involves teaching a function to match an input to an output based on known examples (input-output pairs). For example, if there is a dataset with two variables, age (input) and height (output), a supervised learning model could be implemented, to predict a person's height based on his age.

d.1) Regression

In regression models, the output is continuous. The idea of linear regression is simply to find a row that best fits (or matches) the data. Extensions to linear regression include multiple linear regression (for example, finding a design that fits best) and polynomial regression (for example, finding a curve that fits best).

-Decision Tree (DT):

This decision support or data exploration tool allows to represent a set of choices in a graphical form of a tree. It is one of the most popular supervised learning methods for data problems. Concretely, a decision tree models a hierarchy of tests to predict a result.

The possible decisions are located at the branches ends (the "leaves" of the tree) and are reached based on decisions made at each stage. A decision tree works by iteratively applying very simple logical rules (typically data separation), each rule being chosen according to the result of the previous rule. Decision trees have the advantage of being easy to interpret, very quick to train, being non-parametric, and requiring very little data preprocessing. This algorithm can also extract logical rules that did not appear in the raw data.

-Random Forests (RF):

It is a model that relies on decision trees during the training phase. The decisions of the majority of trees are the final decision of the random forest. In the growing step at each node, a fixed number of input variables are randomly selected and the best split is calculated only among them, the second selection step is performed so all the forest trees are maximal trees.

The main difference between DT and RF is that a decision tree is a graph that uses a branching method to illustrate each possible outcome of a decision, while a random forest is a set of decision trees that gives the final result based on the results of all its decision trees. When the dataset becomes much larger, a single decision tree is not enough to find the prediction. A random forest which is a collection of decision trees is an alternative to this problem.

The aim of this model is to reduce the risk of individual tree error by relying on a majority (i.e. majority wins) prevalence model.

-Multilayer Perceptron (MP):

Like in biology, the perceptron is a set of neurons organized in layers. From one layer to another, the input signal propagates to the output, activating or inhibiting neurons progressively.

The principle is to look at the output compared to what was expected and to update the connections between neurons (strengthen them or inhibit them) to improve the final result, which will be a prediction from the network.

The perceptron is organized into three layers:

-The input layer = a set of neurons that carry the input signal.

-The hidden layer or more often THE hidden layers (hidden layer 1, hidden layer 2, etc.). This is the heart of the perceptron, where the relationships between the variables are highlighted.

-The output layer: this layer represents the final network's result, its prediction.

-Generalized Linear Model (GLM)

GLM makes it possible to study the link between a dependent variable or response Y and a set of explanatory or predictor variables X1, X2, etc ...

This group includes the log-linear model; logistic regression; Cox-regression; Poisson regression; ... etc. These generalized linear models are mainly formed from three components: The response variable to which a probability law is associated; the explanatory variables used as predictors and form the deterministic component; the link that describes the relationship between the variables and the mathematical expectation of the response variable. These techniques are usually used with categorical response variables.

-Fast Large Margin (FLM):

The Fast Large Margin operator applies a fast margin learner based on the linear support vector learning scheme proposed by R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Although the result is similar to those delivered by classical logistic regression implementations, this linear classifier is able to work on a dataset with millions of observations and attributes.

-Gradient Boosted Trees (GBT):

The sets are built from decision tree models. The trees are added one by one to the set and adjusted to correct for prediction errors made by previous models. This is a type of overall ML model called boosting.

Boosting is an ML algorithm in which weak learners are converted into strong learners. Weak learners are classifiers that always perform slightly better than chance “random”, regardless of the distribution on the training data. In Boosting, predictions are sequential where each subsequent predictor learns from the previous predictors’ errors. Gradient Boosting Trees (GBT) is a commonly used method in this category.

Performance comparison aims to reduce bias and variance - Bagging has many uncorrelated trees in the final model, which helps in reducing variance. Stimulation will reduce variance in the process of building sequential trees. At the same time, its goal remains to close the gap between the actual and predicted values by reducing the residuals, which also reduces the bias.

d.2) Classification:

In classification models, the output is discrete. Here are some of the most common types of classification models.

-Logistic regression (LR):

It is similar to linear regression, but it is used to model the probability of a finite number of outcomes, usually two. A logistic equation is created such that the results values can only be between 0 and 1.

Therefore, this predictive model is used to assess the probability of a certain class or event like alive/dead. It aims to build a model making it possible to predict / explain the values taken by a qualitative target variable (binary) from a set of quantitative or qualitative explanatory variables.

-Deep Learning (DL):

From a set of variable measurements, this algorithm tries to find a deterministic relationship between variables and results, represented by a mathematical equation. Learning simply consists of calculating the connection coefficients (weights) between the different layers so that the outputs of the neural network are, for the examples used, as close as possible to the desired outputs.

-Support Vector Machine (SVM):

This predictive model can get pretty complicated but it is pretty intuitive at the most basic level. The objective of the SVM algorithm is to find the separation between two classes of objects with

the idea that the wider the separation, the more robust the classification. In its simplest form, that of a linear separation and separable classes, the algorithm selects the hyperplane that separates the set of observations into two distinct classes so as to maximize the distance between the hyperplane and the most common observations closer to the learning sample. This distance is also called “margin” and SVMs are thus referred to as “wide margin separators”, the “support vectors” being the data closest to the border.

-Naive Bayes classifiers

The naive Bayesian classification is a type of simple probabilistic classification based on (called naive) independence of the assumptions.

Simply put, a naive Bayesian classifier assumes that the existence of one characteristic for a class is independent of the existence of other characteristics. A fruit can be considered an apple if it is red, rounded, and about ten centimeters. Even if these characteristics are related in reality, a naive Bayesian classifier will determine that the fruit is an apple by independently considering these characteristics of color, shape and size.

Despite their simplicity, they have strengths: they need a small amount of training data and they are very fast compared to other classifiers.

-K-Nearest Neighbors

Once the learning phase has been completed, to make a prediction from a new unknown observation, the algorithm finds the observation that is closest to it in the learning data set. Then, the latter assigns the label of this training data to the new unknown observation.

The k in the formula "k nearest neighbors" means that instead of just the nearest neighbor of the unknown observation, we can consider a fixed number k of neighbors from the training set.

Finally, we can make a prediction based on the majority class in this “neighborhood”.

To measure the proximity between observations, we must impose a similarity function on the algorithm.

This function that calculates the distance between two observations estimates the affinity between the observations like this: “The closer two points are to each other, the more similar they are.

While this algorithm is easy to understand, its main drawback is its cost in terms of complexity. Indeed, the algorithm searches for the k nearest neighbors for each observation. Thus, if the database is very large (a few million rows), the calculation time can become extremely long. Special attention must therefore be paid to the size of the input dataset.

e) Predictive analytics models in Machine Learning: Unsupervised Learning

Unlike supervised learning, unsupervised one is used to draw conclusions and find trends from input data without labels. These returns labeled results and brings up "categories". The two main methods used in unsupervised learning include clustering and dimensionality reduction.

-Dimensionality reduction:

Dimensionality reduction is the process of reducing the number of random variables considered by obtaining a set of principal variables. Put simply, it is the process of reducing the size of a feature set (even more simply, reducing the number of features). Most dimensionality reduction techniques can be classified into two categories: feature removal or feature extraction.
A common method of reducing dimensionality is called Principal Component Analysis (PCA).

-Clustering:

It is an unsupervised technique of grouping data points as already explained in detail.

f) Predictive Models evaluation:

The goal of the ML models is to learn which trends lend themselves well to generalization for unseen data instead of just memorizing the data they may have seen during their training. Once a model is elaborated, it is important to check if it behaves correctly on unpublished data that were not used for training it. To do this, the model predicts the response on the assessment dataset (data set aside) and then compares the predicted target to the actual response (ground truth).
With ML, we feed an algorithm with data in order to teach the computer to perform specific tasks. The performance of such an algorithm essentially depends on its ability to predict the results in a relevant way. To ensure that these correspond to reality, a confusion matrix is used. Under this somewhat barbaric designation hides a relatively simple concept, but formidably effective.

The confusion matrix is like a summary of the prediction results for a particular classification problem. It compares the actual data for a target variable to that predicted by a model. Correct and false predictions are revealed and distributed by class, allowing them to be compared with defined values.

Also known as a contingency table, the confusion matrix is used to evaluate the performance of a classification model. It therefore shows how confusing a certain model can be when making predictions. In its simplest form, it is a 2x2 matrix.

The confusion matrix makes it possible to know on the one hand the different errors committed by a prediction algorithm, but more importantly, to know the different types of committed errors. By analyzing them, it is possible to determine the results that indicate how these errors occurred. Therefore, knowing the type of error is a major advantage of the confusion matrix.

It is not only in the context of ML that confusion matrices are used. They are also used in the field of statistics, artificial intelligence or data mining.

Generally speaking, they excel in quickly analyzing statistical data and simplifying the deciphering of results through data visualization. This is enough to assess the performance of a model and identify the trends that can make it possible to modify the parameters.

Computing a confusion matrix requires having at its disposal a set of test data (test dataset) and another set of validation data (validation dataset) with the expected values of the results. A prediction is then made for each row of the test dataset.

From the expected results and the predictions, the number of correct predictions for each class is calculated, as well as the number of incorrect predictions. These different values are then organized in a contingency table according to well-defined rules.

In the confusion matrix, each row indicates the number of a reference occurrences (or real) class. The columns represent the number of estimated class occurrences. Each column therefore contains a class predicted by the ML algorithm as well as the rows of the actual classes.

The results of a confusion matrix are classified into four main categories:

- ✔ *The true positives or TP:* indicate the cases where the predictions and the actual values are indeed positive.
- ✔ *The true negatives or TN:* indicate on the other hand the cases where the predictions and the actual values are both negative.
- ✔ *False positives or FP:* indicate a positive prediction contrary to the real value which is negative.
- ✔ *False negatives or FN:* refer to cases where the predictions are negative while the actual values are positive.

Various metrics are used in ML to measure the predictive accuracy of a model based on its elaborated confusion matrix. Faced with the multiplicity of modeling methods, each of which has its own statistical quality indicators, statisticians have looked for universal metrics for model performance. The following metrics are the most successful and the most widespread.

f.1) The Area Under the Receiver operating characteristic (ROC-Area):

To visualize the discriminating power of a scoring model a curve called the Receiver Operating Characteristic (ROC) curve is used. It represents a model's ability to discriminate between the outcome of a dependent variable. An area under the ROC curve (AUC) of 1 represents a model that makes all predictions perfectly. An AUC of 0.5 represents a model as good as random. Receiver operating characteristic (ROC) curve is one of the most effective evaluation metrics because it visualizes the accuracy of predictions for a whole range of cut-off values. In order to get ROC, we just need to derive two ratios from the confusion matrix: True Positive Rate (TPR) or Sensitivity, and True Negative Rate (TNR), or called Specificity. TPR and FPR changes as cut-off value changes. one can calculate various TPR and FPR for different cutoff values. When we plot the TPR along the y-axis and FPR along the x-axis, we get the ROC curve. The ROC chart is a great visual exhibit to compare models. If we had a perfect model, the ROC curve would pass through the upper

left corner — indicating no error. A better model is when the ROC is close to the upper left corner. The most important parameter that can be obtained from a ROC curve is the Area Under the Curve (AUC). For a perfect model, the area under the curve would therefore be 1.

f.2) Cohen's Kappa coefficient:

It is a statistical score varying between 0 and 1, used in particular to assess the degree of agreement (of concordance) between two prediction models as to how to classify a set of observations. It can be interpreted as the proportion of agreements (or concordant judgments): proportion of elements classified in the same way by the two models. This index reflects a high level of agreement that is higher if its value is closer to 1, which also means that there is no error. The smaller kappa is less agreement between the truth and predictions.

f.3) Sensitivity:

It measures the model's ability to correctly identify, in a target population, patients who really have the desired characteristic (positive cases). The degree of sensitivity therefore indicates the probability of a model to correctly identify a "case" or the probability that a given "case" is correctly identified by the model. The notion of sensitivity therefore relates to the model's detection capacity.

f.4) Specificity:

It measures the ability of a prediction model to identify, in a target population, individuals who do not have a given specific characteristic ("non-cases"). The degree of specificity of a model therefore indicates the probability that the latter correctly identifies a "non-case" or the probability that a given "non-case" is correctly predicted by the model. The notion of specificity therefore relates to its ability to discriminate them. Specificity is determined by the proportion of people identified by the model as not having a given characteristic (negative result) among people who really do not have this characteristic ("non-case"). These are called true-negatives. A prediction model having a specificity problem identifies subjects who do not actually possess a given characteristic as having it. These discrimination errors are called false positives.

f.5) Gain:

Gain is a measure of a classification model's effectiveness, calculated as the ratio between the results obtained with and without the model. It is mainly used to assess its ability to target accurately. It indicates the degree to which targeting using the model gives a better 'hit-rate', than simply and randomly guessing membership of the target group.

f.6) Accuracy:

This measure reflects a model's ability to provide true value / true ranking. It is the percentage of correctly classified instances out of all instances.

f.7) Matthew's correlation coefficient (MCC):

Matthew's correlation coefficient is a statistical measure used for model evaluation. Its job is to gauge or measure the difference between the predicted values and actual values and is equivalent to chi-square statistics for a 2 x 2 contingency table (Brian Matthews in 1975).

f.8) recall curve area (PRC Area):

Computes the area under the Precision-Recall curve (PRC). The latter can be interpreted as the relationship between precision and recall (sensitivity). Once an assessed model is built for robust predictions, it is needed to decide whether it is a good enough model to solve the classification problem. Therefore, classification accuracy alone is typically not enough information to make this decision. Also, it has to be signaled that as a performance measure, accuracy is inappropriate for imbalanced classification problems. The main reason is that the overwhelming number of examples from the majority class (or classes) will overwhelm the number of examples in the minority class, meaning that even unskillful models can achieve accuracy scores of 90 percent, or 99 percent, depending on how severe the class imbalance happens to be. An alternative to using classification accuracy is to use precision and recall metrics.

f.9) Precision:

The precision shows the percentage of real positive instances (as opposed to false positive instances) among the recovered instances (those that should be positive). In other terms, precision quantifies the number of positive class predictions that actually belong to the positive class.

f.10) Recall:

The recall measure quantifies the total number of actual positive cases that were correctly predicted

f.11) F-measure:

The F1 measurement represents the harmonic mean between the predicted positive rate and the sensitivity. It provides a single score that balances both precision and recall in one single score.

f.12) Classification error:

Classification error is a measure used when a prediction model misclassified an observation into a category to which it does not belong initially. For example, if a classification model trains on a dataset containing a set of X inputs, grouped into a number of categories: A and B. If the model classifies an X1 instance in A when it actually belongs to B category, the model is misclassifying by assigning X1 to a category it doesn't belong to initially.

Therefore, it is a metric that is calculated by summing all the incorrect predictions over the total number of data (positive and negative). The lower it is, the better. The best possible error rate is 0, but it is rarely achieved by a model in practice.

Classification errors can have a variety of causes, including poorly designed classification models, deterministic or random errors in the metrics used to create the dataset, or the effect of finite data size (the more a dataset is small, the more likely it tends to make misclassifications).

Table8: ML evaluation metrics with their respective calculation methods

Metric	Derivations
True Positive (TP)	we predicted "yes", and it's "yes"
True Negative (TN)	we predicted "no", and it's "no"
False Positive (FP)	we predicted "yes", but it's "no"
False Negative (FN)	we predicted "no", but it's "yes"
Negative Predictive Value (NPV)	NPV = TN / (TN + FN)
False Positive Rate (FPR)	FPR = FP / (FP + TN)
False Negative Rate (FNR)	FNR = FN / (FN + TP)
Sensitivity (TPR)	TPR = TP / (TP + FN)
Specificity	TNR = TN / (FP + TN)

Precision	PPV = TP / (TP + FP)
Accuracy	ACC = (TP + TN) / (P + N)
F1-measure	F1 = 2TP / (2TP + FP + FN)
Classification Error	ERR= 1-ACC
ROC	X-axis = FPR = 1 – specificity / Y-axis = TPR= sensitivity
Matthews Correlation Coefficient	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

2. Important variables selection:

In this section we will be interested in a phenomenon that is observed when the dimensional space of the variables (that is, the number of variables) grows so fast that the data it includes becomes scattered and distant. The usual statistical methods will tend in these situations to give distorted and biased results: this is the "scourge of dimension".

The increase in the representation space of the data poses comparison and interpretation problems. Indeed, the dimension increase tends to make the data sparser and thus to distort the traditional ways of data analysis. All statistical methods that require the principle of statistical significance are impacted by the lack of data density in space.

Also, classifying data often corresponds to a grouping of individuals with similar properties. In large dimensions, dissimilarities are accentuated just as individuals move away from each other. We thus lose the method's ability to find individuals who resemble each other.

Size reduction seems to be a major issue in order to cope with a phenomenon which is still little understood by the scientific community but increasingly present in biostatistical applications coupled with genomics.

Thus, the scourge of dimension requires dimension reduction techniques in order to be able to represent the data in a suitable space and more easily interpretable by the usual distances and classical data analysis algorithms.

One solution to this scourge is to use dimension reduction methods. In this category, we will evoke variable selection techniques that aim to select the most important variables among all available ones.

2.a) Minimal Depth

This method assumes that variables with high impact on the prediction are those that most frequently split nodes nearest to the root node, where they partition the largest samples of the population. Within each tree, node levels are numbered based on their relative distance to the tree root (with the root at 0). Minimal depth values indicate that the variable separates large groups of observations and therefore has a large impact on the forest prediction. The distribution of the minimal depth reveals a "ceiling effect" in which a tree simply cannot be grown deep enough to properly identify predictive variables (John Ehrlinger, 2015).

2.b) VIMP

A function that calculates the difference in OOB prediction error before and after permutation. A large VIMP value indicates that misspecification detracts from the predictive accuracy in the Random Forest. VIMP close to zero indicates the variable does not contribute to predictive accuracy, and negative values indicate the predictive accuracy improves when the variable is misspecified. In the latter case, we assume noise is more informative than the true variable. As such we ignore negative variables or equal to zero, relying on large positive values to indicate that the predictive power of the forest is dependent on those variables (John Ehrlinger 2016).

In VIMP, prognostic risk factors are determined by testing the forest prediction, ranking the most important variables according to their impact on the predictive ability of the forest.

Permutation: The idea of "permutation feature importance" consists in opposing the performance of the model in prediction with and without the variable to be evaluated. To neutralize the variable, it is recommended to randomly mix the values inside the vector and, therefore, to break the link that it may have with the class to predict (and the other variables at the same time).

The permutation importance measure was introduced by Breiman (2001) to compensate for the lack of interpretation of random forests. Recall that the trees that make up a random forest are constructed from bootstrap samples of the data. For each tree, the set of observations that are not retained in the bootstrap is called the Out-Of-Bag (OOB) sample. These samples are used to measure the importance of the variables for the prediction of Y. More precisely, a variable X is

considered important if by breaking the link between X and Y, the prediction error increases. To break this link, Breiman proposes to randomly swap X's achievements in the OOB samples. The importance of X then is the average increase in the prediction error over all trees.

The variable with the largest decrease in accuracy / largest increase in error is considered the most important variable.

2.c) Attribute Evaluation using Pearson's Correlation:

This method calculates the correlation between each attribute and the output variable, then, only selects those attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1) while dropping low correlation attributes (value close to zero).

2.d) Attribute Evaluation using Information Gain (IG).

This method calculates the information gain (also called entropy) for each attribute for the output variable. Entry values vary from 0 (no information) to 1 (maximum information). Those attributes that contribute most will have a higher information gain value and can be selected, whereas those that do not add much information will have a lower score and can be removed.

The idea of IG is simple: the more the Entropy being reduced after splitting (that is, the more the dataset being clear after splitting, or says, the information gained by split), the more the Information Gain.

2.e) Symmetrical Uncertainty Attribute Evaluation:

Symmetrical Uncertainty Attribute Evaluation uses the symmetrical uncertainty with respect to the class. In other terms, this method proposes a measure of the relevance between the attribute and the class label. The average normalized interaction gains of attribute “a”, every other attribute and the class label, is calculated to reflect the interaction of attribute “a” with other features in the attribute set A.

2.f) CfsSubset Evaluation:

Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

2.g) Gain Ratio Attribute Evaluation

Gain Ratio attempts to lessen the bias of Information Gain on highly branched predictors by introducing a normalizing term called the Intrinsic Information. The Intrinsic Information (II) is defined as the entropy of sub-dataset proportions. In other words, it is hard for us to guess in which branch a randomly selected sample is put into.

2.h) Relief F Attribute Evaluation:

has the potential to capture feature dependencies in predicting the endpoint, (interactions). Relief calculates a proxy statistic for each feature that can be used to estimate feature 'quality' or 'relevance' to the target concept. Relief calculates a feature score for each feature which can then be applied to rank and select top scoring features for feature selection. Alternatively, these scores may be applied as feature weights to guide downstream modeling.

2. i) OneR Attribute Evaluation:

Evaluates the worth of an attribute by using the OneR classifier. OneR, short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, we construct a frequency table for each predictor against the target. It has been shown that OneR produces rules only slightly less accurate than state-of-the-art classification algorithms while producing rules that are simple for humans to interpret.

Chapter 1: Identification of a minimum number of genes to predict TNBC subgroups from gene expression profiles.

Background

Triple-Negative Breast Cancers (TNBC) affect approximately 15% of women with mammary tumors. The so-called TNBC is an immunohistochemical definition corresponding to the absence of estrogen (ER) and progesterone (PgR) receptors expression, and of the human epidermal growth factor receptor 2 (HER2) amplification. The retained thresholds by the American Society of Clinical Oncology guidelines, of negativity are less than 1% of labeled cells for hormone receptors (Hammond et al. 2010), and 0, 1+ or 2+ scores for HER2 labeling but without Fluorescence in situ hybridization (FISH) amplification for ERBB2 (Hwang and Gown 2016).

TNBC are large, high-grade ductal carcinomas with a high Ki67 mitotic index and numerous nuclear atypia on anatomo-pathological examination (Carey et al. 2007).

These cancers are often related to the basal subtype, introduced for the first time by (Perou et al. 2000) and (Perou et al. 2000; Podo et al. 2010) in their princeps work, and have similarities with cancers developed on germline BRCA mutation. The basal-like subtype (BL) is characterized by basal cytokeratin gene overexpression and the absence of estrogen, progesterone and HER2 coding genes expression. BRCA1/2 gene mutations are found in approximately 30% of cases (Matros et al. 2005). TNBC are usually large high-grade tumors associated with a younger age at diagnosis, with aggressive profile and high rates of p53 gene mutations, accompanied by strong immunohistochemically-detected p53 (Dent et al. 2007). They therefore present a high risk of relapse, despite greater sensitivity to chemotherapy, and of metastatic recurrence in the first three years after diagnosis. They are not eligible for treatments targeting hormone receptors or HER2. However, in addition to chemotherapy, these cancers may benefit from new treatment options, depending on the tumor's nature. Since 2005, the intensive development of high-throughput technologies to analyze gene mutation status and/or expression, has increased the knowledge of the genotypic and phenotypic profile of TNBCs (Geyer et al. 2009).

First, several subcategories can be identified by analyzing their morphology and some have either a particular prognosis, or a specific therapeutic response. Second, high-tech throughput technologies,

thanks to the analysis of thousands of genes, have begun to show TNBC molecular subclasses, exhibiting specific molecular abnormalities associated with response to treatment and/or to survival. Thirdly, evidence has accumulated, showing that TNBC microenvironment, the cells and molecules present in the tumor stroma, play a significant role in disease progression. Thus, the characteristics of the microenvironment can serve as a new TNBC sub-classification basis with a potential therapeutic impact (H. Zheng et al. 2021).

In 2011, a group of researchers led by (Lehmann et al. 2011) at Vanderbilt University, evaluated a new classification, named TNBCtype-6, through which they conducted the identification of six TNBC subtypes, based on gene expression profiling of several hundreds of TNBC samples. Various expression abnormalities related to cell cycle regulatory genes, such as *BRCA2* and *TP53* DNA repair ones, were detected in the BL1 (basal-like type 1) subtype. The second basal-like subtype (BL2) was more associated with abnormal activation of other signaling pathways, such as EGFR, MET, cell migration, extracellular matrix-receptor interaction and differentiation. Contrariwise, the MSL (mesenchymal stem cell) subtype was more associated with underexpression of cell proliferation and overexpression of mesenchymal stem cell related genes. The IM (immuno-modulatory) subtype was mainly recognized by immune signal transduction pathways, such as NK, B, dendritic and T cell ones. The M subtype, on the other hand, was enriched in cell migration-related signaling pathways as well as extracellular matrix-receptor interaction and differentiation pathways. The LAR (luminal androgen receptor) subtype was very different from all the others: although ER receptor negative, it expressed the Androgen Receptor (AR) and/or its downstream effectors, and was highly associated with hormonal-related signaling pathways, such as steroid synthesis and androgen/estrogen metabolism.

In 2016, the same researchers group refined the aforementioned classification as they observed a significant presence of tumor infiltrating lymphocytes (TIL) and stromal cells in the IM and MSL subtypes, respectively. Thus, the previous TNBC subtypes were refined into BL1, BL2, M and LAR, which resulted in the TNBCtype-4 classification (Lehmann et al. 2016).

Thereafter, Burstein supervised another study to identify separate markers that characterize each TNBC subtype. It was found that in addition to copy number variations (CNV) analysis, Genomic profiling techniques may be employed to furthermore stratify triple negative mammary tumors. Consequently, four different subtypes were found, with distinct and stable prognosis, labeled as follows: LAR, Mesenchymal (MES), Immunosuppressed Basal Type (BLIS) and Immune Activated Basal Type (BLIA) (Burstein et al. 2015).

On the other hand, in a more recent study by Jézéquel et al. by transcriptomic profiling techniques, three distinct subtypes were highlighted. The first is recognized by an apocrine molecular phenotype showing favorable prognosis, the other two groups had more basal properties: while one was more aggressive and coupled with an immunosuppressive phenotype, the third showed adaptive immune response (Jézéquel et al. 2019; Ensenyat-Mendez et al. 2021)

Finally, another study developed by Liu et al. and based essentially on Long-non-coding RNAs (lncRNAs) to classify TNBC tumors, resulted in the development of the Fudan University Shanghai Classification System (FUSCC). Four subtypes were recognized: IM, LAR, MES, and BLIS, showing upregulation of proliferative pathways and the worst OS. (Y.-R. Liu et al. 2016)

However, the potential driving molecular events within each TNBC subtype, as well as their response to treatment, remain seldom explored. Further insights into the underlying genomic alterations, as well as towards a standardized and easily applicable subclassification, are therefore needed.

The efforts made over the past 10 years to better understand the biology of TNBC have led to an important conclusion: the term “triple-negative” covers different cancers. Some of them have a completely defined “molecular portrait”, which can be identified by genomic methods. However, if the path to the integration into clinical practice of a molecular and morphological portrait is still long, it will nevertheless have to be done to offer a more accurate diagnosis as well as a more personalized treatment for patients.

Under the same perspective and starting mainly from Lehman's classification, we aimed at identifying a limited number of genes that can serve as a genetic signature for the prediction of the different TNBC subtypes.

Materials and Methods

- TNBC datasets

Two TNBC datasets were downloaded from public repositories. The first one was retrieved from the Gene Expression Omnibus (GEO) and refers to whole transcriptome RNA sequencing (RNAseq) performed on pre-treatment research biopsies from the BrighTNess phase III study (AFT-04). This dataset contains log-normalized RNA-seq expression values with clinical stages II to III data. It will be called GEO-TN. (Loibl et al. 2018)

The second one was retrieved from the Genomic Data Commons (GDC) Data Portal of the National Cancer Institute and refers to the cancer genome atlas (TCGA) project: TNBC samples only were selected, based on their ER, PgR and HER2 negative immunohistochemical status, which left us with a total of 63 TNBC records out of 1093 IBC records. This dataset contains log-normalized RNA-seq expression values and their respective clinical data. It will be called TCGA-TN.

The third dataset refers to 72 TNBC samples from Italian patients surgically treated at the Hospital of Biella and at the Policlinico Gemelli in Rome, that underwent gene expression profiling at the Genomics Lab of Fondazione Edo ed Elvo Tempia, Biella (Italy). It will be called Italian-TN. Sample collection was approved by the Ethical Committees of Novara and Policlinico Gemelli (Prot. 861 CE 149/19 and Prot. 3559, respectively). After tumor area selection at the Pathology Department of Biella hospital, macro-dissection and section cut from tumor blocks was carried out at the Molecular Oncology lab of Fondazione Edo ed Elvo Tempia. Total RNA extracted from tissue sections was reverse-transcribed to corresponding cDNA and then in vitro transcribed in order to amplify, label it and allow for gene expression profiling, using whole genome Agilent SurePrint G3 Human GE 8x60K V3 microarrays (Agilent Technologies) containing probes for

26,803 Genes and 30,606 long non-coding RNAs (lncRNAs). After hybridization and scanning, array image analysis was carried out using the Agilent Feature Extraction Software v12.1, and then raw expression data was pre-processed by background subtraction followed by within and between array normalization, using the LIMMA (LInear Models for Microarray Analysis") package in R software. This dataset contains log normalized intensities.

- TNBC subtype prediction

Pre-processed data from the GEO-TN, TCGA-TN and Italian-TN datasets were uploaded in the TNBCtype online tool (X. Chen et al. 2012). This tool first investigates the presence of any hormone receptor positive sample and filters it out. Then it uses Spearman correlation of each candidate tumor sample to compare it with each of six centroids of TNBCsubtypes previously determined: BL1, BL2, LAR, IM, M, MSL. Then it assigns it to the subtype that is most correlated to it and admits it as the final predicted subtype. It also determines the statistical significance of the correlation coefficient. UNS is assigned to unstable samples, with very low and non-statistically significant correlation with any subtype. UNS samples were excluded from downstream analysis.

- Data cleaning

For the GEO-TN dataset, there were 23 ER+ detected and 64 UNS predicted samples, which were discarded. Accordingly, the final number of samples obtained was 395. This dataset was used as a training set. As for the TCGA-TN dataset, it initially consisted of 63 records from which we discarded 13 unstable ones, resulting in 50 TNBC samples. 17 samples were predicted as UNS and were therefore automatically eliminated from the Italian dataset, which resulted in a final number of 55 samples. The two latter datasets were considered as validation sets. An additional filter was used to remove non-expressed genes in all the samples (with at least one zero expression value).

- Gene signature determination:

This step was elaborated by "R software for Statistics v.4.1.0" and based on the calculation of Differentially Expressed Genes (DEGs) specific to each TNBC subtype, in contrast to the others. Two different methods were selected to have the best DEGs pick. The first one was gene

expression analysis with the LIMMA package, where differentially expressed genes between each predicted TNBC subgroup and the remaining samples were obtained by combining a modified t-test with empirical Bayes modeling, in order to moderate the standard errors of the estimated log-fold changes. The detection of differential gene expression was done by applying a cut-off to the Benjamini & Hochberg adjusted p-values (alpha = 0.01). The second method used was the mean difference based on Mann-Whitney U (MWU) test, using the same method to adjust p-values for multiple test comparisons. The detection of differential gene expression was done by applying a cut- off to the adjusted p-values (< 0.01) and to the difference in median expression between subgroups (>1 and <1) for Up- and Down- regulated genes, respectively. Both method outcomes were combined by the "*merge*" function from the "*dplyr*" package on R for further analysis.

- Subtype prediction according to the genetic signature

This step was assessed by "*Weka v3.9.3 software for data mining*". The "subtype membership" was considered as the variable of interest, while all the other attributes (selected genes) were used as predictive variables. Relevant ML algorithms were therefore selected to compare and evaluate the model performance. The following models were used: Naive Bayes (NB); Logistic Regression (LR); Decision Tree (DT); Random Forest (RF); Support Vector Machine (SVM); K-nearest neighbors (KNN) classifier; Multilayer Perceptron (MP).

The analysis includes an automatic feature engineering, which is based on a k-fold cross-validation (Jung and Hu 2015), where the original sample is partitioned into k subsets. The model is trained on all but one subset (k-1), then evaluated on the subset that was not used for training. This cross-validation process is systematically repeated k times (the folds), where each of the k subsets is used exactly once as validation data (and excluded from training) each time. The k fold results are then averaged (or otherwise combined) to produce a single final estimate.

- Prediction evaluation metrics

Each prediction model was evaluated by ten different metrics, which are:

-True positive (TP) rate; False positive (FP) rate; Accuracy; Cohen's Kappa; Precision; Recall; F-measure; Matthews correlation coefficient (MCC); The Area Under the Receiver operating characteristic (ROC) curve; Precision-recall curve area (PRC Area).

- Best attribute selection

This step is useful to choose a small subset of features (genes) that is sufficient enough to classify the target class (TNBC subtype) effectively, by reducing computational cost and improving accuracy. Accordingly, the quality of each gene of the training dataset was considered to detect worthless ones in prediction. Genes that provide less value (voted by the majority rule of different attributes selection algorithms) were discarded, the goal being to produce a genetic signature computationally faster and composed of a lower number of genes.

Consequently, seven different attribute selection algorithms were used. Their central hypothesis is that the important attribute sets are strongly correlated with the target class, and uncorrelated attributes are less important. Further, strong correlation among attributes makes only one of them important and the other one can be removed. If two or more attributes have the same importance to the target class values, it will be good to consider only one of them.

1. Attribute Evaluation using Pearson's Correlation;
2. Attribute Evaluation using Information Gain;
3. Symmetrical Uncertainty Attribute Evaluation;
4. Cf Subset Eval;
5. GainRatioAttributeEval;
6. ReliefFAttributeEval;
7. OneRAttributeEval;

The final attributes selection methods list gathers the results of the ranking of all the attributes from the most to the least important. Only genes that were ranked as unimportant by at least four out of seven algorithms were then highlighted as the least important attributes.

- TNBC subtypes network analysis and identification of druggable targets

Process analysis of the genes that were differentially expressed was performed using the web-based algorithm, MetaCore™ version 22.1. software suite (Clarivate Analytics, Philadelphia, PA, United States). Gene network analysis was carried out using Dijkstra's Shortest Path algorithm to find the shortest path between gene (or gene product) pairs, in each direction. The original genes were linked with additional objects from the database along a directed path, using a predefined maximum number of steps (1 or 2).

Results

1. TNBC subtypes prediction and gene signature determination

All the three TNBC datasets were subtyped using the TNBCtype online tool. For the GEO-TN dataset, there were 23 ER+ detected and 64 UNS predicted samples, which were discarded. Accordingly, the final number of samples obtained was 395. This dataset is by far the largest and was used as a training set. The TCGA-TN dataset initially consisted of 63 records from which 13 unstable ones were discarded, resulting in 50 TNBC samples. 17 samples were predicted as UNS and were therefore automatically eliminated from the Italian-TN dataset, which resulted in a final number of 55 samples. The two latter were used as validation sets. Subtyping results for the three datasets are detailed in (figure 15). The IM and M subtypes are the most prevalent, while BL2 and LAR are the least frequent, which can give us an idea about the subgroup imbalance.

The two tests used to determine differentially expressed genes converged on the most significant genes within each subgroup in contrast to the others. Subsequently, two gene lists were generated, the first with the 120 most up-regulated (Table 1_Supplementary Material) and the second with the 81 most down-regulated genes (Table 2_Supplementary Material).

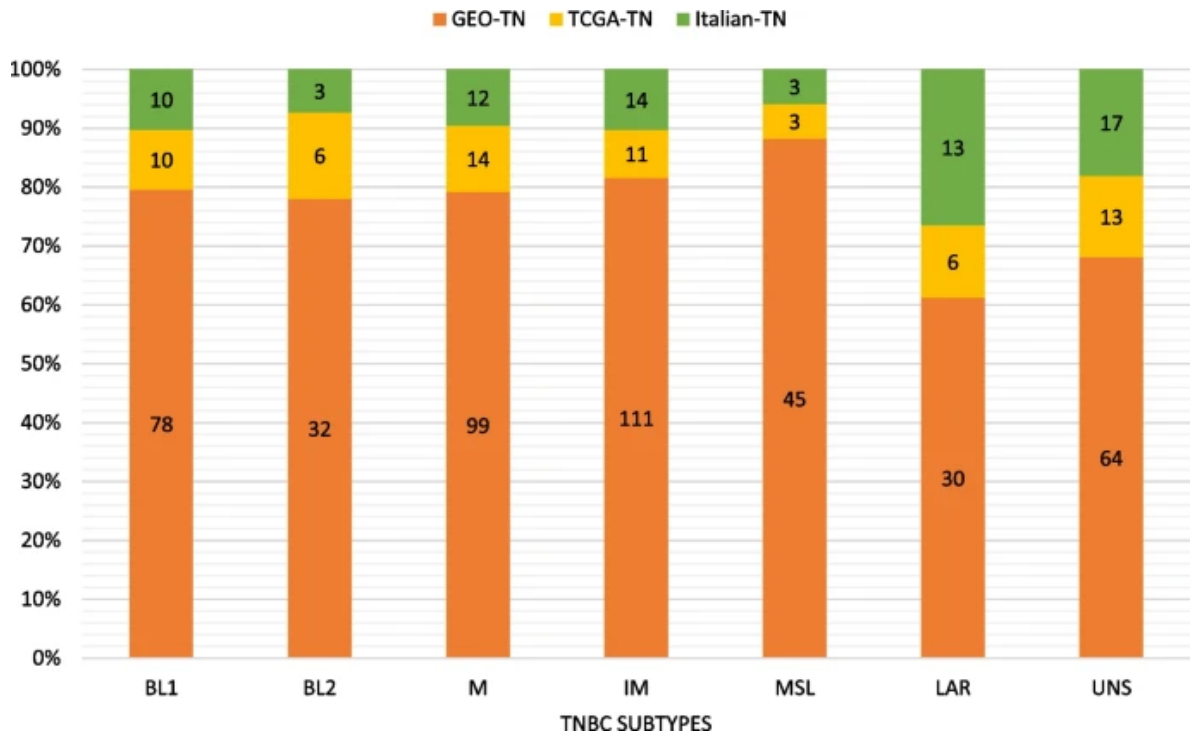

Figure 15: Predicted subtypes count in GEO-TN; TCGA-TN and Italian-TN datasets by TNBCtype tool.

2. TNBC subtype network analysis

It is of great interest to look for genetic interactions within the few TNBC subgroup signature genes. This can lead to a better understanding of the TNBC-subtype specific phenotypes than by just considering single gene effects. To identify complex pathways that control essential functions in TNBC subtype-specific cancerogenesis, we analyzed gene networks using the "shortest paths" function of the Metacore analysis suite, allowing for maximum 2 steps (one extra element as intermediary) to connect the genes in the path. We found interactions between each subtype-specific gene (or its product) and other entities such as binding proteins, enzymes, transcription factors, protein kinases and receptors with enzyme activity, through different regulation mechanisms. All the BL1 up-regulated genes except *KLRG2* are connected via one or two transcription factors (Table 3_Supplementary Material), with *ELF5*, *PADI2*, Matrilysin (*MMP-7*), *COBL* and *CLSP* being the most interconnected signature genes and HNF3-alpha, Androgen and Estrogen Receptors the most interconnected intermediary transcription factors (Figure 16).

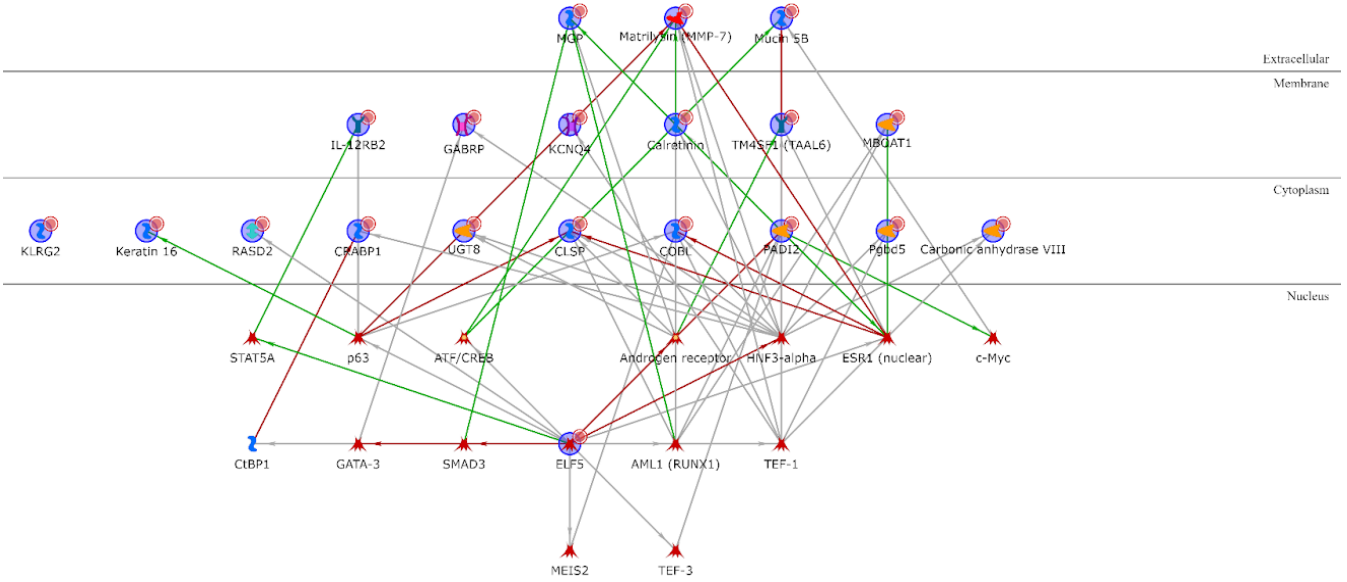

Figure 16: BL1 up-regulated genes network analysis

Red arrows refer to inhibition, green arrows to activation and grey ones to unspecified effects, while red circles refer to uploaded differentially expressed genes

Among the BL1 down-regulated genes (Table 4_Supplementary Material), only *IGF-2* and *PRSS11* (HtrA1) are connected via *Vitronectin* or *IBP* and the location of all the four proteins is extracellular (Figure 17).

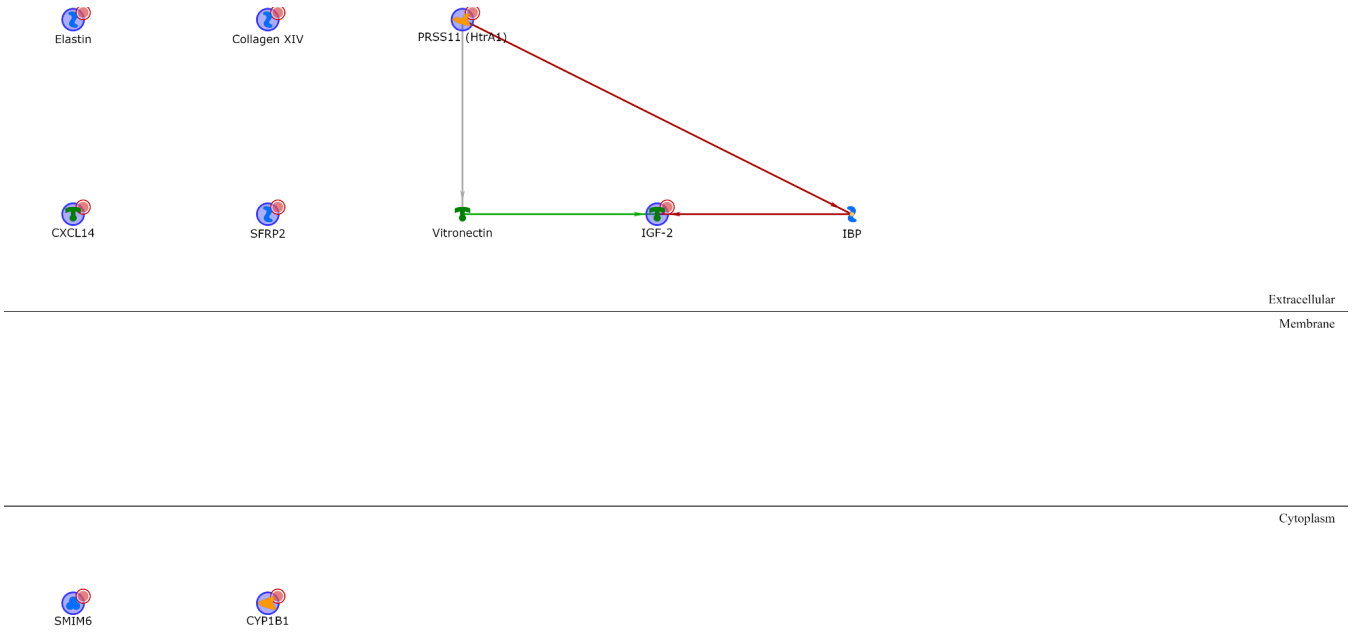

Figure 17: BL1 down-regulated genes network analysis

Concerning BL2 up-regulated genes (Table 5_Supplementary Material), most of them encode for cytoplasmic proteins transcriptionally regulated by a few intermediary transcription factors (p53, STAT3, RAR-alpha, Androgen Receptor, FKHR), except for cytoplasmic Calgranulin A that is directly linked to extracellular Calgranulin B via an autoregulatory loop (mutual activation by binding). *S100-A16* is not connected to any other up-regulated gene, while the only other extracellular product, *Stromelysin-1*, is transcriptionally regulated by several intermediary transcription factors and is also a therapeutic drug target (see chapter below). The only nuclear product is SFN and there are six membrane proteins, all controlled by a few intermediary transcriptional factors (Figure 18).

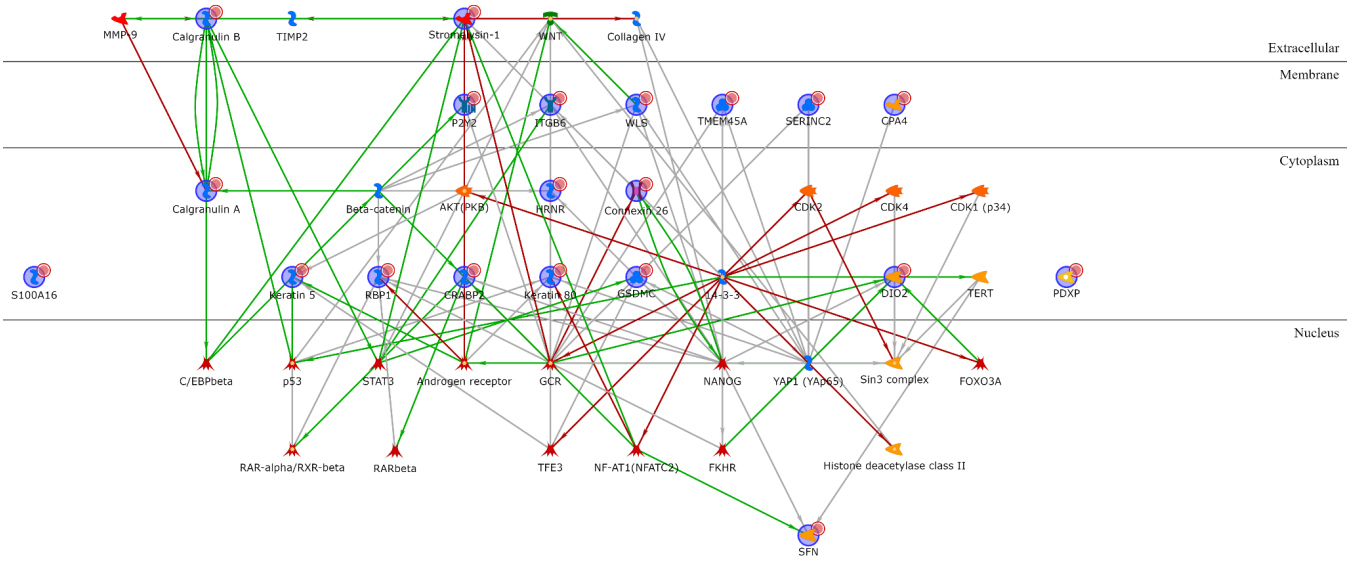

Figure 18: BL2 up-regulated genes network analysis

Among the BL2 down-regulated genes (Table 6_Supplementary Material), the most interconnected proteins are NDRG2 and COBL, both cytoplasmic, BAMBI and MBOAT1, both located on the cell membrane, and EHZF that is located in the nucleus (Figure 19).

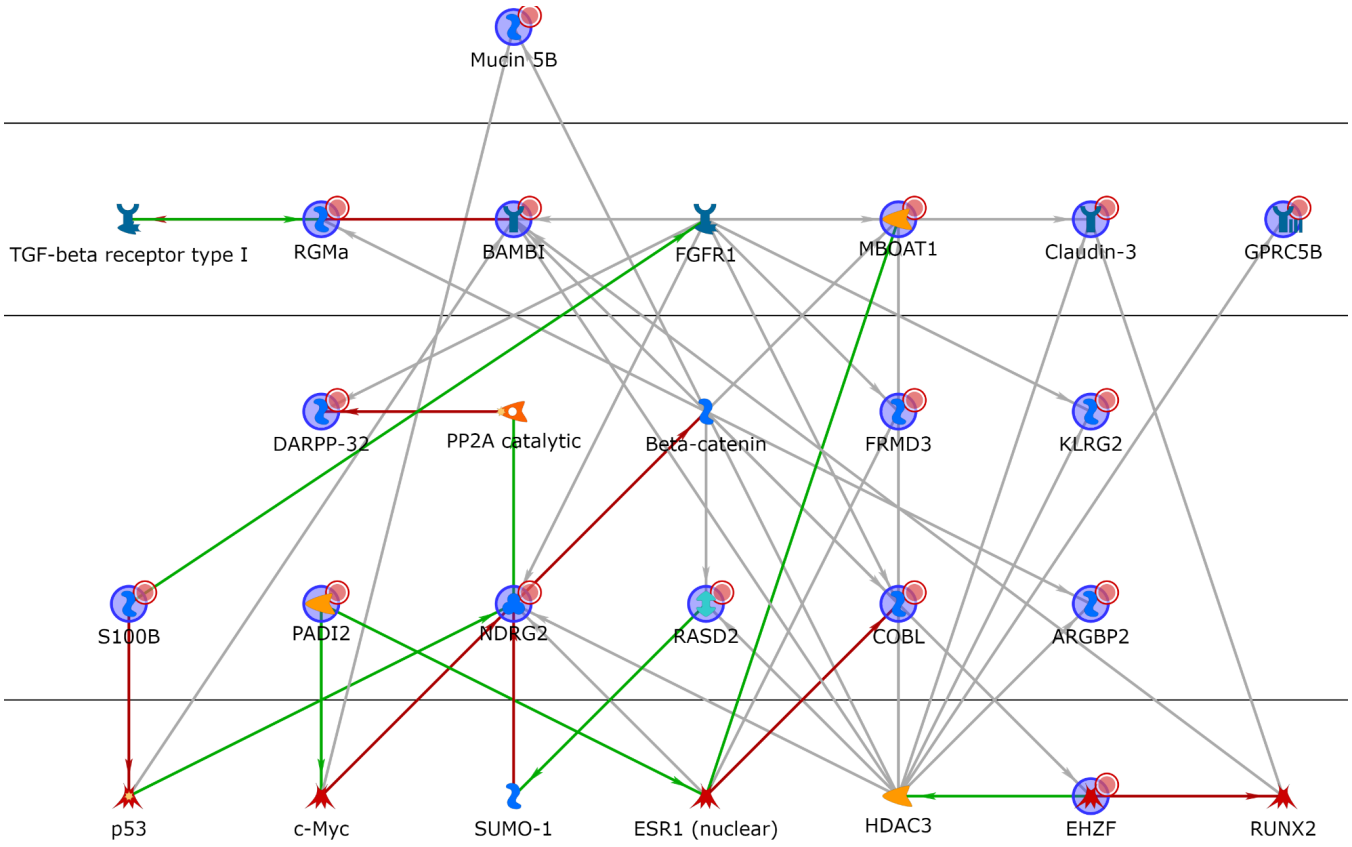

Figure 19: BL2 down-regulated genes network analysis

Twelve out of the twenty LAR up-regulated gene products are directly regulated by the Androgen receptor, that is in the LAR signature it-self (Table 7_Supplementary Material). These include: the Amphiregulin extracellular protein; four membrane proteins (alpha-ENaC, CD166, TSPAN1, STEAP4); seven cytoplasmic proteins (ALOX15B, FLJ20184, KIAA1324, ATAD4, CRAT, FASN, CYP19) (Figure 20).

Figure 20: LAR up-regulated genes network analysis

Thirty-one out of thirty-five proteins encoded by the LAR down-regulated genes are directly connected without any intermediary (Table 8_Supplementary Material), with the transcription factors LBP9, c-Myc and CXXC1 controlling most of the signature genes (Figure 21).

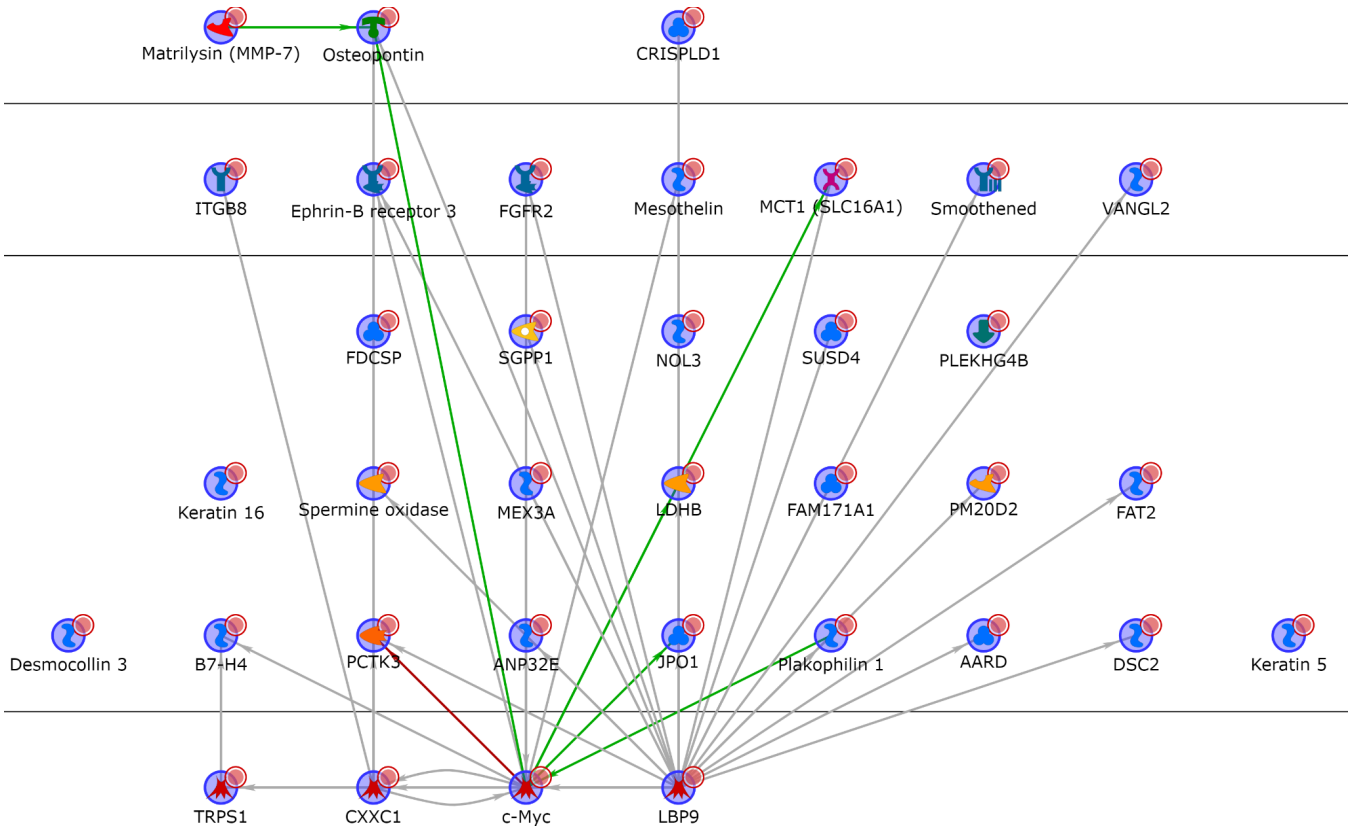

Figure 21: LAR down-regulated genes network analysis

None of the proteins encoded by the M subtype up-regulated genes are directly connected with any of the others (Table 9_Supplementary Material), but they are all connected if one intermediary is added, with SOX6 and ID4 (nuclear), MDFI and Desmocollin 3 (cytoplasmic), and the BAMBI transmembrane glycoprotein being the most interconnected network hubs. The network involving the proteins encoded by the down-regulated M genes (Table 10_Supplementary Material) is not easily interpretable (Figure 22).

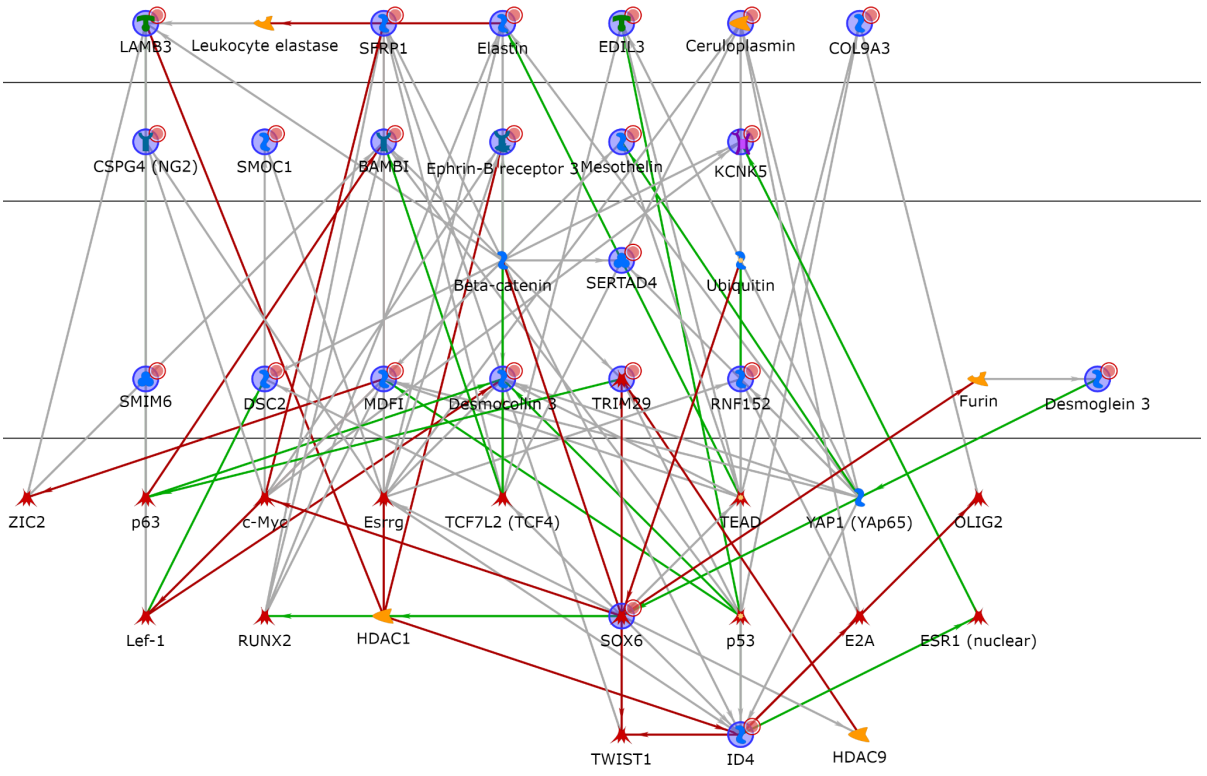

Figure 22: M up-regulated genes network analysis

The network involving the proteins encoded by the down-regulated M genes (Table 10_Supplementary Material) is not easily interpretable (Figure 23).

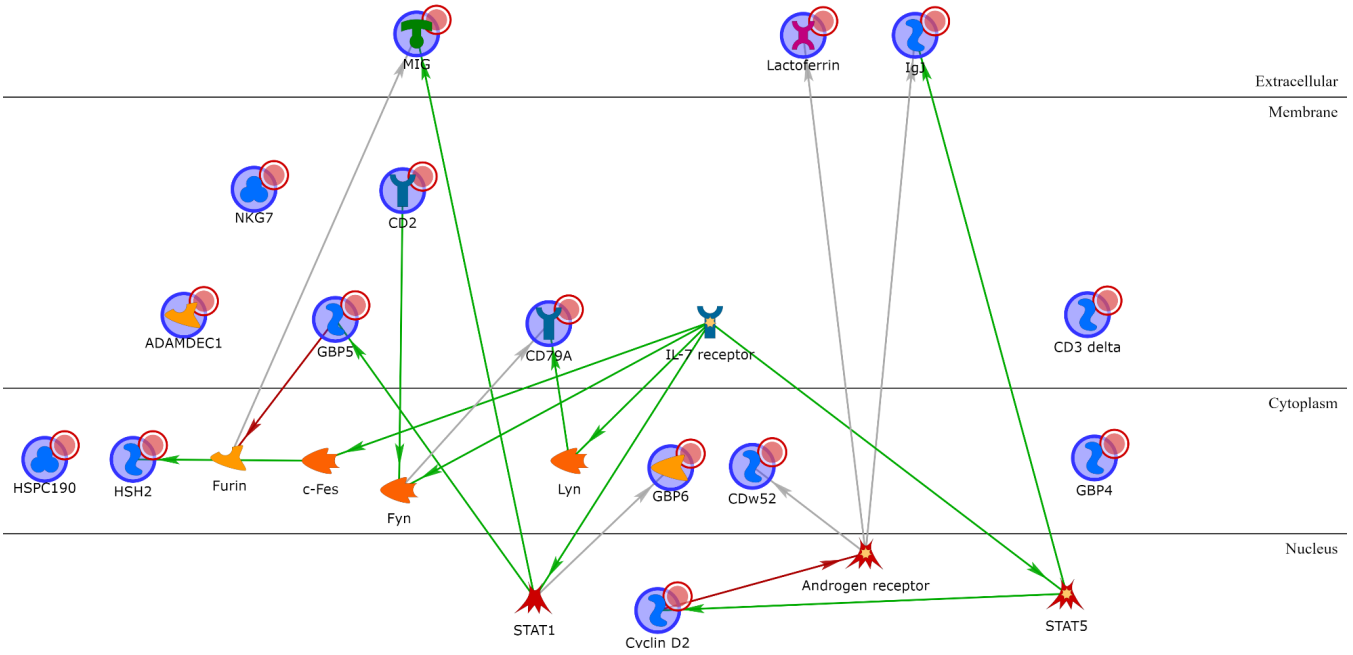

Figure 23: M down-regulated genes network analysis

As for the IM subtype, the only two up-regulated genes encode for two transcription factors (Table 11_Supplementary Material), SPI-B and Aiolos, that are among the most interconnected within the network when one intermediary is included. The majority of intermediaries converge towards IP-10, MIG or I-TAC, three extracellular chemochines, or to CD38, a type II transmembrane glycoprotein, all overexpressed in the IM subtype. Another central node of the IM network is Granzyme B, a protease secreted by natural killer cells and cytotoxic T lymphocytes (Figure 24).

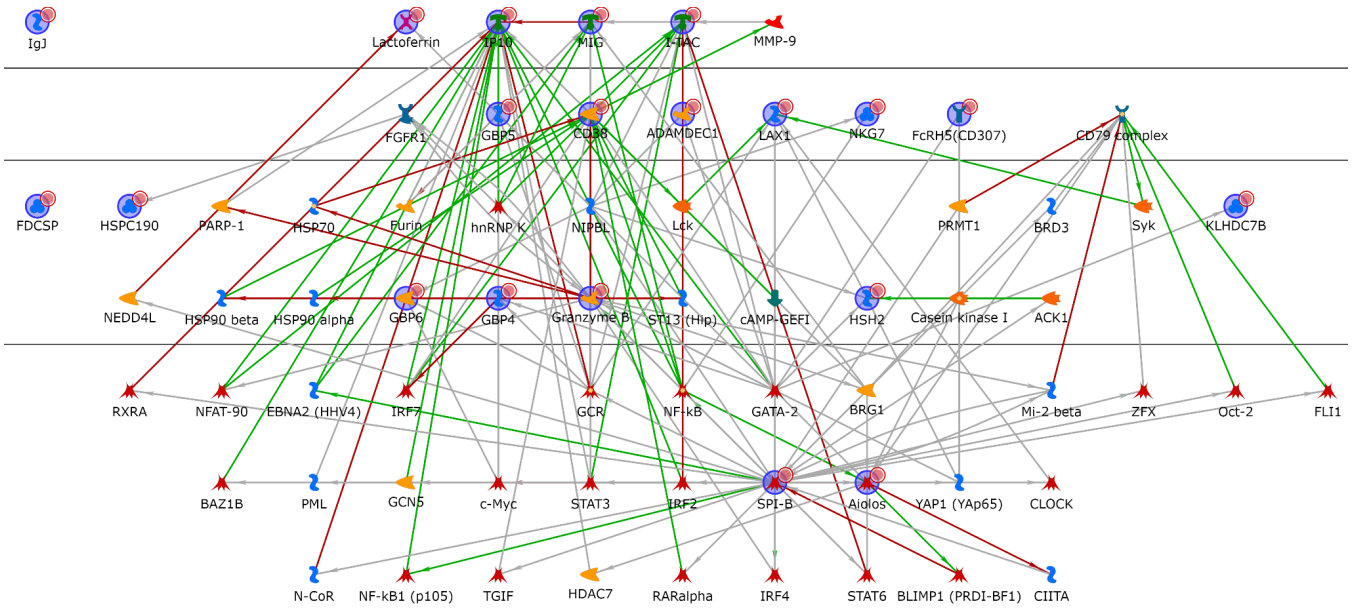

Figure 24: IM up-regulated genes network analysis

The IM down-regulated genes (Table 12_Supplementary Material) are *ID4*, *MDFI*, *KRT81*. Only the proteins encoded by the first two are connected, via either the transcription factor p53 or the demethylase JMJD2A (Figure 25).

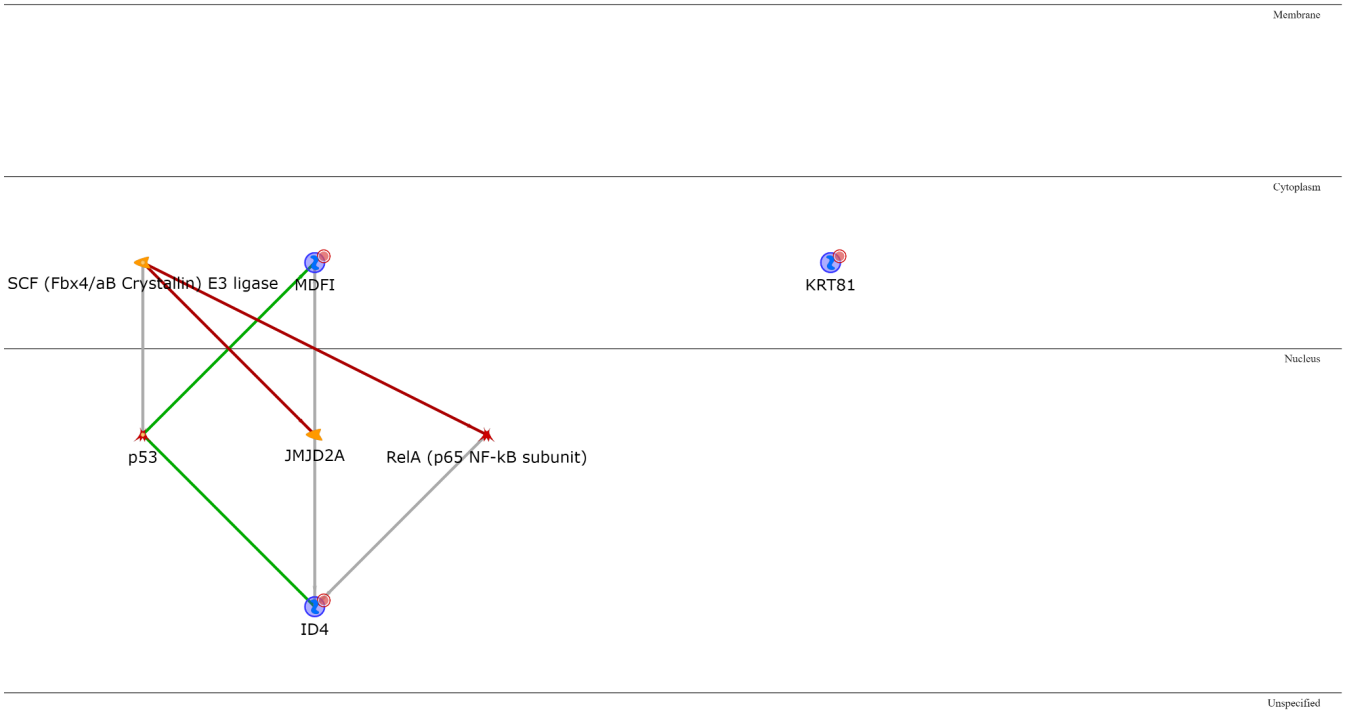

Figure 25: IM down-regulated genes network analysis

Finally, the non-coding gene *MEG3* is the central element in the network resulting from the MSL up-regulated genes (Table 13_Supplementary Material) and is linked to IGF-I and IGF-II via inhibition of several microRNAs (miR-218-3p, miR-96-5p, miR-19-3p, miR-493-5p, miR-665-3p, miR-129-5p, miR-18a-5p, miR-129-3p, miR-181a-5p) targeting the two extracellular growth factors (Figure 26).

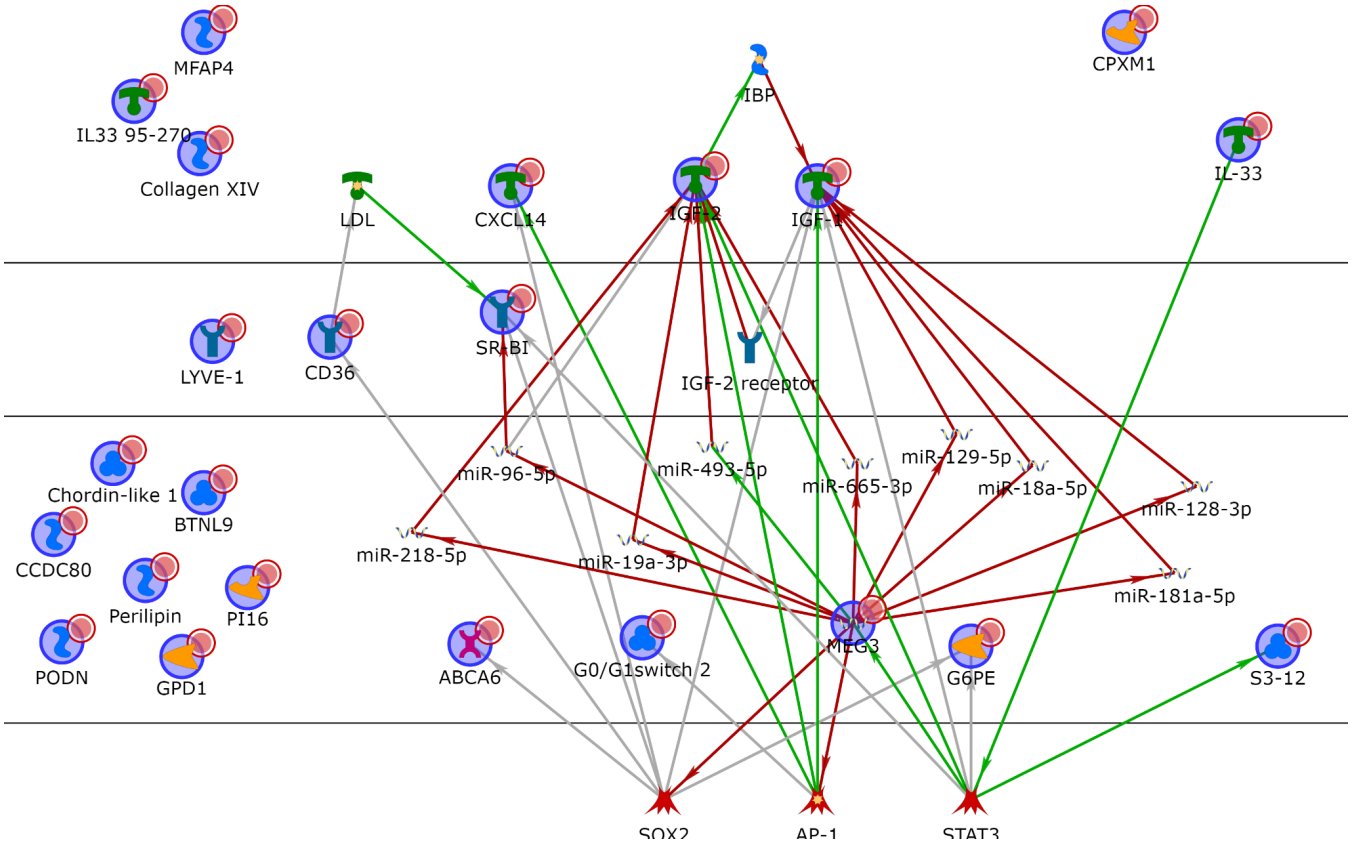

Figure 26: MSL up-regulated genes network analysis

On the other hand, cell cycle controlling elements such as *CDK1* and *CDKN2A* (Table 14_Supplementary Material) have a central role within the MSL down-regulated genes (Figure 27).

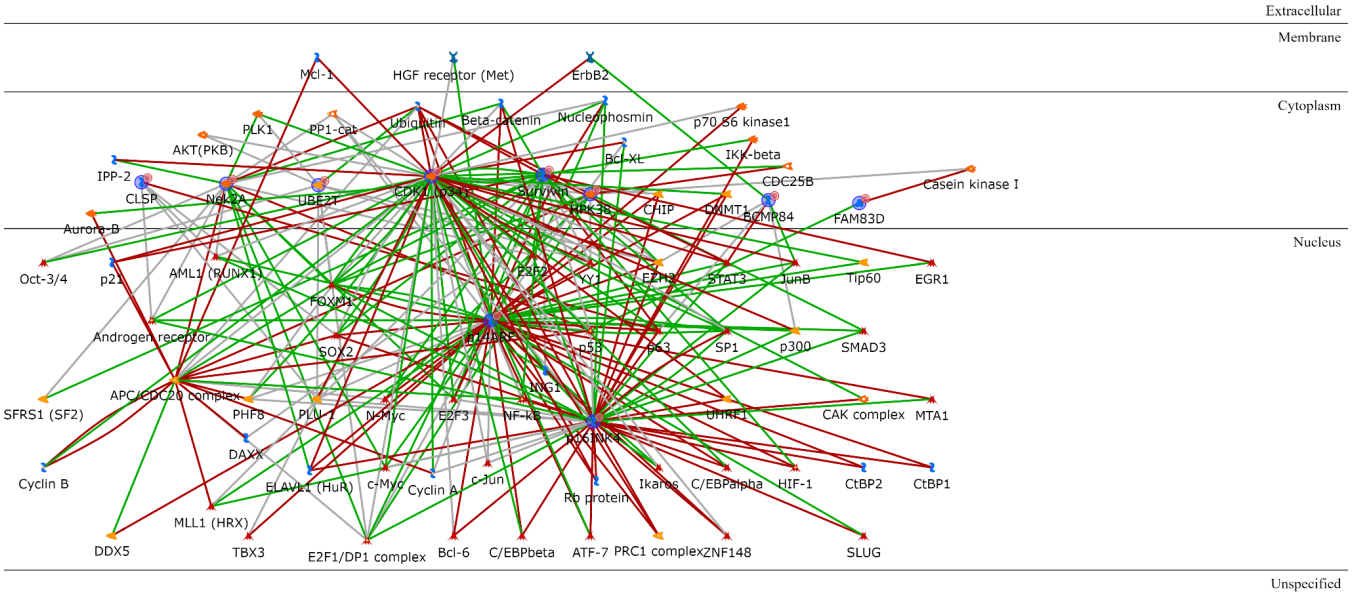

Figure 27: MSL down-regulated genes network analysis

3. Identification of druggable targets

The genes differentially expressed in each subtype were subsequently analyzed with Metacore, to look for any druggable target. The most overexpressed BL1 druggable target is Matrilysin, encoded by *MMP7* and targeted by several therapeutic inhibitor drugs, such as: Batimastat; Marimastat, and Rebimastat (Table 15_Supplementary Material).

As for the BL2 subgroup, the main therapeutic drug-target inhibitory interaction concerns Stromelysin-1 encoded by *MMP3* and targeted by Doxycycline, and Tanomastat (Table 16_Supplementary Material).

On the other hand, one of the most recurrent and potentially important up-regulated LAR druggable targets is Androgen receptor encoded by *AR* and inhibited by Bicalutamide, Diethylstilbestrol, Drospirenone, Finasteride, Flutamide, Metandienone, RU58841, Silibinin, Zanoterone. The second is CYP19 encoded by *CYP19A1* and targeted by several aromatase inhibitors, such as Aminoglutethimide, Anastrozole, Exemestane, Letrozole, and Testolactone. Then GGT1, targeted by Acivicin and by Oxiglutathione; GGTF-I-beta, encoded by *PGGT1B* and targeted by L-778,123; ALDR, encoded by *AKR1B1* and targeted by Tolrestat; alpha-ENaC, encoded *SCNN1A* and targeted by Amiloride (Table 17_Supplementary Material). As for the M, IM and MSL subtypes (Tables 18_, 19_, 20_Supplementary Material), no specific therapeutic drug-target interactions were spotted. Conversely, several inhibition secondary drug-targets interactions for the up-regulated genes, predicted based on similarities in the structures, were found. Ephrin-B receptor 3, encoded by *EPHB3* and up-regulated in the M subgroup, is a predicted target of several inhibitory drugs such as CC-223, Dovitinib, Nazartinib, Nilotinib and Ponatinib; CD38 in the IM subgroup is a predicted target of Ca ('2+), Fluticasone propionate and Quercetin; SR-B encoded by *SCARB1* and overexpressed in the LAR group is a predicted target of beta-Cyclodextrin, Docosahexaenoic acid and ITX-5061.

Reciprocally, no activating therapeutic drug-target interaction for the down regulated genes was spotted in all the six TNBC subgroups (Table 21_ to Table 26_ Supplementary Material).

4. TNBC subtype prediction

It is very important in any biological study to identify the most meaningful information from complex biological data. It is known that physiological and pathological changes in the tumor phenotype and its sensitivity to specific treatments are generally driven by molecular interactions. Hence, we evaluated if the subtype-specific gene signatures previously described were also able to predict sample classes.

Accordingly, seven different prediction models were applied on the GEO-TN dataset, starting from the lists of up-regulated (Table 1_Supplementary Material) and down-regulated (Table

2_Supplementary Material) genes previously obtained. For both lists, 10-fold cross validation was used as it gives the models the opportunity to train on multiple train-test splits, giving a better indication of how well the models perform on unseen data. The variable to predict was “TNBC subtype” and the explanatory features were the Up/Down regulated genes.

Tables 9 and 10 summarize the weighted averages across the 6 classes of the metrics used to judge each model’s performance in classifying the samples using the up- and the down-regulated genes, respectively.

Table 9: Comparative overview of 7 prediction algorithms according to the 120 up-regulated genes.

(TP) True positive; (FP) False Positive; (MCC) Matthews correlation coefficient; (ROC) receiver operating characteristic; (PRC) precision-recall curve.

	TP rate	FP rate	Accuracy %	Mean absolute error	Kappa	Precision	Recall	F-measure	MCC	ROC	PRC-Area
Naive Bayes	0,863	0,035	86.3291	0.0452	0.8281	0,864	0,863	0,863	0,828	0,981	0,934
Logistic Regression	0,595	0,105	59.4937	0.1343	0.4884	0,596	0,595	0,594	0,492	0,862	0,664
Multilayer Perceptron	0,894	0,027	89.3671	0.0429	0.8658	0,894	0,894	0,893	0,867	0,987	0,951
SVM	0,889	0,029	88.8608	0.2257	0.8597	0,889	0,889	0,888	0,860	0,963	0,845
k-Nearest Neighbours	0,808	0,049	80.7595	0.0677	0.7579	0,811	0,808	0,808	0,759	0,872	0,695
Decision Tree	0,646	0,096	64.557	0.1414	0.5488	0,653	0,646	0,646	0,557	0,845	0,603
Random Forest	0,858	0,046	85.8228	0.134	0.8191	0,865	0,858	0,852	0,821	0,985	0,941

Table 10: Comparative overview of 7 prediction algorithms according to the 81 down-regulated genes

(TP) True positive; (FP) False Positive; (MCC) Matthews correlation coefficient; (ROC) receiver operating characteristic; (PRC) precision-recall curve.

	TP rate	FP rate	Accuracy %	Mean absolute error	Kappa	Precision	Recall	F-measure	MCC	ROC-Area	PRC-Area
Naive Ba	0,846	0,037	84.557	0.0515	0.8055	0,848	0,846	0,846	0,809	0,980	0,926
Logistic Regression	0,668	0,082	66.8354	0.1091	0.5828	0,672	0,668	0,669	0,587	0,919	0,758
Multilayer Perceptron	0,886	0,029	88.6076	0.047	0.8563	0,885	0,886	0,886	0,858	0,988	0,953
SVM	0,861	0,036	86.0759	0.2264	0.8241	0,861	0,861	0,860	0,826	0,958	0,808
k-Nearest Neighbors	0,744	0,066	74.4304	0.0884	0.6777	0,744	0,744	0,740	0,677	0,836	0,615
Decision Tree	0,618	0,098	61.7722	0.1542	0.5131	0,612	0,618	0,610	0,521	0,847	0,582
Random Forest	0,825	0,053	82.5316	0.1386	0.7755	0,827	0,825	0,813	0,781	0,979	0,919

The MP followed by SVM model stand out with the best metrics scores; on the other hand, LR and DT seem to be the least performant among all models, for both lists. Therefore, MP was then picked for further use in external validation on the TCGA-TN and Italian-TN datasets. Consequently, in order to know if any of the genes had a low predictive weight according to the best predictive model (MP), seven different attribute selection methods were elaborated, which voted for slightly different ranking genes. The genes that were voted by the majority of algorithms as unimportant were removed (Table 27_Supplementary Material). Following the two gene lists refinement, a per-subgroup ROC comparison was made, before and after attribute selection, to evaluate if the aforementioned gene elimination altered the prediction performance of the same model. The predictions were first measured on the training set with the 10-folds cross validation option, and then on the two validation sets. Very stable ROC scores were obtained, even after deletion of the least important genes. In terms of the up-regulated genes, despite the removal of 17 genes, the ROC score improved in both the training and the validation datasets, in the majority of cases. The detailed ROC areas by class and the weighted averages are shown in Table 11, for up-regulated (upper rows) and down-regulated genes (lower rows), before and after attribute selection.

Table 11: per-subgroup prediction ROC scores for up-and down-regulated genes, before and after attribute selection.

	per-subgroup prediction ROC metric before attribute selection (Total number of Up-regulated genes=120)							
	BL1	**BL2**	**M**	**IM**	**MSL**	**LAR**	**Weighted average**	**Validation option**
Up- regulated genes	0.979	0.97	0.983	0.992	0.996	0,999	0.987	Cross-validation on GEO set (10-Folds)
	0.862	0.949	0,837	0,890	0.904	0.784	0.852	Validation set: Italian set
	0.958	0,981	0.958	0.958	0.816	0.92	0.948	Validation set: TCGA set
	per-subgroup prediction ROC metric after attribute selection (Total number of Up-regulated genes=103)							
	0.975	0.978	0.984	0.988	0.996	1	0.986	Cross-validation on GEO set (10-Folds)
	0.916	0.955	0.843	0.922	0.962	0.7	0.883	Validation set: Italian set
	0.96	0.955	0.974	0.951	0.801	0.864	0.94	Validation set: TCGA set
	per-subgroup prediction ROC metric before attribute selection (Total number of Down-regulated genes=81)							
Down-regulated genes	0.986	0.977	0.987	0.991	0.988	0.998	0.988	Cross-validation on GEO set (10-Folds)
	0.727	0.808	0.924	0.871	0.897	0.886	0.858	Validation set: Italian set
	0.678	0.788	0.861	0.781	0.858	0.985	0.813	Validation set: TCGA set
	per-subgroup prediction ROC metric after attribute selection (Total number of Down-regulated genes=77)							
	0.984	0.961	0.984	0.986	0.985	0.996	0.984	Cross-validation on GEO set (10-Folds)
	0.742	0.962	0.816	0.815	0.776	0.91	0.83	Validation set: Italian set
	0.743	0.807	0.813	0.776	0.7	0.977	0.802	Validation set: TCGA set

Discussion

The development of a plausible treatment for TNBC subtypes is largely hindered by the high heterogeneity of their different phenotypes. Indeed, TNBC patients are pathologically defined by the triple negative expression of ER, PgR and HER2 receptors, and not positively via specific markers that may represent druggable targets.

In this research study, starting from a large dataset of TNBC records and applying the classification proposed by Lehman and collaborators, which relies on whole transcriptomic profiles, we were able to define two small size classifiers, one based on the most over-expressed and the second on the most under-expressed genes within each of the 6 TNBC subtypes. The models were tested on two independent datasets, in order to evaluate the accuracy of the subtype prediction. The least important genes were discarded, to define a minimum number of genes associated with TNBC subtyping. The final classifiers consisted of 103 up-regulated or 77 down-regulated genes, most of which had been previously found by several authors to be associated with TNBC or to basal-type

BC or to BC in general. Therefore, our results add new important pieces of information that may help clinicians in the classification of TNBC. Knowing that a "one-size-fits-all" treatment approach is questionable for TNBC, molecular subtyping is crucial in determining the best therapeutic option for each single patient.

Concerning the basal-like phenotype, stratified into two further subtypes, we found that genes overexpressed in BL1 tumors are enriched in the major mechanisms that define this particular subtype: cell proliferation and DNA damage response. Most of these genes have been previously associated with the basal phenotype, and our study highlights their BL1-specificity. Specifically, *CRABP1*, which proved to be under-expressed in hormone-dependent tumors but maintained at high expression levels in triple-negative tumors, inhibits Retinoic Acid which should normally inhibit growth and induce apoptosis(R.-Z. Liu et al. 2015). *GABRP* was already proven to be critical for TNBC cell growth(G. M. Sizemore et al. 2014) and its inhibition was reported to suppress basal-like BC progression(Cao et al. 2018). Likewise, Powell et al. reported that the majority of breast carcinomas that stain with *CALB2* are more likely to be high-grade, ER-negative, and display a basal-like phenotype(Powell, Roche, and Roche 2011). *TM4SF1*, as well, is known to be downregulated in hormone-positive tumors(J. Chen et al. 2022), while increased expression of *MMP7* distinguishes the basal-like BC subtype from other triple-negative tumors(S. T. Sizemore et al. 2014)(Kim et al. 2014). Indeed, Matrilysin is a validated target of several compounds that could be proposed to personalize BL1 TNBC. At the same time, *PGBD5* levels were found significantly higher in basal-like BC(Henssen et al. 2017), and the same goes for *CALML5*, one of the top expressed genes in TNBC samples(McQuerry et al. 2019), *PADI2*(McElwee et al. 2012) and *KLRG2*(Lim et al. 2020). Gong et al. demonstrated that the upregulation of *MGP* promotes the proliferation of cancer which probably makes it a novel biomarker or therapeutic target for TNBC patients(Gong et al. 2019). The same was also reported for *KRT16* by Lehmann et al., who showed its differential expression in the basal-like subtype(Lehmann et al. 2011), and confirmed by our Metacore analysis that revealed this basal cytokeratin as the predicted target of L-Triiodothyronine.

Two other predicted drug targets within the BL1 signature are *KCNQ4*, targeted by Bepridil and Fampridine, and CA8 encoding Carbonic anhydrase VIII and targeted by Foscarnet.

Among the seven down-regulated genes in *BL1*, *COL14A1*(Temian et al. 2018), *CYP1B1*(Abdul Aziz et al. 2021) and *ELN* had been previously associated with TNBC. The latter was considered in a TNBC genetic signature(Asztalos et al. 2015), in line with our findings. On the other hand, *HTRA1* was found to be significantly expressed within the breast normal ductal glands and its expression is significantly downregulated in IBC in general(N. Wang et al. 2012). Our study therefore confirms and specifies its down modulation in the BL1 subtype.

The BL2 subtype is mainly defined by the abnormal over-activation of several signaling pathways such as Wnt/β-catenin; indeed, one of the overexpressed genes found in our study is *WNT7B*, also reported by several studies in governing BC generally and TNBC more specifically(Dey et al. 2013). Through the latter, another BL2 gene (*WLS*) promotes the proliferation of BC cells(D. Zheng et al. 2020). In terms of *S100A9/8*, Bergenfelz was the first to report that it can be considered as a novel therapeutic target for patients with ER(-) PgR(-) BC (Bergenfelz et al. 2015) followed by several other studies(Bao, Wang, and Mo 2016). Indeed, our Metacore analysis identified Calgranulin B, encoded by *S100A9*, as the predicted target of Paquinimod as well as of Tasquinimod. Gene expression studies have previously identified *KRT5* mRNA in normal breast and basal-like BC and monoclonal antibodies against *KRT5* have been used to identify basal-like TNBC(Ricciardelli et al. 2017). This basal cytokeratin has been identified as a predicted target of Androstanolone by our analysis, however it is widely expressed in normal gland structures such as salivary and sweat glands and therefore targeting it may be critical. Previous findings indicated that *CRABP2* promotes invasion and metastasis of ER^- breast cancer. No studies to date have demonstrated the direct involvement with the BL2 phenotype of *CPA4*(Y. Wang et al. 2021), *TMEM45A*(Flamant et al. 2012), *S100A16*("Roles of S100 Family Members in Drug Resistance in Tumors: Status and Prospects" 2020), *COL4A5*, *GSDMC*, *MMP3*, *ITGB6*, or *GJB2*(Y. Liu et al. 2019). However, our drug interaction analysis revealed that *GJB2* is a predicted target of beta-Cyclodextrin.

On the other hand, Kloten et al. reported the loss of *NDRG2* protein expression in human BC and low *NDRG2* immunoreactivity in TNBCs(Kloten et al. 2016), which goes in line with the significant downregulation we found in the BL2 subgroup. *SORBS2*, another gene downregulated in BL2, is a tumor suppressor that was reported by Alsafadi et al. as a candidate marker to predict metastatic relapse in BC(Alsafadi et al. 2011). In terms of *PADI2* gene, we found that -as mentioned before- it is significantly highly expressed in BL1 subtype, contrary to BL2 subtype where it is significantly lowly expressed. Therefore, it can be proposed as a potential biomarker for differential diagnosis within the basal-like TNBC tumors. This has also prognostic implications as the BL1 subtype showed a significantly higher response rate to chemotherapy than the BL2 (Lehmann et al. 2016)(Echavarria et al. 2018).

The mesenchymal-like subtype (M) is mainly defined by a variety of signaling pathways, such as extracellular matrix–receptor interactions and gap junctions, which can explain the differential overexpression of *DSG3* compared to the other subtypes(Rötzer et al. 2015). The latter operates by facilitating cancer cell growth and invasion by controlling E-cadherin-Src signaling and cell-cell adhesion. The same goes for *COL9A3*(Del Bano et al. 2019), which is involved in matrix synthesis and controls its degradation. It was also identified as significantly associated with the prognosis of TNBC in an independent prognostic signature(Lv et al. 2019). *MSLN* has been explored by several studies and found to promote epithelial-to-mesenchymal transition and tumorigenicity(Koopmans and Rinkevich 2018). This can explain its overexpression in this particular TNBC phenotype as also reported by Del Bano et al.(Del Bano et al. 2019). *ID4* was reported to be highly expressed in TNBCs by Donzelli et al.(Donzelli et al. 2018) and it acts as an oncogene. Shen et al. found that the majority of ER-negative BC cells expressed moderate to high levels of *KCNK5* protein, whereas minimal/low levels of *KCNK5* were detected in ER-positive cells; as also confirmed by Alvarez-Baron et al.(Alvarez-Baron et al. 2011). *SOX6* has also been investigated by Mehta et al. who found it had an emerging role in BC development and maintenance as well as an involvement in the mesenchymal phenotype(Mehta, Khanna, and Gatza 2019). A set of genes found to have a promoter and primordial role in TNBC related epithelial to mesenchymal transition includes:

EPH(R.-X. Li, Chen, and Chen 2014), *EDIL3*(Gasca et al. 2020) and *TRIM29*. On the other hand, no analysis has explored *MDFI*, *CSPG4*, *CP*, *LAMB3*, *RNF152*, *BAMBI*, *SERTAD4* and *SFRP1* to show their involvement in promoting the mesenchymal phenotype of TNBC, while *ILRG* is a predicted target of Nedocromil.

As for the immunomodulatory subtype (IM), mainly enriched in immune cell markers and signaling, it turned out that all the genes overexpressed in this subtype, according to our analysis, are involved in the tumor immune infiltrate: *CD79A*(Z. Liu et al. 2018), *CXCL10*(Chuan, Li, and Yi 2020), *CXCL9* (Liang et al. 2021) which proved to be a potential biomarker of immune infiltration associated with favorable prognosis in ER-negative BC; *GZMB*("Immunotherapy for Triple-Negative Breast Cancer: A Molecular Insight into the Microenvironment, Treatment, and Resistance" 2021), *KLHDC7B*(Beltrán-Anaya et al. 2019), *LTF*(Chiu et al. 2020), *GBP5*(Cheng et al. 2021) and *CXCL11*(Narita et al. 2016), which was found to be significantly overexpressed in the plasma of BC patients compared to healthy controls; *LAX1*, which was reported by Mamoor et al. as associated with survival in TNBC; *IKZF3*, which contributes to the Immunologic Phenotype of TNBC(C. I. Li et al. 2021). A very recent study showed the prognostic value of tumor-infiltrating B lymphocytes along with *CD38* and plasma cells in TNBC(Kuroda et al. 2021). All the remaining genes have been confirmed to be associated with immune induced pathways along with breast cancer, but not specifically triple negative, thus contributing to a better refinement of TNBC.

Regarding the mesenchymal stem-like subtype (MSL), by definition it expresses low levels of cell proliferation-related genes, and high levels of stemness-related genes(D.-Y. Wang et al. 2019). This is supported by the genes we found downregulated, such as *CDK1*, or overexpressed , such as *IGF1* (Farabaugh, Boone, and Lee 2015) and *IGF2*(Tominaga et al. 2016), as well as *CXCL14*(Sjöberg et al. 2016). The long non-coding RNA *MEG3* is generally down-regulated in BC but it has been found highly expressed in Hs578T TNBC cells(Deocesano-Pereira et al. 2019). Conversely, *ID4* and *MDFI* are highly expressed in the M subtype but down regulated in the IM subtype. On the other hand, *CALML5* is overexpressed in BL1 but down-regulated in the MSL subtype. Ehmsen et

al. reported that *S100A14* is overexpressed in epithelial-like, but not in mesenchymal-like phenotype(Ehmsen et al. 2015), which converges with our findings.

The LAR subtype, even though it does not express the ER receptor, shows highly activated hormonal-related signaling pathways. Lehman et al. reported that tumors within the LAR group expressed numerous downstream *AR* targets and coactivators such as *ALCAM*, *FASN*(Lehmann et al. 2011), which were both contained in our LAR-related signature. We found that six of the up-regulated LAR genes, among which AR it-self, are experimentally validated druggable targets of up to thirty existing compounds. However, AR targeting in TNBC(Brumec et al. 2021)(Mina, Yoder, and Sharma 2017) has not achieved so far the expected efficacy. In an inverse perspective, Bhattarai et al.(Bhattarai et al. 2020) suggested a new refinement of the classification of TNBC by introducing Quadruple-negative BC based on AR expression negativity.

Conclusion

Our study took full advantage of available TNBC datasets to stratify samples and genes into distinct subtypes, according to gene expression profiles. The development of a data mining approach to acquire a large amount of information from several datasets, has allowed us to identify a well-determined number of genes that may help in the recognition of TNBC subtypes. This small number of genes can be tested in the clinics without the need of whole transcriptomic approaches. Most of the signature genes have been previously found to be associated with TNBC and/or have the potential to become novel diagnostic markers and/or therapeutic targets for specific TNBC subclasses.

Potential implications

Overall, our refined genetic signatures for each TNBC subtype may provide a simple clinical tool, affordable by most pathology departments, that might contribute to explore TNBC heterogeneity and identify the appropriate treatment for each patient based on the subtype-specific druggable targets. Novel clinical trials taking into account the molecular portrait of the tumor are in fact under development, for TNBC as well.

Chapter 2: Ki-67 proliferation index to further stratify invasive breast cancer molecular subtypes: Northern African comparative Cohort-study with external TCGA-BRCA and METABRIC validation

Introduction:

Globally, BC is the most common cancer in women, with approximately 2.2 million new cases diagnosed in 2020 (11.7% of all cancers, both sexes, all ages included). Its incidence rate varies significantly between regions of the world, with its peak in Asia followed by the central, eastern, and western parts of Europe, North and Latin America and Africa (The Global Cancer Observatory. Globocan 2020). Mortality within the African continent ranges from 5090 events in Southern Africa as the least concerned region, to 25626 in Western Africa, making it the most concerned region of Africa by this type of cancer.

In 2020, 11747 new BC cases were recorded in Morocco, representing 19.8% of all cancers in women and the first diagnosed cancer. It is the first also in terms of mortality (3695 estimated deaths) and prevalence (31420 cases for 5-years prevalence) (The Foundation Lalla Salma Cancer prevention and treatment. DETECTION GUIDE of EARLY breast and cervical CANCER. 2011 Edition. 2011).

The cancer registry of the greater Casablanca region (2016-2020), according to the latest report developed by the Department of Epidemiology and Disease Control of the Ministry of Health, estimates the frequency of BC at 35.8%, with a peak recorded between 55 and 59 years (RCRGC. CANCER REGISTER).

It is therefore clear that BC is the first female cancer, making it a public health problem in Morocco, as well as around the world.

It is currently estimated that one out of 9 women will develop BC in her lifetime and one out of 27 will die from it, underscoring the importance of this disease in terms of public health. It should be noted that men can also develop BC. These cases are however rare, since they represent only 1% of breast carcinomas (Yalaza, İnan, and Bozer 2016).

Currently, mammography is the best way to detect BC in its early stages. On average, the tumor can be detected 1.7 year before a woman experiences a lump. In the early stages of a localized tumor, chances of survival at 5 years are 95%. These chances decrease in late stages : they are less than

50% when the tumor has disseminated in the lymph nodes and less than 20% when it has disseminated in distant organs. The early detection of cancerous lesions, surgery, with selective removal of the tumor, and various therapies (chemotherapy, radiotherapy, hormone therapy or other targeted therapies) have contributed to significantly reducing mortality due to BC. However, despite these advances, some aggressive and metastatic types of BC are difficult to treat and still remain incurable.

It is then very important to define the full panoply of biomarkers influencing the survival of patients with BC. Statistical methods may help select the best combination of biomarkers to use in order to predict survival and prognosis (Vickers and Cronin 2010). Several studies have been previously carried out using conventional statistical techniques that are limited in terms of generating clear and creative visualizations of the results obtained by the analysis of these factors (Rajula et al. 2020). The limitations of these statistical techniques may have enabled clinicians to use other much robust and deep ML techniques, such as decision trees (DT), Naive Bayes (NB); Generalized Linear Model (GLM) ; Random Forest (RF), Fast Large Margin (FLM) ; Deep Learning (DL); Gradient Boosted Trees (GBT); Support Vector Machine (SVM) (Dubey, Gupta, and Jain 2015);(Hou et al. 2020; Ganggayah et al. 2019).

We assessed and evaluated the same prediction techniques mentioned above on a Moroccan patients dataset with 1266 BC records in this 5-year-follow-up retrospective study.

In this work, we applied those statistical ML techniques on a large cohort of BC patients to explore whether the molecular classification accepted as a reference for the determination of BC subtypes can be refined by statistical partitioning methods. This is especially useful in low- and middle-income countries (LMIC) such as Morocco, where laboratories do not necessarily have wide access to new molecular methods which are proving to be very expensive like the gene expression profiling assays. In addition, our work also consists of knowing whether these outcomes can be reproducible at a certain degree of accuracy.

Every ML technique has some advantages. DT for example, allows the best visualization and illustration of results in the form of a well-represented and easily interpretable DT, something which

is greatly sought when considering a large number of samples. It also provides a set of easy-to-interpret decision rules that are necessary for decision makers (Pathologists, clinicians,) to make timely and appropriate decisions about BC prognosis prediction(Song and Lu 2015). Also, one of its greatest advantages is its ability to handle various types of data (Continuous, categorical, cardinal variables...etc.).

As for the RF algorithm, it is a nonparametric statistical method requiring no distributional assumptions. It has the ability of handling a large number of features and estimates the importance of variables used for classification. It appeared therefore to be a suitable classifier without any over-learning. In general, it has a better performance than decision trees by calculating the direct "Out-of-Bag" error which replaces the cross validation (Breiman and al. 2001).

Finally, GLM was chosen as an appropriate algorithm for binary variables, allowing to evaluate the model accuracy using binary values (Boughorbel, Al-Ali, and Elkum 2016);(Ropo Ebenezer and Lougue 2019).

However, it is necessary to point out that studies on BC using ML techniques have already been developed before by several authors, but the factors studied vary from one study to another, depending on the target population, its geolocation, its lifestyle, the available databases and even the purpose of the study (Ropo Ebenezer and Lougue 2019; “Website” n.d.; Yassin et al. 2018; Dhahri et al. 2019).

We therefore concluded that it is necessary to develop a model for the African and the LMIC context, a model that has never been studied before and, more precisely, in Morocco, in order to study the variables that can govern the survival rate of Moroccan patients with BC through the histo-prognostic indicators usually analyzed routinely in all pathology laboratories. We would also be interested subsequently in the technique of selecting the most relevant variables by using these same ML techniques in the medical field.

The main aim of this study is to explore any further subdivisions that can be found depending on Ki-67 value distribution within BC tumors, to identify the histo-prognostic features that can drive

the survival prediction of BC Moroccan patients, and discover the most influential and important features according to different algorithms of ML techniques.

Material and methods

1) Study design and setting:

This is a comparative retrospective Cohort-study including Moroccan BC patients with 5years of follow-up, the TCGA-BRCA and METABRIC datasets with 13 and 30 years of follow-up, respectively.

All IBC records in the mentioned databases were included. In contrast, benign tumors or tumors of uncertain malignancy; tumor recurrences; BC in men; patients with incomplete/equivocal immunohistochemical status were all discarded.

2) Collected datasets:

a) TCGA-BRCA dataset

It was retrieved from "https://portal.gdc.cancer.gov/" and was initially composed of 963 IBC records. The dataset contains the following features: ESR1, PGR, ERBB2 and MKi-67 (genes expression Z-scores), Lymph Nodes stage, Neoplasm Disease stage, Menopause status, tumor stage, altered genome fraction, metastatic stage, Cancer Type, Sample initial weight, mutation count, micrometastasis detection, ethnicity category, histologic type, race category, Lymph Node Ratio (LNR), OS Status, OS period (over 13 years of follow up), Disease Free period, Disease Free Status, Diagnosis stage, Positive Lymph Node Count, Lymph Node Examined Number. After missing values filtering, the final number of remaining BC records was 625.

24 patients were recorded as Hispanic or Latino; 458 as non-Hispanic or Latino and the status was missing for 143 patients. As their race category: 428, 70, 38,1 were recorded as white, black, or African American, Asian respectively and the race data was missing for the remaining 88 patients.

b) METABRIC dataset

It contains 1885 BC records and was retrieved initially from: https://www.cbioportal.org/. It contains the following histoprognostic features: MKi-67 (genes expression Z-score); ER/PgR/HER2

(Overexpressed / Underexpressed); Age at diagnosis (years); Cancer type; Cellularity; Chemotherapy; Neoplasm histologic grade; HER2 status measured by SNP6; Histologic subtype; Hormone therapy; Inferred menopausal state; primary tumor laterality; Nottingham prognostic index; Radio therapy; Tumor size; patients vital status; Lymph nodes examined positive; mutation count; OS time (over 30 years of follow up); survival status (Censored/ Dead). There were no missing values in this dataset. No social or demographic feature was present in this dataset.

c) Moroccan dataset

General and clinical data of all IBC recorded from January 1st, 2013, to March 30th 2018 at the Pathology Department of Ibn Rochd University Hospital of Casablanca were retrieved, which led to 1266 Moroccan patients with IBC and 165 of them were followed at the King Mohammed VI National Centre for the Treatment of Cancers, where their 5-years follow-Up survival data were collected (from their corresponding Medical Records from the National Population-based Cancer Registries). This is considered the largest public hospital in Morocco with the largest cancer registry. The patient was considered Dead if the death was confirmed on a predefined date. Or confirmed Alive, by her attending physician at the last date of follow up. No demographic/social information was found in the national registry. The histo prognostic features collected are: age at diagnosis; tumor size (TS), lymph node infiltration (NI), SBR grade; Oestrogen receptor (ER); Progesterone receptor (PgR); Ki-67 proliferation index and HER2 receptors status by immunohistochemistry, the absence / presence of vascular emboli (VE), histologic type, TNM staging, first and last dates of follow up, state at the last date of follow up.

These datasets were specifically chosen given the large number of BC samples they contain. The Moroccan dataset was used primarily as an internal dataset on which the analysis focuses. METABRIC and TCGA-BRCA datasets were used to serve as external validation sets publicly available.

3) Immunohistochemistry and scoring

Samples:

The Moroccan study was carried out on breast samples corresponding to resection specimens (mastectomy, lumpectomy) or biopsies.

Tissue preparation

The specimens undergo conventional histopathology techniques.

Fixation

Fixation is done by immersion in a 10% formalin solution. The fixation duration depends on the size of the sample.

Macroscopic analysis (resection specimens)

The tissues are examined, measured and weighed. The representative parts of samples are chosen from the pathological tissue and are put in codified cassettes.

Rinsing and dehydration:

After fixation, the samples are rinsed with running water and then with distilled water to remove excess fixative. Dehydration occurs by passing the samples through successive baths of titrated alcohol croissants (70 °, 95 °, 100 °). Dehydration is followed by impregnation of the tissue with a paraffin solvent: toluene.

Waxing:

Impregnation is carried out by passing the tissues through two successive paraffin baths heated to 56 °C.

Inclusion:

The samples are embedded in paraffin within a stainless-steel mold. After cooling, firm paraffin blocks are obtained containing the processed tissues.

Microtome sectioning :

The paraffin blocks containing the tissues are cut into sections 4 to 5 µm thick using a microtome. The cuts obtained are glued to slides using a 1% albumin solution. The biopsy code is attributed to the blades thus obtained.

Deparaffinization and rehydration:

The slides are placed in the oven for better adhesion to slides at 65 ° C for 30 min. Then they are included in two baths of toluene (4 minutes each) for dewaxing. This step is followed by rehydration by immersion of the slides in alcohol baths of decreasing degree (100 ° up to 70 °) then in distilled water.

Staining:

The sections are stained with Hematein (for 1 min), rinsed, and then stained with Eosin (for 2 min) then rinsed. Staining is automated.

Dehydration and assembly

Dehydration is done by immersing the blades in alcohol baths in increasing amounts then in two toluene baths (2 min each). The sections are then mounted between slide and coverslip using a synthetic resin.

Immunohistochemistry

The immunohistochemical technique makes it possible to complete the data of the HE staining of suspected tumors. It allows to visualize in situ the hormonal and HER2 receptors and assess their status.

Principle of the immunohistochemical technique

The immunohistochemical technique involves locating antigens in tissues using a specific antibody. The specific antibody binding site can be visualized using a tracer attached to the antibody.

Experimental protocol

For the demonstration of the expression of receptors, we used the Herceptest kit (DAKO) which contains the primary antiHER2 antibody, IR657 Monoclonal mouse anti-human Oestrogen Receptor antibody for ER, IR068 FLEX Monoclonal Mouse anti-Human Progesterone Receptor for PR and IR626 FLEX Monoclonal Mouse anti-Human Anti-Ki-67 Antigen for Ki-67. While the visualization is done using the Dako EnVision ™ detection kit. The immunohistochemical technique is carried out using the PT Link automaton and the Autostainer Link stainer. It is preceded by dewaxing and rehydration, and is performed in several stages, as follows.

Heat Induced Epitope Retrieval (HIER)

The dewaxed sections are immersed in the unmasking solution preheated to 65 ° C then incubated at 97 ° C for 20 min. The sections are then cooled for 20min in the same solution up to 65 ° C then rinsed with a buffer.

Blockage of endogenous peroxidase:

Sections are treated with the blocking peroxidase reagent (100 µl per section for 5 min). After rinsing, the slides are placed in a buffer.

Application of the primary antibody:

Each section is treated with 100 µl of primary antibody. The incubation, which lasts 30 min, is followed by rinsing with a wash buffer.

Visualization of the marking:

Each section is treated with 100 µl of Dako EnVision FLEX / HRP visualization coupled to a secondary antibody. Incubation, which lasts 30 min, is followed by a rinse with a wash buffer.

Application of the chromogenic substrate solution (DAB):

The sections are each coated with 100µl of the chromogenic substrate solution, incubated for 10 min then rinsed with distilled water.

Counterstain:

The sections are counterstained by immersion in a Hematoxylin bath for 2 min followed by a rinse with distilled water.

Dehydration and assembly:

Sections dehydrated by passing through toluene are mounted between blade and coverslip using synthetic resin. IHC was performed using the PT Link controller and the Autostainer Link controller. Afterwards, each section was treated with 100µl of Dako EnVision FLEX/HRP visualization reagent coupled to the secondary antibody. The slides were then examined under a LEICA DM1000 optical microscope.

Scoring and histological subtyping:

The pathological evaluation and interpretation of the staining was carried out by the same team of four pathologists at the Pathology Department of ibn Rochd University Hospital, and the final scoring was given by a consensus, leading to consistent pathological reporting.

The histological subtyping and SBR grading were assessed in concordance with standard guidelines. Scoring is based on the ASCO/CAP recommendations for ER and PgR (Harbeck, Thomssen, and Gnant 2013) considering any nuclear staining in at least 1% of invasive tumor cells as positive and according to 2018 ASCO/CAP recommendations for HER2 (Turashvili and Brogi 2017). As for Ki-67, the cut-off was set at 20%, one of the easiest levels of staining to characterize (Turashvili and Brogi 2017; Hammond et al. 2010). The results for ER, PgR and Ki-67 were recorded therefore as the percentage of immunoreactive cells over up to 2000 neoplastic cells.

4) BC molecular classification

For all the datasets, Ki-67, ER, PgR and HER2 variables were extracted for further analysis. Subsequently, BC was systematically classified into five intrinsic subgroups, as follows:

- ➔ LuminalA (LumA): ER+ and/or PgR+; HER2- ; low Ki-67
- ➔ LuminalB HER2+ (LumB HER2+): ER+ and/or PgR+; HER2+; high Ki-67
- ➔ LuminalB HER2- (LumB HER2-): ER+ and/or PgR+; HER2-; high Ki-67
- ➔ Pure HER2: ER- and PgR-; HER2+; irrespective of Ki-67
- ➔ Triple Negatives (TN): ER- and PgR-; HER2-; irrespective of Ki-67.

5) Pre-processing

We first assessed a cleaning step, consisting in normalizing all numeric variables to improve the performance of algorithms that use weighted inputs or distance measurements and make them comparable.

As for METABRIC and TCGA-BRCA dataset which contain z-score variables, positive ones were considered with high expression of MKi-67, ER, PgR and HER2; in contrast, variables with negative z-scores values were considered underexpressed. All missing data were excluded, and their corresponding rows were deleted from the datasets. Hence the filtering of any patient showing at

least one missing value for one of the four following variables: ER, PgR, Ki-67 and HER2 which are the principal features of our interest.

6) Statistical partitioning

Mainly assessed by “STATISTICA software, v10”.

7) Determination of the optimal number of clusters

-Quality indices:

Assed by NbClust package under R software. This library helps determine the right number of classes in both datasets. Thirty proposed validation indices, as well as hierarchical and non-hierarchical classification methods were assessed in order to determine in the most impartial and objective manner the optimal number of clusters. Its advantage is the simultaneous application possibility of several indices to find the best score among all the scores obtained.

-Validation measures:

Validation measures were assessed by the “clValid package” in R, that helps to simultaneously select multiple clustering algorithms, validation metrics, and cluster counts in a single function call, in order to determine the most suitable method and optimal number of clusters.

-Cluster stability measures include: The figure of merit (FOM); The average distance between means (ADM); The average proportion of non-overlap (APN); The average distance (AD). The values of the first three measures range between 0 and 1, with smaller values corresponding to highly consistent clustering results. AD measures range between 0 and infinity, and smaller values are also preferred.

-Clusters Internal Validation measures. The Dunn Index and Silhouette Width are both examples of the compactness and separation; with the connectivity, they comprise the three most important internal measures. As an input, internal validation metrics require the dataset and the clustering partition and use intrinsic information in the data, to assess the quality of the clustering. Stability measurements therefore make it possible to assess the consistency of a clustering result by comparing it to the clusters obtained after removing each column, one at a time.

8) Prediction models for clusters membership

This step was assessed by “RapidMiner Studio software v9.9” which splits the entire Moroccan database in two subsets before running the prediction models.

Eight prediction algorithms were applied, which are a set of ML techniques that search for patterns in sets and use those patterns to predict new records : Naive Bayes (NB), Generalized Linear Model (GLM), Fast Large Margin (FLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT) and Support Vector Machine (SVM). And the used predictor variables were: ER, PgR, HER2 and Ki-67

To explore which model showed the strongest predictive ability, we evaluated them by nine metrics to be able to extract the most robust model: Accuracy, Area Under the Curve Receiver Operating Characteristic (ROC-AUC), Precision, Recall, F-measure, Sensitivity, Specificity, Classification error and Total running time of the model. The Cross-Validation option was used, which helps assess the model's ability to classify new data, to avoid problems like overfitting, and estimates how accurately this predictive model performs.

9) The k-folds Cross-Validation:

This option was used for all the prediction models, as it helps assess the model's ability to predict new data that was not used in estimating it, to avoid problems due to overfitting. It estimates how accurately the predictive models will perform. The cross-validation operator assesses two sub-processes: the first is called training and is used to train a model, the latter is then applied in the Test sub-process. Thus, the performance of the model is measured during the test phase.

The principle behind is to partition the input set into k-subsets of equal size. Of the subsets, only one subset is kept and labeled as a test data set. The remaining k-1 subsets are used as the training data set. The cross-validation process is then repeated k-times, each of the k-subsets being used exactly once as test data. The k results of the k-iterations are averaged (or otherwise combined) to produce a single estimate.

Evaluating a model's performance on independent test sets gives a good estimate of performance on new data sets. It also shows if an "overfitting" is occurring. This means that the model has been trained well, and therefore represents the test data very well. But it may not be well generalized on all new data. Thus, the performance can be much worse on the test data.

As explained previously, the validation of a prediction model requires

-to separate an initial dataset into a training dataset and a validation dataset,

-to calibrate a model with the training set and

-to assess the quality of the model with the validation data set by calculating the indicators defined previously.

To ensure that the quality of the model is well measured, these three steps must be repeated several times. This is because the initial dataset can be separated into a very large number of pairs of training and validation datasets.

For example, in the first iteration, the 70% observations that will be used to calibrate the model will be different from the 70% observations in the training dataset in the second iteration.

The validation metrics can be averaged over all the iterations to have a better characterization of the model. This concept is known as cross-validation because the goal is to find out whether the validation of the model is equivalent when different training and validation datasets are used.

If the validation dataset is set to contain 20% of the observations and several iterations are performed, we say that we have done a 20-fold cross validation. If the validation set includes only 5% of the observations, it is a 5-fold cross validation. Another method often reported is the leave one out cross validation. As the name suggests, validation is done with just one observation.

Validating a prediction model is necessary to ensure that the model is indeed able to accurately predict the values of a variable of interest. Once again, it should be understood that the model will be evaluated the better the more independent the training and validation datasets are. It often happens that people use very consistent training and validation datasets to boost their validation indicators (some even validate their model on the data that were used to calibrate their model ...). It

is recommended to perform these cross-validation procedures and to calculate the validation indicators so that the prediction models are correctly evaluated.

10) Important Variables Selection

-VIMP algorithm

Mainly based on RF model to test the prediction and ranks of the most important variables according to their impact on the predictive ability of the forest. A VIMP value equal or close to zero indicates that the variable does not contribute to the predictive accuracy; on the other hand, negative values indicate that the predictive accuracy improves when the variable is misspecified. Accordingly, we used the "gg_vimp" function associated with the "ggRandomForests" package in R software, which we used to essentially extract VIMP measures for each of the variables used to grow the RF.

-Minimal Depth

It assumes that variables with high impact on the prediction are those that most frequently split nodes nearest to the root node, therefore, the largest samples of the population. Node levels in every tree are numbered based on their relative distance to the root of the tree (which is indicated by level 0). The assumption is that smaller minimal depth values indicate that the variable has a large impact on the forest prediction. The latter was elaborated with the “randomForestSRC” package.

11) Survival analysis

Surviving patients were censored at the date of last follow-up; survival plots according to BC molecular subgroups were drawn using the Kaplan–Meier method to evaluate their OS rates. The log-rank test was applied to assess the survival difference between strata.

12) Prediction models for survival analysis

The survival dataset contains 165 records from the original total dataset for which all the survival information was available. The quality of this dataset was then evaluated by 4 different prediction algorithms: Decision Tree (DT), Random forest (RF), Neural Network (NN), and Generalized

Linear Model (GLM). They were evaluated by five metrics: Accuracy, specificity, sensitivity, ROC curve and Kappa coefficient.

The prediction models were built by RapidMiner Studio online tool, and all the other statistical analyses and processing were developed by R software.

13) Statistical tests used:

To better study how each explanatory variable is governed by the other variables, we used bivariate statistics. A bivariate study between two variables makes it possible to determine the degree of association between these two parameters with a certain threshold of significance. The bivariate analysis was assessed by :

-Kruskal-Wallis test: A non-parametric alternative of ANOVA, except that it is based on ranks instead of means. It is used to compare at least three samples, and to test the null hypothesis according to which the different samples to be compared come from distributions with the same median.

-The G-square test of independence: was used as an alternative of Chi-2 when we have two nominal variables and the aim was to see whether the proportions of one variable are different for different values of the other variable.

-The Mann-Whitney Wilcoxon test: a nonparametric statistical test which tests the hypothesis according to which the medians of different groups are the same.

All tests were two-sided, and we considered Bonferroni adjusted P-value 0.05 as the threshold to declare significance.

14) Softwares and online tools used:

- -STATISTICA from StatSoft, version 10 was exclusively used for EM clustering.
- -WEKA for data mining version 9.9, used for Prediction models and their evaluation.
- -The IBM® SPSS® software , version 20, mainly used for univariate analysis.
- -Exploratory Inc software, version 6.4.1, mainly used for better, clear and personalized figures.

- -R software for Biostatistics and R studio software, used for the entirety of this research work.

Results:

In the first section, we will assess a general overview of the Moroccan dataset, to become more familiar with the variable's distribution. Second, we will address the diverse possible subdivisions that can be found within the routinely established molecular classification based on immunohistochemistry that may govern the distinct prognosis and 5-year OS outcome of patients that belong to them. Finally, in the second section, we will approach the preliminary results of the comparative retrospective Cohort-study concerning Moroccan BC patients validated externally by TCGA-BRCA and METABRIC datasets.

This insight is interesting insofar as it highlights a new refinement of the BC molecular classification and can improve it. Also, this is the first study of its kind in the African context and which extends to an OS analysis of the BC patients depending on their cluster membership.

Section1 : General overview of the moroccan population

After all missing values abstraction, we collected clinical information of 1266 IBC scattered in 566 pre-surgical core biopsies (44%) and 700 surgical specimens (55%). The most frequent histological type is the No Special Type (NST), representing 97% of all records.

The patient's age ranges from 17 to 97 years. 50% of the population is concentrated in the age group between 43 and 58 years old. While the least affected age group is between 58 and 97 years old.

The expression of ER, PgR, HER2 and Ki-67 proteins was positive in 67.85%; 60.9%; 29.85%; 61.84% respectively (figure 28). On the other hand, the SBR classification showed that 222 of the patients had a grade I (17,53%), 556 of the patients had a grade II (43,91%) while 488 had a grade III (38,54%).

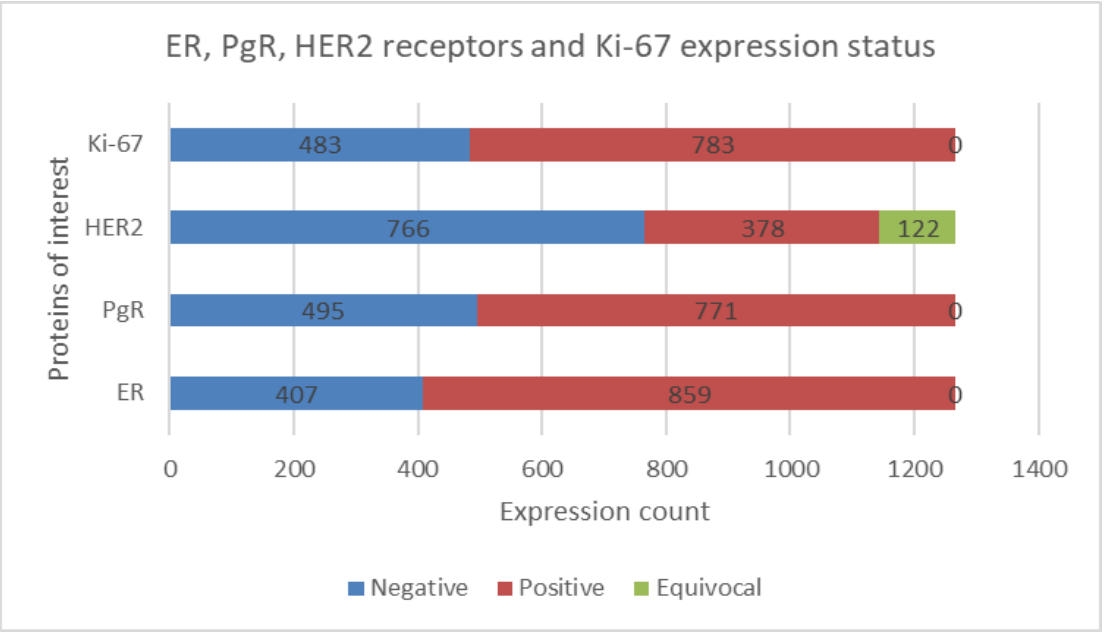

Figure 28: expression status counts according to ER, PgR, HER2 and Ki-67

As reported in table 12, in terms of ER+PgR+ tumors: 67% are HER2- while vascular emboli and lymph nodes are likely to be absent. As the HER2 overexpression increases, the SBR grade severity is also increasing from I to III. Similarly, in ER+PgR- tumors, the grade III increases from 19% to 33% as HER2 immunohistochemistry score increases from 0+ to 3+ and vascular emboli and lymph node invasion become more pronounced.

Conversely, ER- tumors have a remarkable increase in KI-67 index compared to tumors in which ER is expressed (ER+PgR+ and ER+ PgR-). Vascular emboli are less present in hormone-dependent tumors (ER+PgR+, ER+PgR- and ER-PgR+) and the SBR grade is higher for ER-PgR- and ER-PgR+ tumors (grade III).

Table 12: The distribution of patients according to their tumors phenotypes (A: Absent, P: Present)

	HER2	BC type	Ki-67		Age	VE (%)	NI (%)	Grade (%)	Total	
			Low	High					n	%
ER+ PgR+	0+	LumA/B HER2-	50.8	49.1	51.51±10.43	A : 25.8 P : 12	A : 28.2 P : 25.8	I: 27 II : 47.6 III : 25.3	340	45.09
	1+		46.6	53.3	52.91±11.92	A : 18.2 P : 9.7	A : 23 P : 27.2	I : 24 II : 47 III : 27	165	21.88
	2+		46.1	53.8	52.18±11.65	A : 21.8 P : 5	A : 15.38 P : 24.35	I : 16.6 II : 52.5 III : 30.7	78	10.34
	3+	Lum B HER2+	40.3	59.6	49.11±11.85	A : 22.8 P : 17	A : 19.8 P : 17.5	I : 14 II : 53.8 III : 32.1	171	22.67
Total of patients with ER & PgR positive tumors			47.08	52.9	51.34±11.22	A: 23.07 P: 11.93	A: 23.87 P: 24.13	I: 22.4 II: 49.6 III: 27.9	754	59.55
ER + PgR-	0+	LumA/B HER2-	47.2	52.7	51.52±12	A : 19.4 P : 16.6	A : 25 P : 19.4	I : 16.6 II : 63.8 III : 19.4	36	34.28
	1+		55.5	44.4	45.72±10.15	A : 22.2 P : 5.55	A : 11.1 P : 33.3	I : 22.2 II : 38.8 III : 38.8	18	17.14
	2+		33.3	66.6	54.33±11.04	A : 8.33 P : 16.66	A : 25 P : 16.6	I : 8.3 II : 25 III : 66.6	12	11.42
	3+	Lum B HER2+	41	58.97	49.55±10.64	A : 10.25 P : 12.82	A : 5,12 P : 20.51	I : 7.7 II : 58.9 III : 33.3	39	37.14
Total of patients with ER positive & PgR negative tumors			44.7	55.23	50.12	A: 15.23 P: 13.33	A: 15.23 P: 22	I: 13.3 II: 53.3 III: 33.3	105	8.29
ER-PgR+	0+	LumA/B HER2-	28.5	71.4	51.57±5.5	A:28.57 P: 0	A:28.57 P : 42.85	I : 14.28 II : 28.57 III :57.14	7	41.17
	1+		33.3	66.6	49.67±2.08	A : 0 P : 0	A : 0 P : 33.3	I : 0 II : 0 III : 100	3	17.64
	2+		0	100	50.9	NA	NA	II	1	5.88
	3+	Lum B HER2+	16.6	83.3	40.33±6.02	A : 0.5 P : 33.3	A : 16.6 P : 0.5	I : 16.66 II : 0.5 III : 33.3	6	35.29
Total of patients with ER negative & PgR positive tumors			23.5	76.4	47.23±7.16	A:29.4 P:11.7	A: 17.6 P:41.1	I: 11.7 II: 35.3 III: 53	17	1.34
ER-PgR-	0+	Triple negative	24.51	75.5	50.95±11.5	A : 14.2 P : 14.2	A :25.8 P : 25.8	I : 11 II : 31.36 III : 57.4	155	39.74
	1+		33.3	66.6	51.54±10.48	A : 19 P : 9.5	A : 16.6 P : 26.2	I : 19 II : 30.9 III : 50	42	10.76
	2+		33.3	67.7	53.16±13	A : 12.9 P : 19.3	A : 25.8 P : 6.4	I : 6.4 II : 42 III : 51.6	31	7.94
	3+	Pure HER2	9.25	90.7	51.14±10.39	A : 13.3 P : 14.8	A : 19 P : 20.3	I : 6.17 II : 27.7 III : 66	162	41.53
Total of patients with ER & PgR negative tumors			19.7	80.2	51.27 ±11.04	A: 13.84 P: 14.35	A: 22 P: 22	I: 9.48 II: 30.7 III: 59.7	390	30.80
Total of all patients			38.1	61.8	51.16 ±11.12	A : 19.6 P : 12.8	A : 22.5 P : 23.53	I : 17.5 II : 44 III : 38.5	1266 (100%)	

Routinely, when HER2 receptor expression is found to have a 2+ score (considered equivocal), it is essential to carry out a more in-depth analysis by means of the FISH technique, which makes it possible to decide on the positivity / negativity and avoid its ambiguous status. In the Moroccan dataset, for a total of 122 HER2 IHC 2+ cases, FISH could not be assessed. This is the main reason for their exclusion from downstream statistical analysis.

Similarly, the ER-PgR+ subgroup was excluded, because it represented only 1.34% of the study population (17 patients) and there is much debate about the existence of this combination. That left us with a total of 1127 patients for downstream analysis.

Therefore, the new repartition of the moroccan set (excluding HER2 2+ and ER-PgR+ tumors as addressed above), depending on their HER2, ER and PgR profile and limited to 9 phenotypes is {ER + PgR + HER2 0+; ER + PgR + HER2 1+; ER + PgR + HER2 3+; ER + PgR-HER2 0+; ER + PgR-HER2 1+; ER + PgR-HER2 3+; ER-PgR-HER2 0+; ER-PgR- HER2 1+; ER- PgR- HER2 3+}, as shown in figure 29 below.

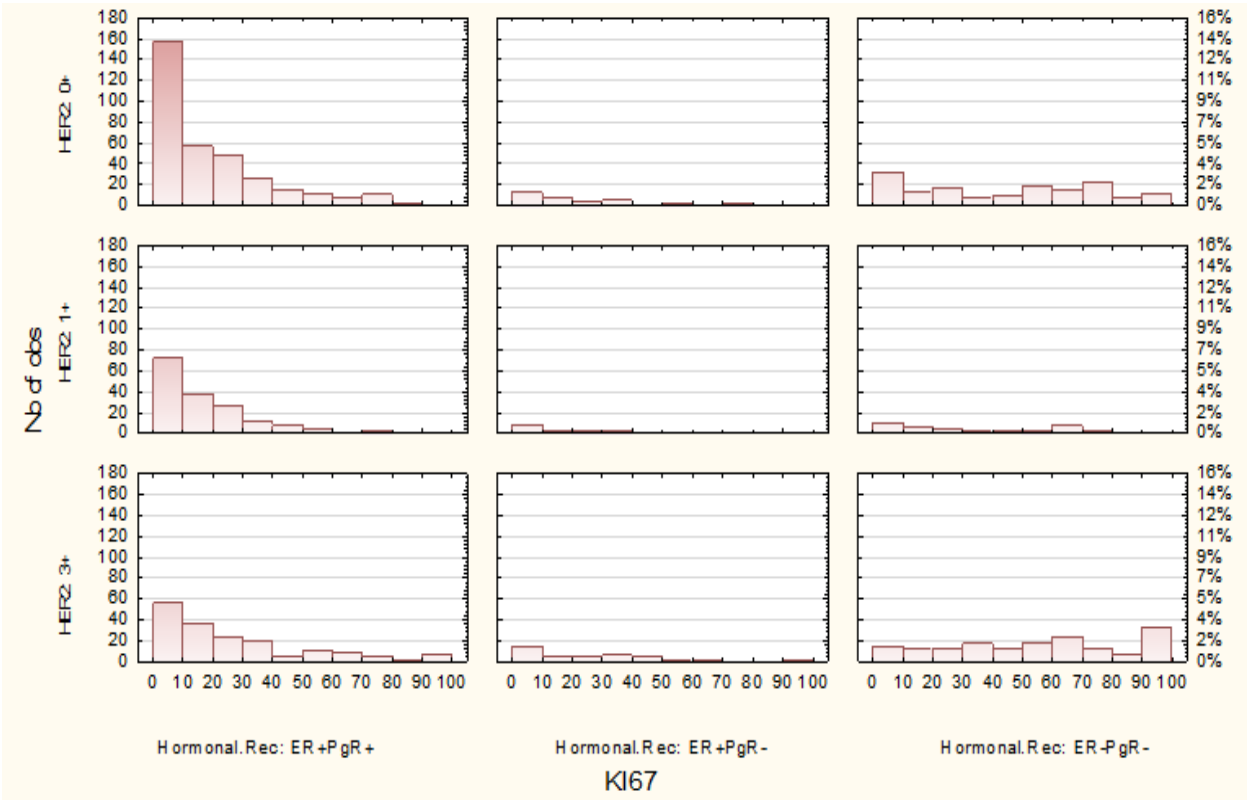

Figure 29: Histogram of Ki-67 immunohistochemical scores according to the 9 tumors phenotypes depending on their hormonal profile

(X-axis : Ki-67 immunohistochemical staining percentages, Y-axis :Number of observations depending on HER2 scores)

We find that the majority of (ER + PgR +, HER2 0+) tumors have low KI-67 proliferation index, and that this effect decreases as HER2 becomes positive (ER + PgR + HER2 3+). Similarly, tumors with ER-PgR- HER2 3+ profiles are much more enriched in high KI-67 than those with low expression of HER2. Thus, the KI-67 proliferation index increases with HER2 over-expression and hormone receptor under-expression.

As shown in the previous figure, the percentage of Ki-67 remarkably varies from one phenotype to another. In order to determine whether these intergroup variations are statistically significant in our set, the Kruskal Wallis test was used, a nonparametric test that can accommodate more than two groups whose observations are independent. The test's results are summarized in figure 30:

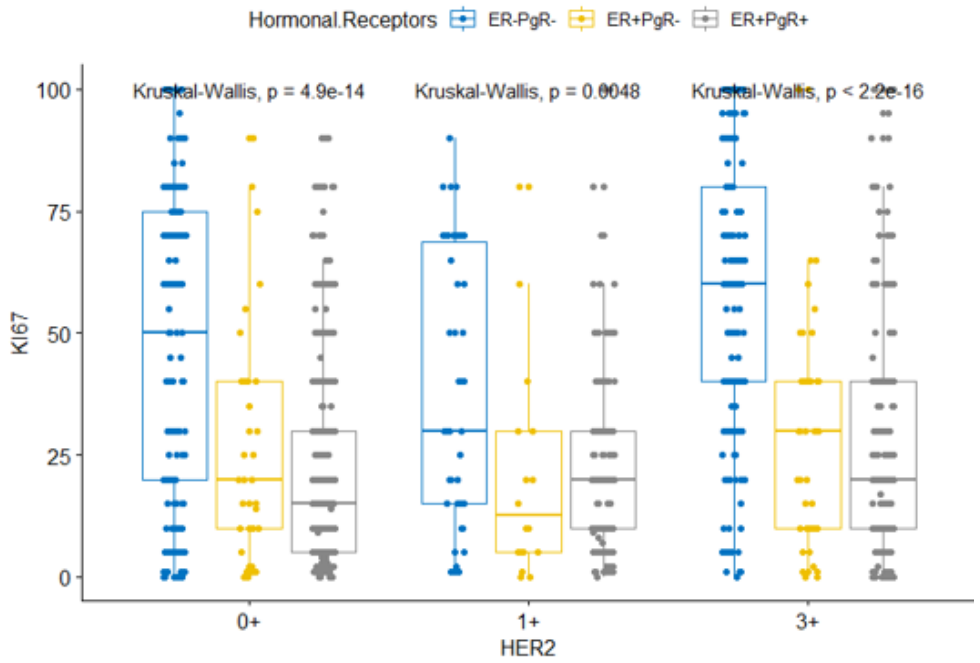

Figure 30: Inter-phenotypes comparison of Ki-67 expression distribution

The above-mentioned figure shows Ki-67 values (y-axis) according to the hormonal (strata) and HER2 phenotype (x-axis). The application of the Kruskal-Wallis test, which is used to determine whether the median varies significantly among the different phenotypes, confirmed this hypothesis. Therefore, we can conclude that:

- Ki-67 expression levels related to HER2 are most significantly different in tumors with HER2 3+ compared to tumors where HER2 is negative or very low (0 and 1+).
- KI-67 is higher in the ER-PgR- tumors, especially when HER2 is overexpressed.

- When ER is expressed exclusively (ER + PgR-), KI-67 is more prominent with HER2 overexpression than without.
- In the (ER + PgR +) phenotype, KI-67 is poorly expressed.
- Thus, Ki-67 is higher with HER2 overexpression and absence of hormone receptors. This confirms the inter-phenotypes heterogeneity that defines BC molecular classification.

A. Moroccan data clustering based on Ki-67; ER; PgR and HER2

In order to further explore the existence of an intra-molecular subgroup heterogeneity, we implemented the Generalized Estimation-Maximization cluster Analysis module in Statistica software, whose purpose is to detect clusters in observations and to assign those observations to the clusters. This method was chosen because it extends this basic approach to clustering in three important ways:

1. Instead of assigning cases or observations to clusters so as to maximize the differences in means for continuous variables (ER; PgR; Ki-67), the EM clustering algorithm rather computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm is to maximize the overall probability or likelihood of the data, given the (final) clusters.
2. Unlike the classic implementation of k-Means clustering, the latter can be applied to both continuous (ER; PgR and Ki-67) and categorical (HER2) variables.
3. A major shortcoming of k-Means clustering has been that there is a need to specify the number of clusters before starting the analysis (i.e: The number of clusters must be known a priori). Instead, the Generalized EM Cluster Analysis module uses a modified v-fold cross-validation scheme to determine the best number of clusters from the data. This extension makes the Generalized EM Cluster Analysis module an extremely useful data mining tool for unsupervised learning and pattern recognition. The unsupervised learning method was chosen because the outcome variable of interest (number of clusters) cannot be directly observed. Instead we want to detect some "structure" or clusters in the data that may not be trivially observable.

Hence the use of this specific method for this separate dataset. This hypothesis was tested on all 9 phenotypes in an unsupervised manner. During the EM clustering, the "v-fold cross-validation" algorithm used to automatically determine the appropriate number of clusters in our population was applied. We found that each phenotype was statistically divided into two further subdivisions as follows:

- ★ C1 (for Cluster 1): includes patients with a low Ki-67 proliferation index (16.26 ± 11.9 as mean percentage).
- ★ C2 (for Cluster 2): includes patients with a high Ki-67 proliferation index (68.8 ± 18 as mean percentage).

As shown in figure 31 and as long as ER is expressed, C1 was statistically found to be the most frequent. C2 becomes the most frequent when there is no expression of hormone receptors, especially when HER2 is positive.

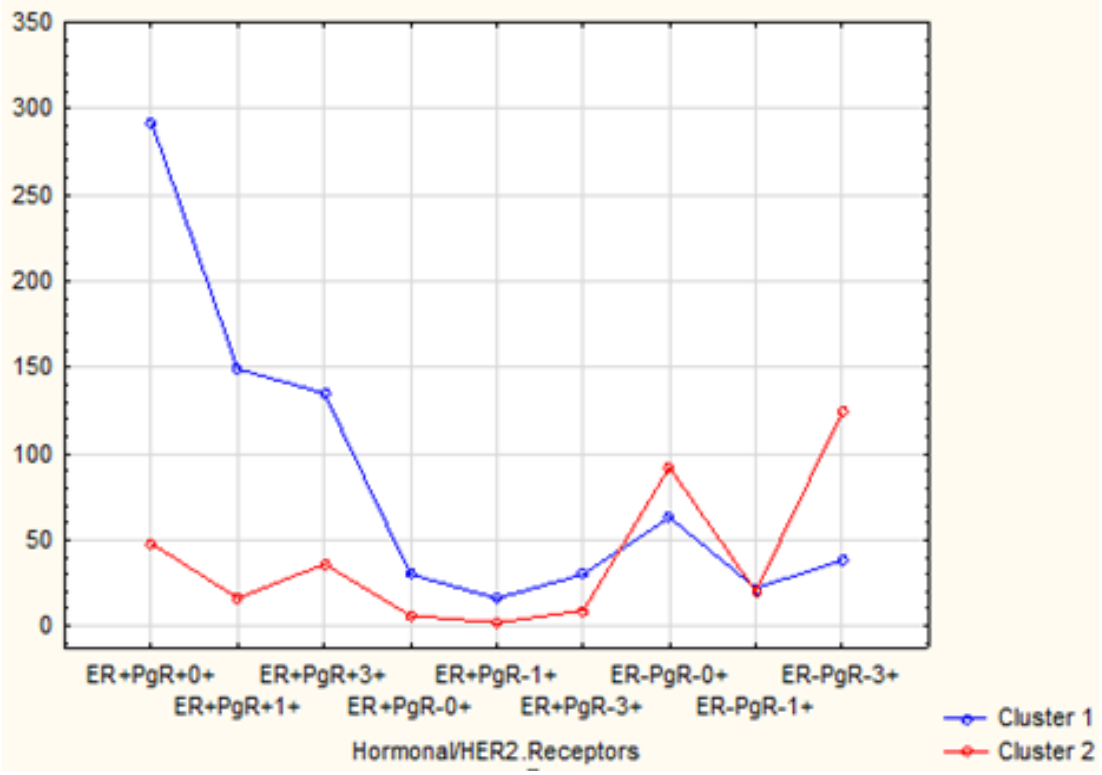

Figure 31: Cluster1 and Cluster2 counts according to ER/PgR/HER2 tumors phenotypes

(X-axis: Molecular subgroups; Y-axis: Observations counts)

Converting the same results to the molecular classification instead of tumor phenotype (figure 32), we see that the highest frequencies of C1 are found in Luminal tumors (A, B HER2 +/-) ; on the other hand, hormone-independent tumors (HER2 and TN) mainly belong to C2.

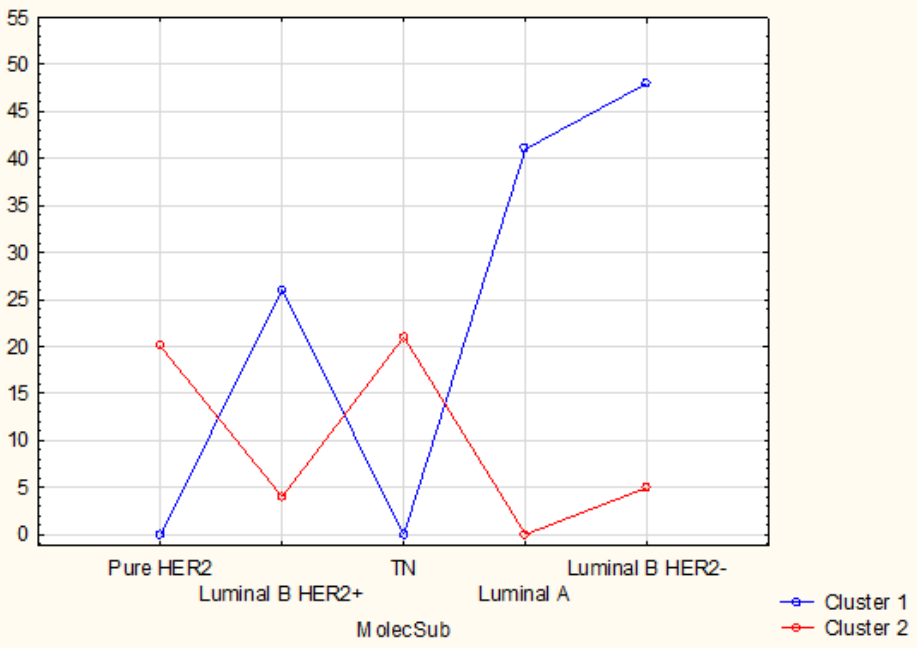

Figure 32: Graph of clusters counts depending on molecular subgroups membership

(X-axis: Molecular subgroups; Y-axis: Observations counts)

To further explain what this means in terms of numbers, and to better understand the distribution of KI-67 within these new subdivisions (C1 and C2), the results are grouped in the following table 13.

B. KI-67 distribution depending on cluster's membership:

Table 13: KI-67 distributions in Cluster1 and Cluster2 according to ER/PgR/HER2 status

(For convenience and clarity of the clutter-free table, the LuminalA HER2- subgroup has been combined in the same cell as the LuminalB HER2- subgroup because they both form the luminal type with a common criteria of underexpression of HER2.)

	HER2 status					
ER-PgR-	0+		**1+**		**3+**	
EM clusters	**Cluster1**	**Cluster2**	**Cluster1**	**Cluster2**	**Cluster1**	**Cluster2**
Patients (n)	63	92	22	20	38	124
Patients (%)	40.6	59.3	52.3	47.6	23.4	76.5
Mix-Max	0-40	45-100	1-50	60-90	0-70	75-100
Mean Ki-67	16.4	73	20.4	71.7	42.3	92.5
St dev	13	15.4	16.3	8.2	21.4	9.2
Total	**155**		**42**		**162**	
Molecular subgroup	Triple Negative				Pure HER2	
ER+PgR-	**0+**		**1+**		**3+**	
EM clusters	**Cluster1**	**Cluster2**	**Cluster1**	**Cluster2**	**Cluster1**	**Cluster2**
Patients (n)	30	6	16	2	30	9
Patients (%)	83.3%	16.6%	88.8%	11.1%	77	23
Min-Max	0-40	50-90	0-40	60-80	0-65	100-100
Mean Ki-67	16.86	68.33	14.7	70	26.3	100
St dev	13.32	15.7	13.7	14	20.2	0
Total	**36**		**18**		**39**	
Molecular subgroup	Luminal A/ B HER2-				Luminal B HER2+	
ER+PgR+	**0+**		**1+**		**3+**	
EM clusters	**Cluster1**	**Cluster2**	**Cluster1**	**Cluster2**	**Cluster1**	**Cluster2**
Patients (n)	292	48	149	16	134	36
Patients (%)	85.8	14.1	90.3	9.7	79	21
Min-Max	0-20	25-90	0-30	40-100	0-40	45-100
Mean Ki-67	9.5	45	14.8	51.7	17.5	71.1
St dev	6.3	18	9.1	15.1	12.5	16.4
Total	**340**		**165**		**170**	
Molecular subgroup	Luminal A/ B HER2-				Luminal B HER2+	

- In ER+PgR+ tumors, the most frequent cluster have a low Ki-67 mean value.
- In ER+PgR- tumors, the most frequent cluster have a low Ki-67 mean value.
- In ER- PgR- tumors, the most frequent cluster have a high Ki-67 mean value.

- Thus, in all phenotypes, KI-67 decreases with the overexpression of hormone receptors and simultaneous absence of HER2.

Overall, these results suggest that BC tumors can be furthermore divided into two levels of mitotic activity with different mechanisms behind ER+/-, HER2+/- tumors. These observations grouped by their respective molecular subgroup are summarized in table 14 :

Table 14: Summary of clusters frequency and molecular subgroups membership

	C1		C2		Total	
Molecular subgroups	**N**	**%**	**N**	**%**	**N**	**%**
Pure HER2	38	23.5	124	76.5	162	14.3%
Luminal A/B HER2-	487	87.11	72	12.8	559	49.6%
Luminal B HER2+	164	78.5	45	21.5	209	18.5%
TN	85	43.14	112	56.8	197	17.4%
Total	774	69	353	31	1127	100%

The percentage of patients belonging to C2 is higher when it comes to pure HER2 (76.5%) and TN (57%) molecular subgroups and may explain the high mitotic activity of these specific molecular subgroups, therefore the worst prognosis that they have compared to the luminal groups.

On the other hand, C1 is the most frequent (87% and 78% in Lum A/B HER2- and LumB HER2+ respectively) and it may explain their low mitotic activity and therefore better prognosis.
Still, there are some cases, albeit less frequent, of tumors in the TN and pure HER2 molecular subgroups that still belong to C1 (43% and 23.5%, respectively). Conversely, there are tumors belonging to C2 but found within the LumA / LumB HER2- and Lum B HER2 + molecular subgroups (12,8% and 21%, respectively).
A possible interpretation of the rare tumors classified as Luminal but belonging to C2 may be due to gene mutations that influence major oncogenic pathways and lead to a more severe phenotype. Or by the influence of an unknown pro-mitotic mechanism on the pathogenesis of these particular molecular subgroups that are supposed to be of a favorable prognosis.

As for the tumors classified in TN and HER2 but belonging to C1, it may be due to certain heterocellular gene signatures that can reveal this heterogeneity in terms of prognosis.

C. Predicting cluster's membership

For this prediction step, modeling was performed on the total dataset with 1127 records. After splitting the survival dataset into a training subset (70% = 789 records) and test subset (the remaining 30% = 338 records). The quality of data was then compared using prediction algorithms which are a set of ML techniques that search for patterns in sets and use those patterns to predict new records. The eight different prediction algorithms used are: NB; GLM; FLM; DL; DT; RF; GBT; SVM. The use of the multiple above-mentioned prediction algorithms at this stage, is mainly due to counterbalance the limitations of each and enjoy the benefits of others. And also to evaluate them together in order to choose the most suitable one for our dataset and types of data, both explanatory and that to be predicted. Obviously, all the chosen models have been meticulously concocted to our analysis which is above all a classification problematic. To achieve this, for algorithms, the variable of interest (variable to predict) is either the patient belongs to C1 or C2, and the explanatory variables are ER ; PgR ; Ki-67 ; HER2.

D. Evaluating the clusters membership predictions:

To explore which model shows the strongest predictive ability of cluster's membership (C1 or C2) in BC patients, we evaluated them by nine metrics to be able to extract the most robust model: Accuracy, Area Under the Curve Receiver Operating Characteristic (ROC-AUC); Precision; Recall; F-measure; Sensitivity ; Specificity ; Classification error; and Total running time of the model. The variable of interest is "cluster's membership" which refers to whether the tumor belongs to C1 or C2; and the explanatory variables are Ki-67, ER, PgR and HER2 features.

Table 15 summarizes the feature engineering run results. Each column represents a different evaluation metric and each row represents a different model.

Table 15: evaluation summary of all 8 prediction models for clusters membership

Prediction models	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F- measure (%)	Sensitivity (%)	Specificity (%)	Classification Error (%)	Scoring time (ms)
NB	80 ± 2.4	80	81.4	91.9	86.2	91.9	54.1	20	190
GLM	80.8 ± 2.2	83.1	79.2	97.7	87.4	97.7	44.9	19.2	246
FLM	80.5 ± 3.4	78.4	85.3	86.4	85.8	86.4	68.1	19.5	159
DL	79 ± 1.7	78.6	78.4	96.0	86.2	96	41.8	21	429
DT	81.4 ± 3	81.6	80.9	95.5	87.6	95.5	51.1	18.6	166
RF	80 ± 4.3	80.3	82.9	89.1	85.9	89.1	60.4	20	657
GBT	81.4 ± 2.5	81.1	82.7	92.4	87.2	92.4	57.5	18.6	1000
SVM	68.4 ± 0.7	52.5	68.4	100.0	81.2	100	0	31.6	2000

All the models show approximately good prediction scores. DT and GBT present very similar metrics except that the latter requires much more running time; therefore DT remains the most suitable algorithm for our dataset, where it has the best metrics scores and the least classification error, therefore the best prediction accuracy.

E. Univariate analysis between Clusters membership and histoprognostic features in the survival subset:

The cross-tabulation below summarized in table16 shows whether belonging to a certain category of a histo prognostic feature makes a case more likely to be in a particular category of the dependent variable (belonging to C1 or C2). This allows us to examine the association between these two categorical variables and the same for all the others. Patterns of association can be examined simply by comparing the observed frequencies across rows of the table, and comparing it to the calculated expected frequencies using the G–test of goodness-of-fit (also known as the likelihood ratio test, the log-likelihood ratio test, or the G2 test of independence). The latter is the most appropriate test for this univariate analysis because the variables are nominal. It will help us see whether the number of observations in each category fits a theoretical expectation. The null hypothesis is that the relative proportions of one variable are independent from the second variable.

Table 16: Bivariate analysis between histopathological features and cluster's membership

Attributes	C1	C2	Total	G-square	P-value
Age Category					
<40	6	14	20	15.4652	0.000084
>40	109	36	145		
TNM staging					
T1	45	14	59	4.7571	0.092684
T2	45	17	62		
T3	25	19	44		
ER status					
Negative	1	45	46	151.3062	0.00001
Positive	114	5	119		
PgR status					
Negative	9	44	53	107.3396	0.0001
Positive	106	6	112		
HER2 status					
Negative	26	24	50	10.2545	0.001363
Positive	89	26	115		
Ki-67 status					
Negative	64	48	112	32.4218	0.00001
Positive	51	2	53		
Lymph Nodes Infiltration (NI)					
Negative	91	48	139	9.1514	0.002485
Positive	24	2	26		
Vascular Emboli (VE)					
Negative	81	47	128	13.2965	0.000266
Positive	34	3	37		
Molecular subgroups					
Pure HER2	0	20	20	145.7436	0.00001
Luminal B HER2+	26	4	30		
Triple Negatives	0	21	21		
Luminal A	41	0	41		
Luminal B HER2-	48	5	53		

Some histo prognostic features originally in the form of numeric variables were converted into nominal/categorical variables to fit the univariate analysis under the same statistical test. The “Age” variable has been split into two scored sub-categories:

- <40: referring to patients under 40 years old.
- >40: referring to patients older than 40 years.

This stratification has been maintained under the findings of several studies that have confirmed 40 y.o as an adequate threshold that delineates an aggressiveness, survival rate, cancer biology and metastasis incidence that differ in women under the age of 40y.o with BC faced by older women.
As for ER and PgR receptors, their status is considered negative if the expression is less than 1%, otherwise the status is considered positive. On the other hand, Ki-67 status is considered positive if its expression is greater than 20%. In terms of HER2 status, all 3+ scores are taken for positive, the remaining observations with 0+/1+ are considered negative.
The vascular emboli are scored positive if they are present, in case of absence their status is considered negative. The same goes for the lymph node infiltration status.
The results show that the histo prognostic features and cluster membership are effectively associated. Therefore, the next main objective is to assess Kaplan meier survival curves depending on cluster’s membership and then predict the membership to those clusters for any BC patient for which all the histo prognostic features are available; that is why predictive models are of great interest.

F. Kaplan-Meier survival curves grouped by clusters’ membership

The main interest of using Kaplan Meier analysis is to study whether the probability of survival depends only on the time after the initial event (BC screening) or depends also on clusters belonging. The curves are initially considered to be stable with respect to absolute time. This means that observations entering the study on different dates should have the same behavior. To do this, we use the equality test (Log Rank) on the survival distributions on the different levels of our strata factor (C1 and C2). This helps obtain stratified survival curves, as well as essential statistics such as the median residual time of survival which estimates the survival functions, without requiring that

the time intervals be regular, as well as to analyze the evolution of the records belonging to either C1 or C2 depending on survival time.

Therefore, this chapter evokes survival analysis as a set of several statistical methods aiming to process data where the feature to be explained is time until an event occurs (variable of interest). The latter is assessed based on two related mathematical functions.

First, the survivor function which is referring to the probability that an individual survives from the time of origin to some time beyond time T. This is why we used the Kaplan-Meier method. The log rank test was also used to evaluate differences within groups in terms of the survival curves.

Second, on the other hand, the hazard function gives the instantaneous potential of the event occurring at a specific time Tx.

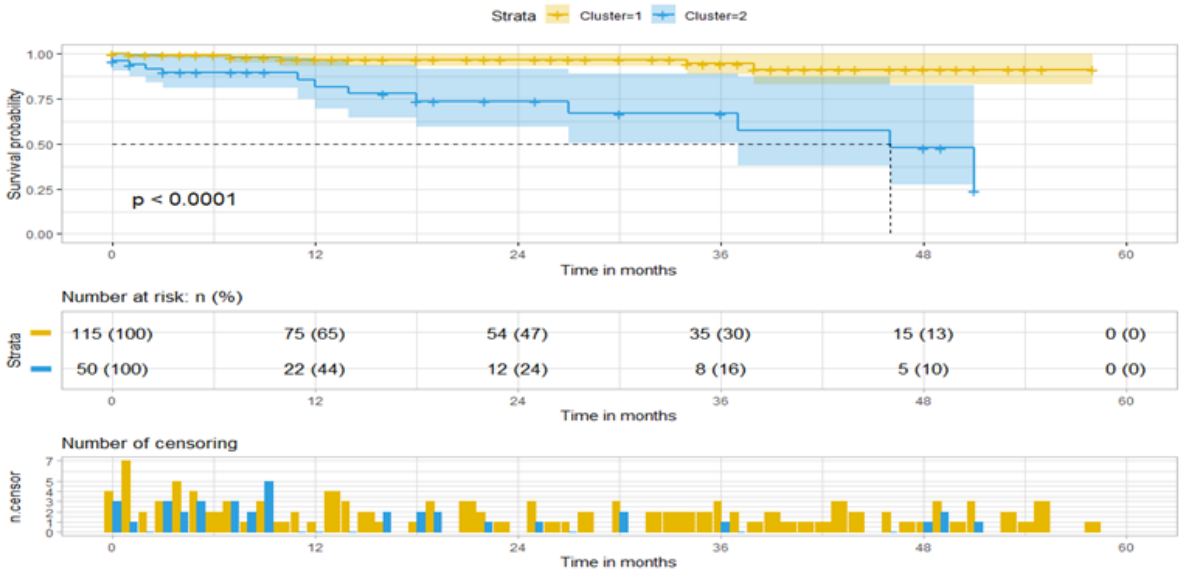

Figure 33: Summary of Kaplan-Meier survival curves grouped by clusters from the survival subset

For Kaplan Meier survival analysis, the data processing was performed on the survival dataset with 165 records and the associated histo-prognostic features were considered predictors of the survival rate. It was mainly carried out by grouping the patients by their cluster's membership, without taking in consideration their molecular subgroups. The mean survival of C1 was found to be approximately 52 months while it was 37 months for C2 suggesting a poorer survival for the latter compared to C1.

The log rank test showed high significance of the separation between the two survival curves (p=0.0001), while the separation of both curves by Ki-67 alone showed a significance with p=0.0063.

G. Kaplan-Meier survival curves grouped by molecular subgroups and clusters belonging:

After asserting that the clusters belonging influences the survival time of patients. We are now interested in studying the synergistic effect of cluster belonging and molecular subgroups membership at the same time on the patient's overall survival.

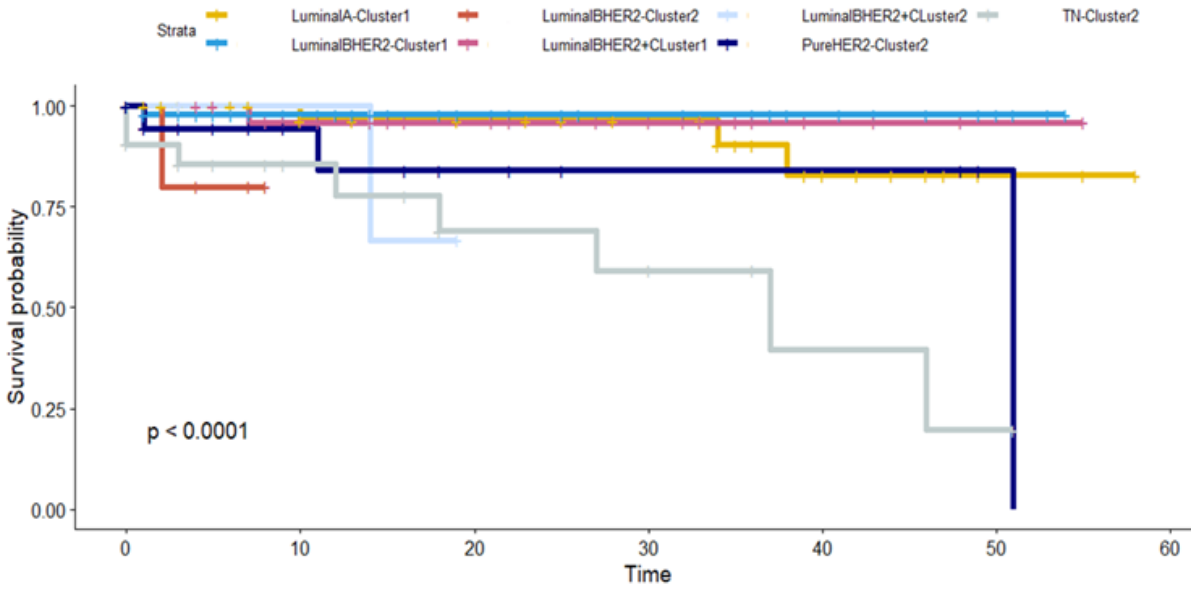

Figure 34: Summary of Kaplan-Meier survival curves grouped by Clustered molecular subgroups

(X-axis: Survival probability, Y-axis: Time in months)

Within this particular survival dataset, it has to be highlighted that there was no tumor classified in the Pure HER2 and TN molecular subgroups and simultaneously belonging to C1, conversely, we did not find any Luminal A type tumor belonging to C2.

The plotted summary of the 5-year OS rates presented in figure 37 shows the survival of the same patients presented in figure 36 but based on their belonging to the clusters within each molecular subgroup. The curves diverged early especially for the C2 grouped ones: patients belonging to C2 of TN have the worst survival outcome which reaches less than 25% survival probability after 4 years of follow-Up, followed by C2 of Luminal B HER2-; Luminal B HER2+, and then C2 of Pure

HER2 molecular subgroup. Therefore, we can conclude that all defavorable prognosis found are attributed to tumors classified as C2 whatever their molecular subgroup membership is.

Conversely, the survival curves representing the luminal molecular subgroups belonging to Cluster1 (LuminalA-Cluster1, LuminalBHER2-Cluster1, LuminalBHER2+Cluster1) all converge at the beginning of the 5year Follow-Up, since they are superimposed until 3 years of follow-Up. The luminal A curve subsequently reached 87% probability of survival compared to the two Luminal B curves (LumBHER2+Cluster1) and (LumBHER2-Cluster1), which reached approximately 95% probability of survival at the end of the study. The log rank test for difference in survival indicates that the survival curves differ significantly.

Generally, the partition of survival curves depending on molecular subgroups AND clusters belonging is more precise and distinctive of the survival outcome of BC patients and statistically significant. Especially for the luminal B molecular subgroups, the patients belonging to C2 of Luminal BHER2- have a significantly poorer survival outcome compared to patients of the same molecular subgroup but belonging to C1. The same thing was remarkable for Luminal B HER2+. Therefore, this new classification refines and distinguishes the prognosis of patients belonging to the same BC molecular subgroup.

H. Predicting Breast Cancer patient's survival depending on cluster's membership

After validating that within the same molecular subgroup exist clusters with different prognosis and survival, we are now interested in predicting such cluster membership.

For the prediction to be precise, the model to set up must be seriously substantiated. The objective is to prove that the model generates good estimates of the variable of interest values (survival based on clusters belonging). To achieve this, it is necessary to work with at least one training dataset and one validation dataset. Quite simply, the training data is used to calibrate the model while the validation set is used to show that the model is reliable and relevant. More generally, we separate a basic initial set of samples into a training dataset and a validation dataset.

Therefore, the prediction modeling was performed on the survival dataset after splitting it into a training subset (70% of the survival dataset = 115 records) and a test subset with the remaining 30% (= 50 records).

The quality of data was then compared using 4 algorithms: DT; RF; NN; LR.

To assess the predictive performance of these models, we need performance metrics that are easily usable for clinical analysis, in the sense that they must account for the capacity of the models to distinguish high-risk patients from low-risk ones.

In the field of classification, metrics are defined from a contingency table, which counts the results according to the actual and the predicted values.

We therefore redefine five performance metrics: area under the ROC curve (AUC); accuracy, specificity, sensitivity and Cohen's Kappa coefficient.

A model that obtains a good score on one of these metrics will necessarily have equally good results on the others. A sophisticated survival prediction model should score high on all of these metrics.

These metrics are useful because they help describe a model's ability to answer different questions, as follows:

- **ROC-AUC**: Is a patient likely to die within a certain observation period depending on the cluster he belongs to?
- **Accuracy**: what is the correct prediction survival rate of the patients?
- **Sensitivity**: how good is the ability of the model to predict true positives ?
- **Specificity**: how good is the ability of the model to predict true negatives ?
- **Cohen's Kappa**: To what degree can one assess the degree of agreement (concordance) between several evaluators as to how classify a set of patients into one cluster or another?

Once the model has been created with the training dataset, it is necessary to calculate objective indicators to assess whether the model has generated relevant predictions for the studied variable. The "true" values of this variable are assumed to be known for all training and validation datasets. Intuitively, for each sample in the validation dataset, we want to know if the values predicted by the model are close to the true values of the validation dataset.

As shown in table 17, RF generated one of the best accuracies and performed better in terms of the other metrics also, therefore we used this algorithm for further analyses as we found it to be the most suitable for the survival data. On the other hand, we see that GLM provides the lowest performance especially in terms of the kappa statistic which turns out to be negative. The latter means that the number of agreements observed is fewer than would be expected by chance. In other words, there is less agreement than would be expected by chance according to GLM.

Table 17: evaluation summary of all the prediction models of survival

Algorithm	Accuracy (%)	Kappa (%)	ROC (%)	Sensitivity (%)	Specificity (%)
Decision Tree	94	92.26396	96.21212	99.09091	93.33333
Random Forest	97.7	90.2173	100	97.5	100
Neural Network	90.8	37.092391	96.42857	99.16667	35.71429
GLM	71.58	-5.556183	46.65179	65.35714	25

I. Random survival forest:

When it comes to modeling non-linear dependencies between variables and improving learning, Random Survival Forests (RSF) can be used. This technique overcomes several limitations by facilitating the discovery of complex data structures. The RSF method defines two particular principles in the context of survival analysis. The first is the use of the logrank test to construct decision trees. The second is the construction of mortality sets to predict survival probability. In general, this algorithm can be summarized by the following steps:

1. Draw B bootstrap samples.
2. Develop a survival tree based on the data from each of the bootstrap samples.
    a) At each node of the tree, define a subset of predictor variables.
    b) Among all the binary splits defined by the predictor variables selected in (a), find the best split into two subsets (the daughter nodes) according to an appropriate criterion for right-censored data, such as the log-rank test.
    c) Repeat (a)-(b) iteratively on each daughter node until a stopping criterion is satisfied.
3. Aggregate the information from the terminal nodes (nodes with no further splits) of the B survival trees to obtain a risk prediction set.

In survival analysis, the log rank test makes it possible to estimate the survival functions of two groups at each time of interest (i.e. at each survival or censoring time). This is a non-parametric test for censored data in the context of non-informative censoring. It is used to test the null hypothesis that there is no difference in the probability of survival between the two groups at each time of interest. The statistical calculation is based on the time of each event. For each of these times, we take the total number of events observed (for each group) and the number of individuals at risk. We can then estimate the number of events expected for each group. Much like Cox regression, RSF falls into the context where censorship is non-informative. Indeed, the log rank test statistic depends only on the number of events at each instant.

In the following dataset, censored data are not informative (i.e. do not carry information concerning the evolution of the probability of survival). This emerges from the fact that the censorship mechanism is independent of the observed event. Indeed, in the case of a clinical study, a patient is censored either if he abandons the study in progress, or if he reaches the end of the study without having undergone or experienced the event of interest.

For example, if the follow-up of a patient ceases during the study because the patient moves to another country, the reason for his leaving is independent of his death risk. This means that at each time, the censored patients have the same prospect of survival as those who continue to be followed. However, the approaches presented above (survival random forests) only consider the times (or the order of times) at which an event is observed. Data censored only intervenes in the counting of patients who are still at risk at the dates when an event occurs, which makes it a very adequate method to predict survival probability.

The RF model whose outcome is visualized in figure 38, built seven different predicted survival curves. BC patients of C2 within the TN molecular subgroup have the worst predicted outcome, which reaches 25% of survival probability outcome, followed by C2 of LuminalBHER2- and C2 of LuminalBHER2+. These last two curves turned out to be superimposed and reached 40% of survival probability.

Conversely, C1 of Luminal B HER2- has the best predicted prognosis, which approaches 98% of survival probability, followed by C1 of LuminalB HER2+ and Luminal A; the latter have a better predicted survival probability converging towards the luminal groups. This was expected because luminal BC is known to have better prognosis and survival rate. But it has to be highlighted that even within those particular luminal subgroups, some patients may present poor prognosis.

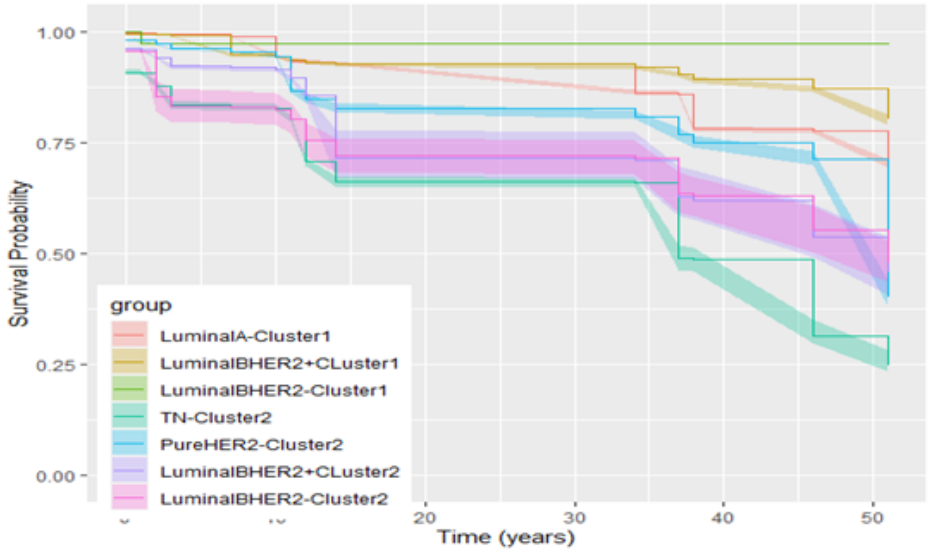

Figure 35: Kaplan Meier survival estimates predicted by Random Forest algorithm according to clusters belonging and molecular subgroups membership

Predicting survival according to clusters belonging, reveals a certain accurate comparison of the actual survival curves. This testifies the great capability of this model, based on just four histopathological indicators (ER; PgR; HER2 and Ki-67) to predict a patient's survival.

This fairly stable prediction ability allows us to suggest that this new 2-clusters subdivision should be taken into consideration to refine patient prognosis. It also means that the histo prognostic features considered in Pathology departments as standard indicators of BC and used in this analysis as explanatory variables, are sufficient enough to characterize each cluster.

This is the main outcome of this study aimed at finding new methodic insights that may help pathologists to refine the molecular classification of BC established in routine, using the same panoply of histoprognostic features.

J. Cumulative force of mortality of cluster 1 and cluster 2 depending on molecular subtypes membership

In this section, we computed the same survival curves using simultaneously the combination of two factors “Clusters” and “molecular subgroups” belonging.

The output visualization was assessed using the “survminer” package on R. The plot below (Figure36) shows survival curves by either the patient’s belonging to C1 or C2, faceted according to the molecular subgroup he belongs to.

The predicted survival curves depending on the molecular subgroups variable faceted according to the cumulative force of mortality, can be defined as the cumulative probability that the event occurred before 5-years of the follow-Up period. This incidence proportion of death is the most pronounced in C2 and more especially in the hormonal non-dependent tumors, mainly TN and pure HER2 tumors and still high compared to the other tumors belonging to C1. Plus, there is a remarkable difference in the force of mortality between Luminal B HER2+ where it is higher than the Luminal B HER2- one. It can be concluded that the simultaneous presence of hormonal receptors; presence of HER2 receptor and C2 membership increases the force of mortality.

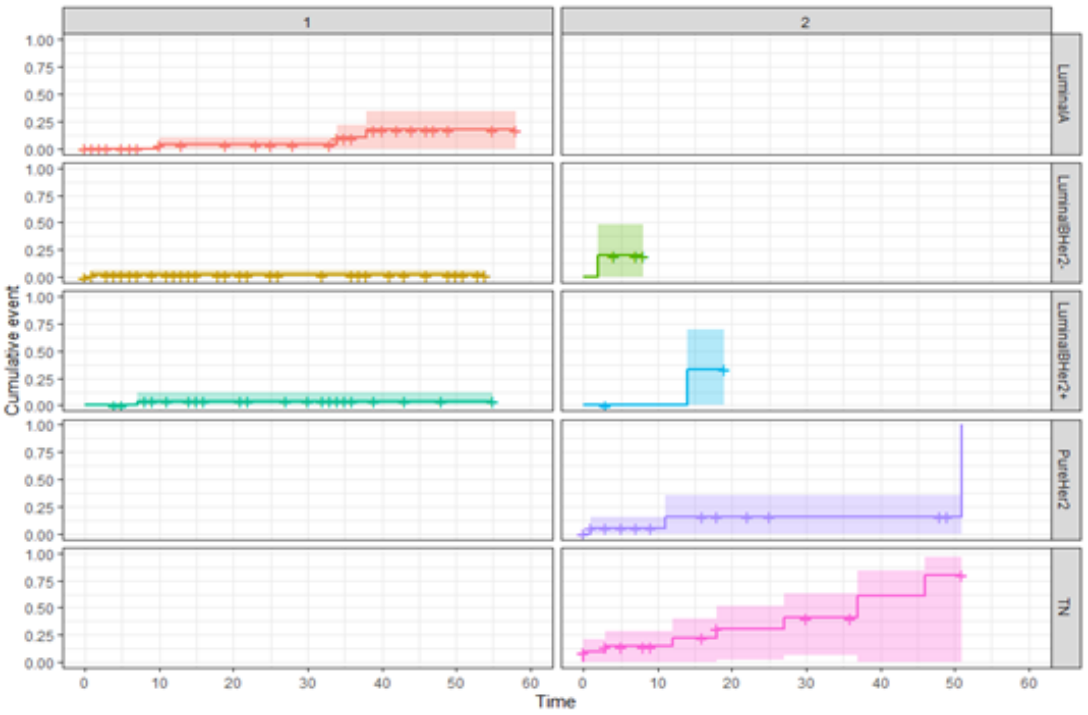

Figure 36: Cumulative force of mortality of Cluster1 and Cluster2 depending on molecular subgroups membership for 5 years of follow-Up

(Left column 1: Cluster1 membership; right column 2: cluster 2 membership)

K. Important variables selection

Tree based predictive models such as RF are very popular for analyzing large sets of data, in particular because of their good predictive precision. However, they are not intrinsically interpretable since their prediction results from the average of several hundreds of decision trees. A classic approach is to calculate the importance of explanatory variables, which are used to assess the predictive impact of each input variable and thus play a large role in data analysis.

After determining the 2-clusters partition classification, and proving its effectiveness in terms of the patient's survival prediction; subsequently the refinement of their prognosis, we would like to assess its importance compared to the other histopronostic parameters recorded in routine by pathologists. The variable importance selection algorithm was therefore addressed, which helps us determine the most important explanatory variables that govern the patient's survival.

Thereafter, we create a new variable called "ClustersANDMolecSubgroup" that refers to the simultaneous combination of "clusters belonging" and "molecular subgroup membership". It tells us about the cluster AND the molecular subgroup to which the patient belongs at the same time.

a) VIMP algorithm

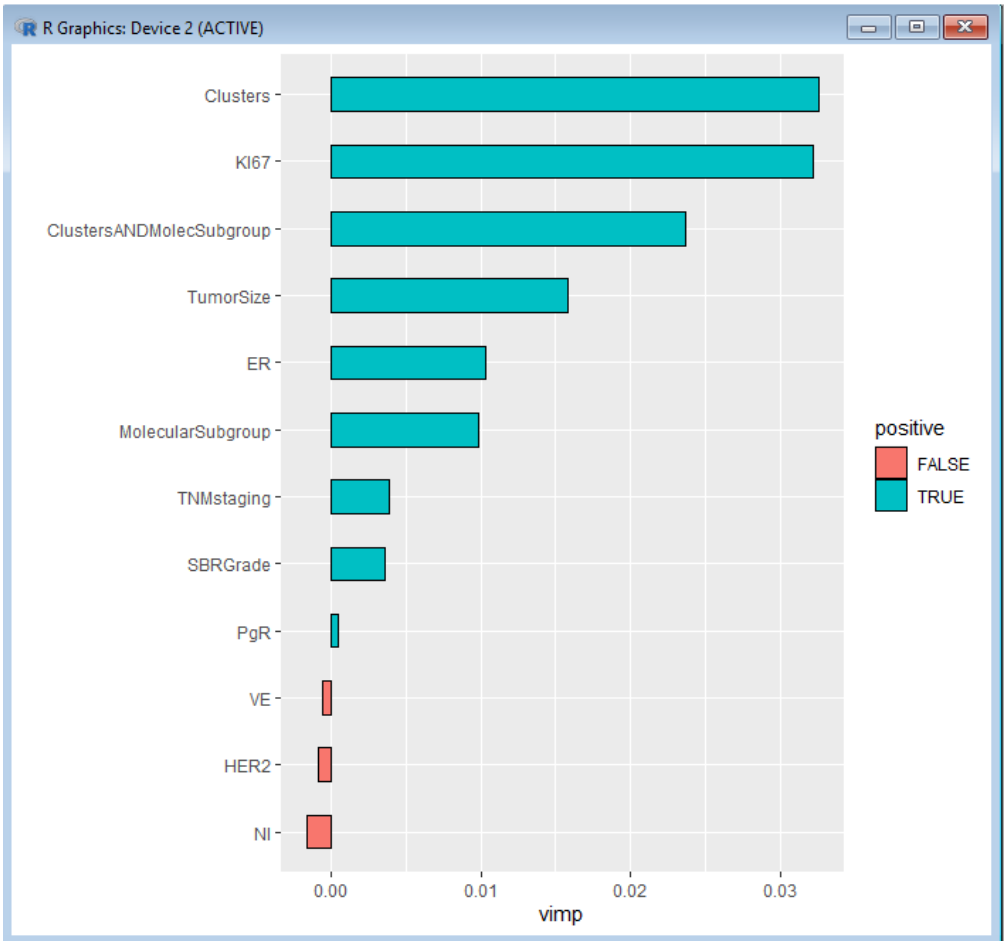

Figure 37:Important variables selection by Random forest VIMP algorithm

(Blue bars indicate positive VIMP, red indicates negative VIMP. Importance is relative to the positive length of bars.)

The function's output that we visualized in figure 37, shows the variables, in VIMP rank order, labeled with the corresponding named vector.

VIMP measures are shown using bars to compare the scale of the error increase and colored by the sign of the measure (red for negative values) which tends to generate False positive records of the RF predictive model, therefore decreasing its accuracy.

The VIMP plot details variable importance ranking, from the largest which is “Clusters” as the top first important feature in terms of predicting BC patients’ survival outcome; followed by the proliferation index Ki-67; ClustersANDMolecSubgroup; Tumor Size; ER; Molecular subgroups membership; TNM staging; SBR grade and PgR. Those are the only important histo prognostic features (as highlighted in green) which refers to their importance and their influential potential in survival prediction, according to the VIMP algorithm.

We can deduce therefore that the largest VIMP value goes to the cluster the patient belongs to. The latter was found to be more important than all the other histo prognostic variables.

This further proves that the 2-clusters partition we propose in this work is quite revealing in terms of predicting overall patients survival, more than molecular subgroups as a prognostic reference.

Additionally, the second partition that we propose based on the simultaneous effect of clusters belonging and molecular subgroups membership “ClustersANDMolecSubgroup” is ranked third, which also confirms its veracity.

On the other hand, nodes infiltration, HER2 status and vascular emboli presence turns out to be the weakest VIMP ranks since they are the closest to zero, and therefore contributes nothing to the predictive accuracy of the forest, indicating that the predictive accuracy improves if these three features are misspecified. Which makes them less informative than noise.

b) Minimal Depth variable selection:

In order to evaluate the relevance of the previous results, we used another algorithm called minimal depth. The latter assesses a simple optimistic threshold value, classifying variables with minimal depth lower than this threshold as important in forest prediction as visualized in figure 38 below.

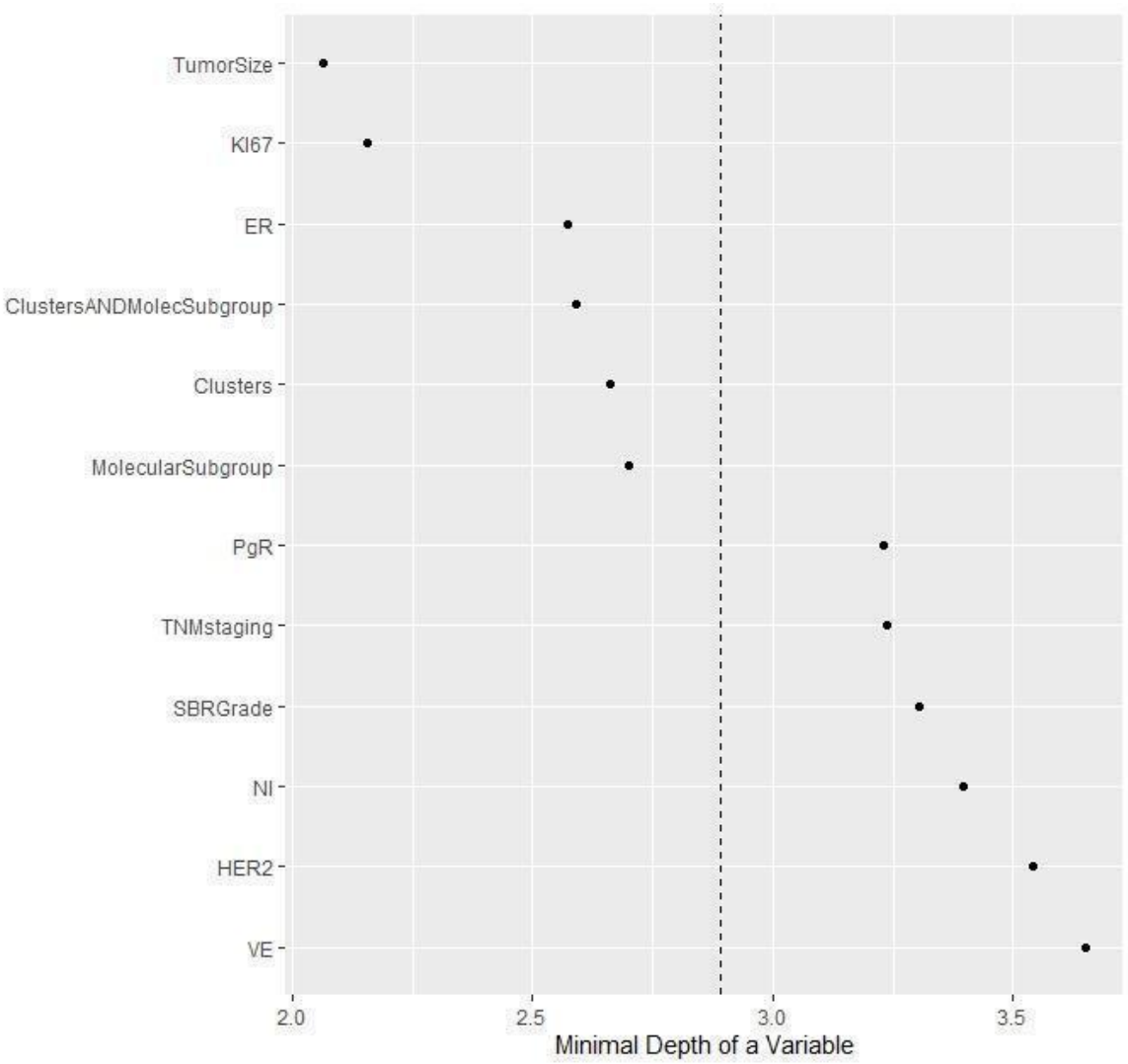

Figure 38: Important variables selection by Random forest's minimal depth algorithm

(Minimal Depth variables in rank order,
most important at the top. Vertical dashed line indicates the maximal minimal depth for important variables.)

The minimal depth plot based on RF indicates that variables are ranked from most important at the top (minimal depth measure), to least at the bottom (maximal minimal depth). The latter votes for a total of six important variables which are the tumor's size; Ki-67 proliferation index; ER; the partition according to both clusters and molecular subgroups simultaneously; followed by the 2-clusters partition, and at last the molecular classification.

On the other hand, PgR; TNM staging; SBR grade, nodes infiltration; HER2 status and Vascular Emboli contribute nothing to the prediction value of the model.

c) Important Variables selection comparison (VIMP vs Minimal Depth)

Since the VIMP and Minimal Depth measures use different methods, we expect the variable ranking to be somewhat different between both their outputs. Therefore, we used the "gg_minimal_vimp" function to compare variable rankings between both methods in figure 39.

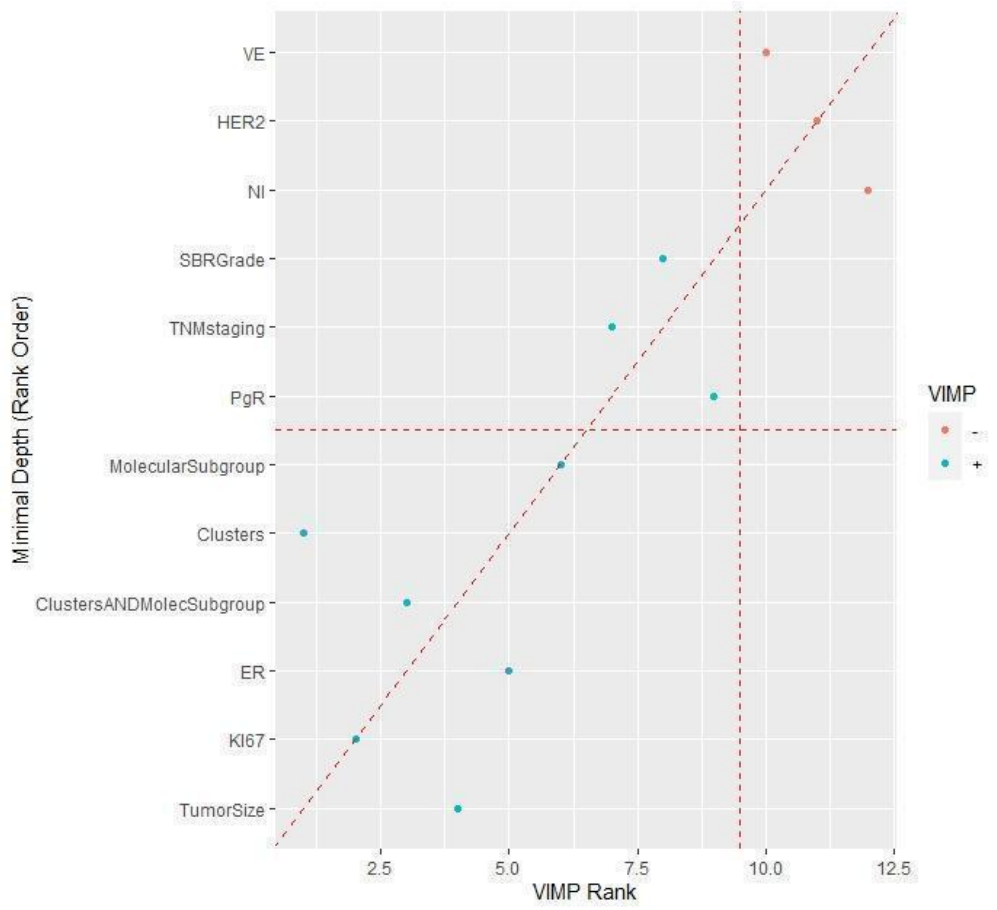

Figure 39: VIMP vs. Minimal depth rankings comparison

The comparative figure shows points along the red dashed line that indicate where the measures are in agreement. Points above the red dashed line are ranked higher by VIMP than by minimal depth, indicating the variables are more sensitive to misspecification. Those below the line have a higher minimal depth ranking, indicating they are better at dividing large portions of the population. The further the points are from the line, the more the discrepancy between measures.

Therefore, we can conclude from figure 39 that both algorithms have different outcomes in terms of the important variables ranking but still agree on the fact that both the partitions we created (either based on clusters/ or on clusters AND molecular subgroups) are among the five most influential variables on BC patients survival.

In other terms, they both confirm our partitions are more important than other histo prognostic features, and that they contribute more to predict the patient's survival than HER2 status, Nodes invasion and vascular emboli presence, the latters were confirmed by both algorithms. The whole panoply of 14 explanatory variables initially considered is therefore reduced to a manageable subset containing the six most influential features on patient's survival.

d) Variable / Response dependence

So far, the two different methods mentioned above have been used to reduce the panoply of prognostic features to a manageable subset. Once we have an idea of which variables contribute the most to the predictive accuracy of the random forest, we would like to know how the response variable depends on the clusters and molecular subgroups partition we created.

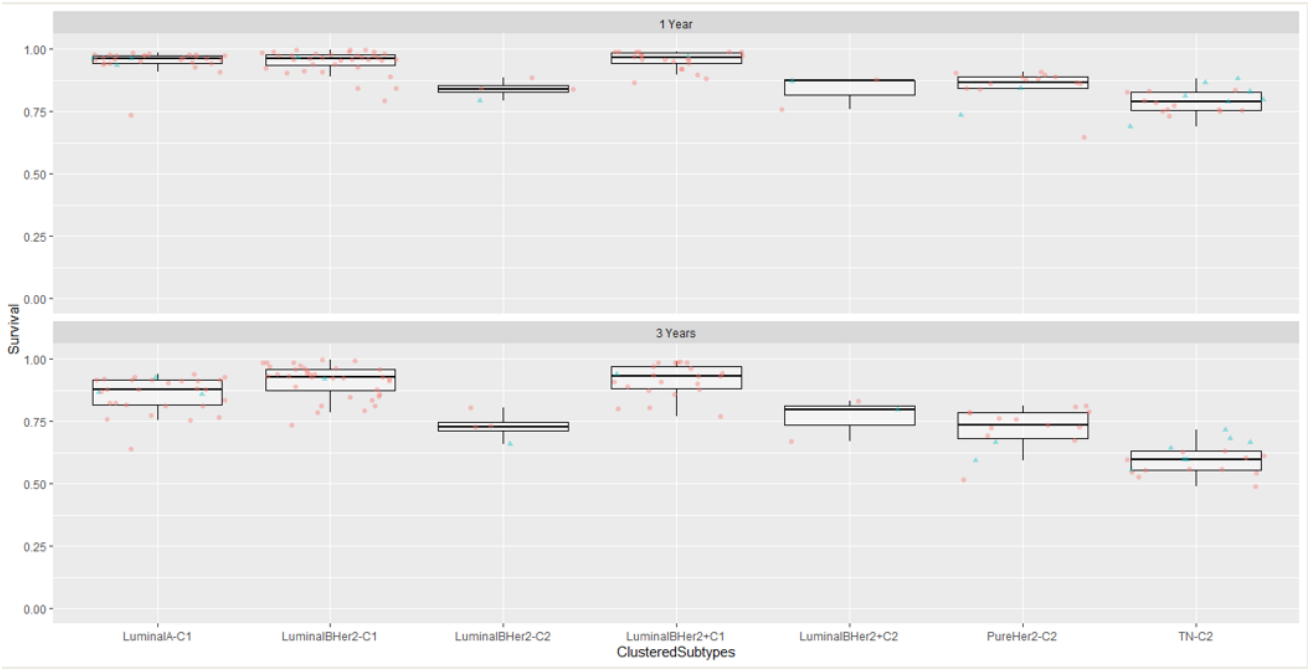

Figure 40: Survival dependence plots at 1 and 3 years for clusters membership.
Individual cases are marked with green circles (dead) and red (alive or censored).
Boxes indicate distributional properties of observations in each group.

The variable dependence plot represented in figure 40 visualizes the survival rate dependency on the "ClustersANDMolecSubgroup" feature, and examines the forest predicted survival dependency on the clustered subtypes previously found, at the 1- and 3-year survival time. The boxes are shown with horizontal bars indicating the median, 75th (top) and 25th (bottom) percentiles.

The predicted survival depending on the clustered subtype affiliation shows that the worst survival boxplot is relative to patients belonging to C2 of TN molecular subgroup, whose survival outcome diminishes from 82% survival probability at 1year to 60.5% survival probability at 3years of follow Up. Followed by LumBHER2-Cluster2 showing a survival probability of 84.5% at 1 year and 72.5% at 3 years of follow-Up. Patients affiliated to LumB HER2+Cluster2 and PureHER2-Cluster2 have also a poor survival outcome that diminishes from 87.5% to 81.5% and 75% respectively after 3years.

Conversely, the survival of patients assigned to LumA-Cluster1 decreases from 95% to 87%; from 96% to 93.5% for LumBher2-Cluster1 and from 97.5% to 93.5% for LumBHER2 + Cluster1 after 3 years of followUp.

It is to be highlighted that the survival probability within LumBHER2- clusters is very distinct, therefore they have different outcomes and shouldn’t be considered as one group. After 3 years, patients belonging to C1 of LumB HER2- decreased only by 2.5% while it decreased by 12% for patients affiliated to Cluster2 of the same molecular subgroup.

Section 1 main findings:

- -The distinction of a new intrinsic subdivision to the molecular classification, rated C1 and C2.
- -C1 moderately includes all BC patients with low Ki-67 and a less aggressive and proliferative tumor profile than those belonging to C2.
- -The distinction between two BC patients belonging to the same molecular subgroup seems therefore to be essential because there is a certain significant degree of heterogeneity within the same molecular subgroup, which can be related to a different prognosis.
- -Among all BC histo-prognostic features, hormone receptors support belonging to C1, while Ki-67 along with HER2, lymph node invasion, the presence of vascular emboli and the SBR grade support belonging to C2, since they confer a proliferative and aggressive power to the tumors that converges with the profile found in tumors in C2.
- -The mean OS of records belonging to C1 is much more favorable compared to C2.
- -The OS of patients belonging to the C1 of LuminalB HER2- is much closer to that of patients belonging to the C1 of luminal B HER2+ than to that of C2 patients of luminal B HER2-. The same goes for the other clustered sub-groups, hence the need to split each sub-group into two subdivisions C1 and C2 which refine their prognosis and survival prediction.
- -We assessed several survival predictive models and evaluated them with different metrics, by applying the cross-validation option. The results show that the survival can be effectively

predicted by all the models which presented good accuracy scores, but mainly the Random Forest one which preceded them and successfully predicted the survival for each cluster within each molecular subgroup.

- -The application of two different important variable selection algorithms showed that the most important histopathological variables in terms of predicting BC patient survival are: Cluster & molecular subgroup, Ki-67; Tumor Size, the clusters belonging; hormonal receptors status.
- -As a result, it was possible to reduce the panoply of histopathological parameters having the greatest importance in dividing BC patients and in predicting their prognosis and OS after 5 years to the 4 most important and explanatory variables.

Section2: A comparative study with external TCGA-BRCA and METABRIC validation:

In this section, the main goal is to validate the same “2 clusters partition” within other independent BC databases.

To do this, it will be necessary to explore the optimal number of hidden clusters in new databases with different numbers of records and different histopathological; transcriptomic and/or genetic features, but in the most impartial and objective way. This is the reason why unsupervised ML techniques are used. The latter have the advantage of finding by themselves the best number of clusters so as not to prevail over the number of clusters (k = 2) obtained in the internal Moroccan database.

For this aim, TCGA dataset was used, which is freely accessible and retrieved from the CBioPortal For Cancer Genomics. It contains a great wealth of information: a combination of histopathological; RNA-seq and survival data for 625 BC records.

1) TCGA dataset:

A) optimal number of k-clusters:

First, one of the most important things is our work reproducibility on other datasets. While different distance metrics and grouping methods can lead to different interpretations of the data, at least we strive to achieve the same interpretations for the same method in the same dataset. For that, we want to choose the optimal number of clusters ``k” in the TCGA dataset and not necessarily just align to the previous Moroccan outcome (k=2).

Accordingly, the goal is to have clustering that explains the TCGA dataset the most, because with too many segments we lose our ability to interpret clusters, while on the other hand, with too few clusters we could risk generalizing the distribution of TCGA BC records.

For this aim, we applied two main types of cluster validation measures: “internal” and “stability” measures. Likewise, we used the most performant clustering algorithms.

Therefore, we thought of combining both validation measures and clustering methods, to simultaneously evaluate several clustering algorithms while varying the number of clusters, which helps determine the most appropriate method and number of clusters for this particular dataset. We

also made sure the data did not have any predefined classified input or clustered groups so as not to alter our analysis.

Thereafter, twenty-four combined quality index calculations will vote for the best optimal number of clusters at a time, therefore we can rely on this majority voting rule and decide on the optimal value of “k”, as determined in figure 41.

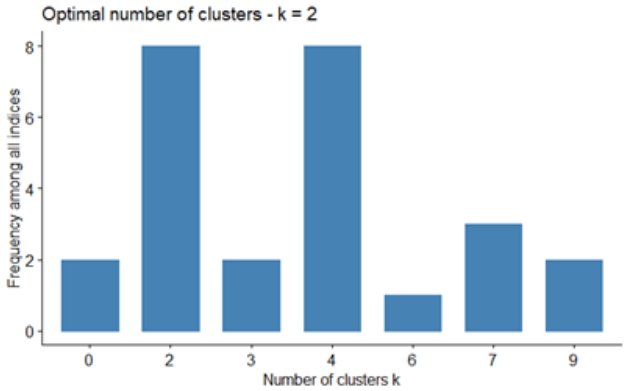

Figure 41: Determination of k-clusters optimal number by several quality indices

As we can see in figure 41, among all indices, 8 proposed k=2 as the best number of clusters and according to the majority rule, the best number of clusters is therefore k = 2. However, k = 4 also seems to be a potential candidate. As we can see from the two approaches, we can to a certain extent choose either k=2 or k=4 as the optimal number of clusters.

Since 8 indices voted for k = 2 and 8 others voted for k = 4 equally, we can decide on the cutoff which - in our opinion - explains the maximum data distribution. But to ensure the stability of the results and prevent running this clustering several times may produce different results, we will rely on a new method that aims to assess clustering stability between the two optimal numbers. To say it another way, if we get different clusters every time, our work isn't reproducible.

Therefore, using a new package called “clValid” allows us to further analyze two other types of measures: “internal” and “stability” measures.

b) Internal measures

As an internal validation, measures reflecting compactness, connectedness, and separation of clusters partitions were chosen. As already mentioned in the methods section, the Silhouette score is used to measure how dense and well-separated the clusters are by taking into consideration the intra- and inter- cluster distance with data points within the same cluster, and within the next nearest cluster respectively.

The range of the silhouette score falls between [-1, 1]; when it equals 1, this means that the clusters are nicely separated because they are very dense. On the other hand, the score of 0 refers to overlapping clusters. But when the score is negative, it means that the cluster's membership of each data point may be wrong/incorrect. Observations from the Silhouette index and the other internal measures show that the k-cluster value =2 is the best pick, which supports the proposals of the previous indices.

The plots of the connectivity, Dunn Index, and Silhouette Width are given in table 18 below. Recall that the connectivity should be minimized, while both the Dunn Index and the Silhouette Width should be maximized. Thus, it appears that hierarchical clustering outperforms the other clustering algorithms under each validation measure. For hierarchical clustering the optimal number of clusters is clearly two. For PAM, a case could be considered by using three clusters.

On the other hand, stability measures including APN, AD, ADM, and FOM should be minimized in each case. For the APN and ADM measures, PAM clustering with two clusters again gives the best score. However, for the other two measures, k means with five clusters has the best score. Though PAM clustering with two clusters has the best score, hierarchical clustering with four clusters is a close second. The AD and FOM measures tend to decrease as the number of clusters increases. Here, both PAM and k-means with five clusters have the best overall score. For the ADM measure PAM with two clusters again has the best score.

In table 18, a tabulated summary of all the previously mentioned measures is displayed, along with the clustering methods and number of clusters corresponding to the optimal score for each measure. Therefore, we can clearly see that Hierarchical clustering method with two clusters performs the best in each case.

Clustering method	Internal measures	2 clusters	3 clusters	4 clusters	5 clusters
Hierarchical	Connectivity	2.9290	10.4544	24.9115	28.7694
	Dunn	0.2603	0.0933	0.0915	0.0915
	Silhouette	0.463	0.2450	0.2619	0.2328
K-means	Connectivity	62.1687	152.5885	133.0683	149.8202
	Dunn	0.0349	0.0183	0.0315	0.0335
	Silhouette	0.4114	0.2430	0.3040	0.3036
PAM	Connectivity	63.6940	69.7341	161.0004	194.0429
	Dunn	0.0437	0.0418	0.0214	0.0328
	Silhouette	0.4163	0.4331	0.2987	0.2888
	Stability measures	APN	AD	ADM	FOM
	Score	0.0808	1.6819	0.3595	0.8451
	Clustering method	PAM	K-means	PAM	K-means
	Optimal number of clusters	2	5	2	5

Table 18: summary of clustering methods and internal/ stability measures for optimal number clusters calculation in TCGA-BRCA dataset

d) Rank Aggregation:

As we saw previously, the order of clustering algorithms on each validation measure is rarely the same. Rank aggregation is helpful in reconciling the ranks and producing a "super"-list, which determines the overall "winner" and also ranks all the clustering algorithms based on their performance as determined by all the validation measures simultaneously. We therefore cluster the TCGA data using the hierarchical, K-means, and PAM algorithms with one to five clusters. Both internal and stability measures are used for validation.

The "getRanksWeights" function extracts the validation measures and order of the clustering algorithms for each validation measure to use as input for RankAggreg. The validation measures are used for calculating weighted distances. The top three ranking algorithms for each measure are given in figure 42 below:

```
> print(res$ranks[,1:3], quote=FALSE)
                1              2              3
APN             pam-2          hierarchical-2 kmeans-2
AD              kmeans-5       pam-5          kmeans-4
ADM             pam-2          hierarchical-2 pam-3
FOM             kmeans-5       pam-5          pam-4
Connectivity    hierarchical-2 hierarchical-3 hierarchical-4
Dunn            hierarchical-2 hierarchical-3 hierarchical-4
Silhouette      hierarchical-2 pam-3          pam-2
>
```

Figure 42: Ranking of the optimal number of clusters based on clustering algorithms order and validation measures

We can deduce from the second column which reflects the first order, that the partition on 2 clusters for both clustering methods (PAM and hierarchical) is the best choice according to 5 measures from 7. Therefore, two clusters perform best on all seven measures, so picking an overall winner is relatively straightforward in the case of this TCGA-BRCA validation dataset.

e) Clustering tree:

At this stage, we were able to obtain columns containing the cluster assignments from clustering TCGA-BRCA data using different methods of clustering and validation methods. This clustering information is all we need to build a clustering tree. Each column must consist of numeric values indicating which cluster each sample has been assigned to. To plot the tree, we just pass this information to the "clustree" function. This method produces a single score by considering how the samples may change grouping as the number of clusters increases. This is useful for showing which clusters are distinct and are unstable. It doesn't explicitly tell us which choice of optimal clusters is, but it is useful for exploring the possible choices.

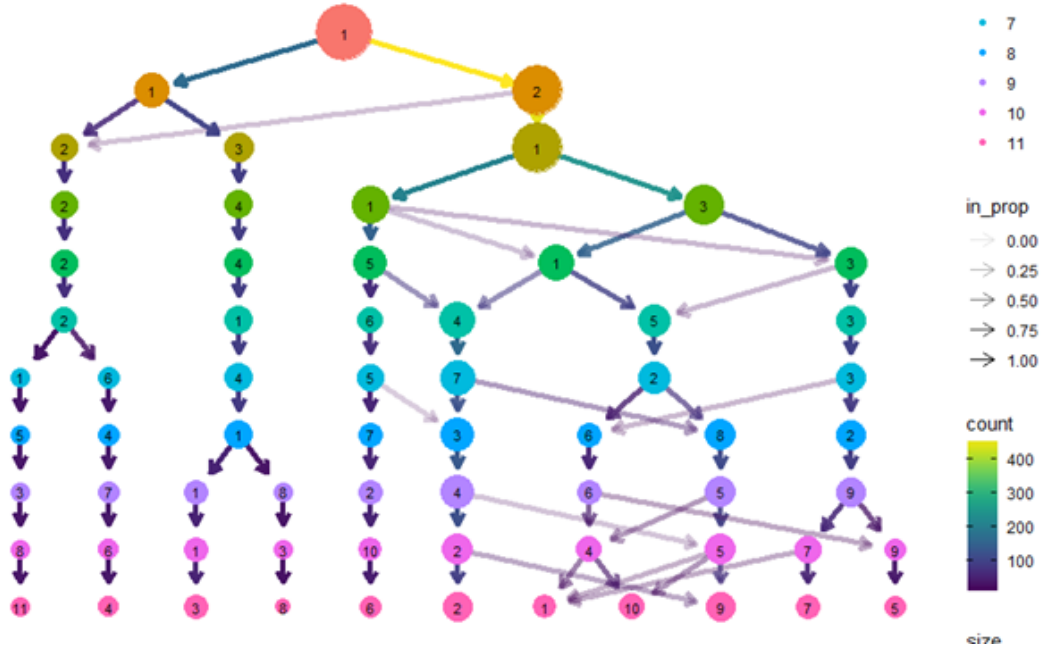

Figure 43: "Clustree" TCGA-BRCA stability plot

The size of each node is related to the number of samples in each cluster and the color indicates the clustering resolution.

Edges are colored according to the number of samples they represent and the transparency shows the incoming node proportion, the number of samples in the edge divided by the number of samples in the node it points to.

Stability index shows how stable each cluster is across the selected range of k. The stability index varies between 0 and 1, where 1 means that the same cluster appears in every solution for different k.

We can see that the first cluster of the first node (left side of the tree) is much more stable than the second one (right side of the tree). The latter is very unstable and does change with the value of k because it splits into further clusters after which the tree becomes messier and there are nodes with multiple incoming edges. This is a good indication that we have over clustered the data at this point. Therefore, over-clustered data have higher numbers of overlapping clusters. Therefore, k=2 is a good choice revealing stable clustered data.

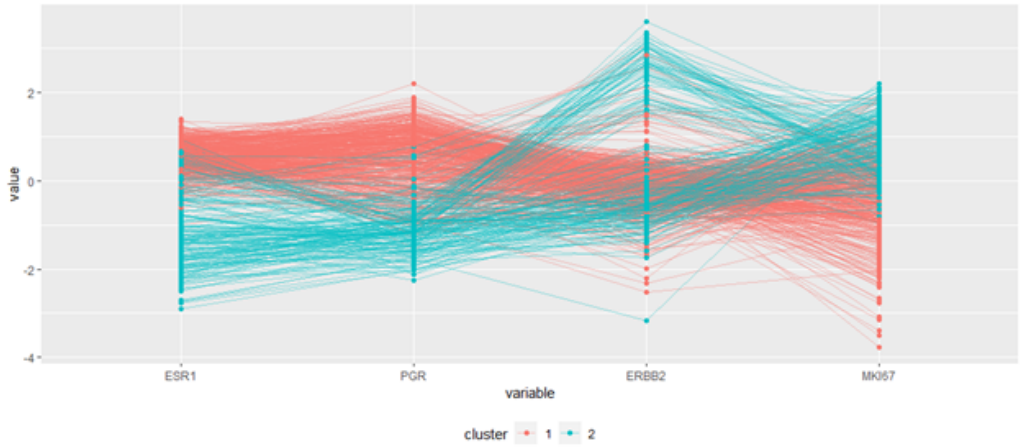

Figure 44: Parallel coordinate plot for the TCGA dataset

We plot the final decision, which refers to k=2 clusters as the most suitable number of clusters in the TCGA-BRCA dataset, in the Parallel coordinate displayed on figure 47, which is a type of visualization used to represent multivariate numeric data. Parallel coordinate plots are ideal for comparing many variables together and seeing the relationships between them.

Each of the four variables is given its own axis, all the axes are placed in parallel to each other and they all have the same scale since they have been standardized and therefore uniform. The values are plotted as a series of lines that are connected across all the axes. This means that each line is a

collection of points placed on each axis that have all been connected together. We can thus deduce that BC records clustered within the first group (C1) are much more abundant when expression values of ESR1 and PGR are positive (red lines). On the other hand, BC records clustered within the second group (C2) are more related to ERBB2 and MKi-67 overexpression (Blue lines).

B) Statistical overview of TCGA-BRCA dataset clusters:

The outcome shows that C2 contains 166 records and is mostly recognized by an underexpression of hormonal receptors (shown as negative values) genes while in the other hand, C1 contains 459 records and is associated with high expression of hormonal receptors (=0.53) genes and underexpression of ERBB2 and MKi-67 genes (-0.12 and -0.3 respectively). Then based on the expression status of these same four genes, we applied the molecular classification into Luminal A, Luminal B HER2 +, Luminal B HER2-, HER2 and Triple Negatives; then visualized the final partitions (based on molecular subgroups AND clusters membership) as represented in figure 45 below:

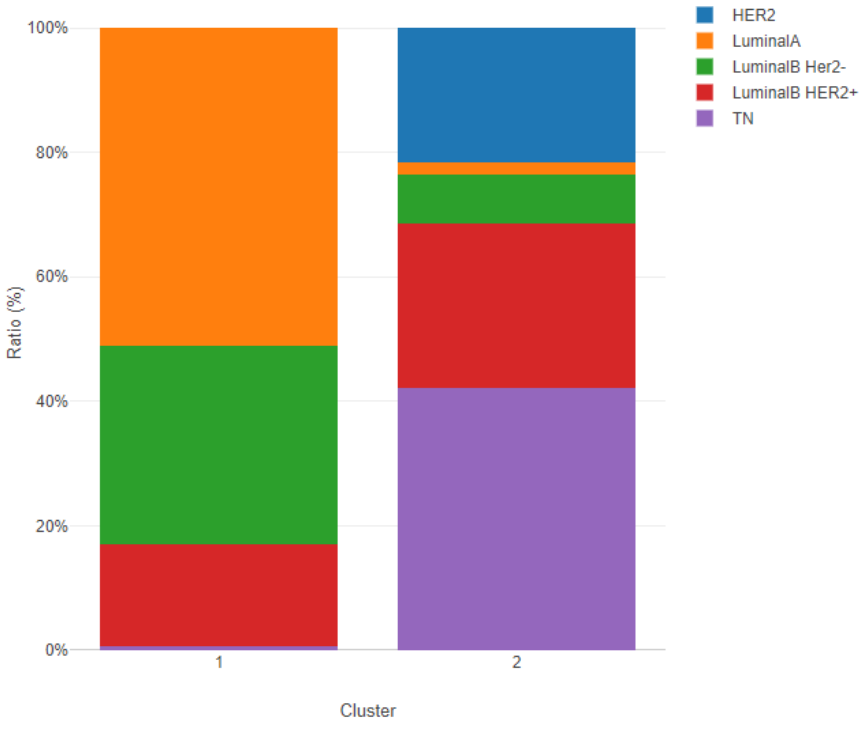

Figure 45: TCGA-BRCA records counts according to clusters and molecular subgroups membership

Figure 45 shows that all luminal subgroups belong mostly to C1. First, LuminalA constitutes 50.98% of C1 records, followed by LuminalB Her2- (32.03%), LuminalB Her2+ (16.34%) and then

TN (0.65%). Contrary, the molecular subgroups belonging to C2 are: TN; LuminalB HER2+; HER2; LuminalB Her2-, then Luminal A (42.17%, 26.51%, 21.69%, 7.83% and 1.81% of total C2 records respectively). This converges with the results obtained in the Moroccan database, which support the fact that C1 has a lower mean value of Ki-67 index, and that it is mainly associated with hormone dependent tumors (Luminal subtypes). On the other hand, C2 has a high mean value of Ki-67 index and negative mean values of hormonal receptors' gene expression. The latter is also more associated with records classified as hormone independent tumors.

a) Inter-heterogeneity of clusters distribution within molecular subgroups based on MKi-67 gene expression score:

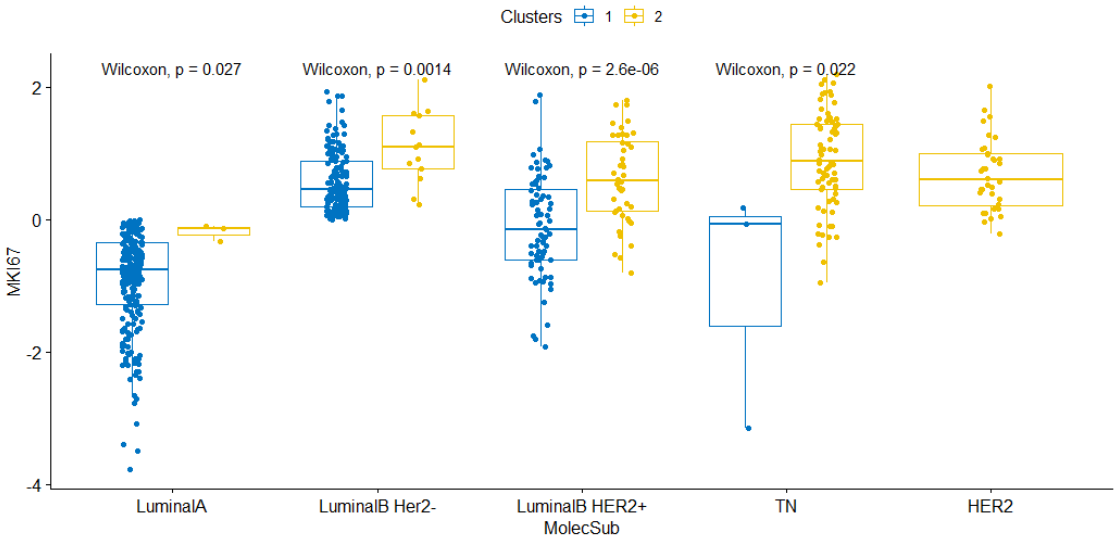

Figure 46: Distribution of TCGA-BRCA records according to their Molecular subgroups (X-axis), and clusters membership depending on MKi-67 gene expression score (Y-axis)

As shown in figure 46, MKi-67 expression score varies remarkably from one subgroup to another. In order to determine whether these intergroup variations are statistically significant in the TCGA-BRCA dataset, the Wilcoxon-test was used. According to the p-values of the statistical test, the difference in MKi-67 scores between the different clusters within the same molecular subgroup is significant.

On the other hand, for the HER2 subgroup, the statistical test was not assessed, because there was no reported HER2 record classified within C1, which is why only the C2 boxplot was visualized. It

can also be clearly seen that MKi-67 gene expression levels are relatively very low in tumors belonging to C1 compared to C2. Plus, the TN, Luminal B and HER2 subgroups have higher MKi-67 scores for both clusters compared to LuminalA.

b) Overall survival analysis on TCGA-BRCA dataset:

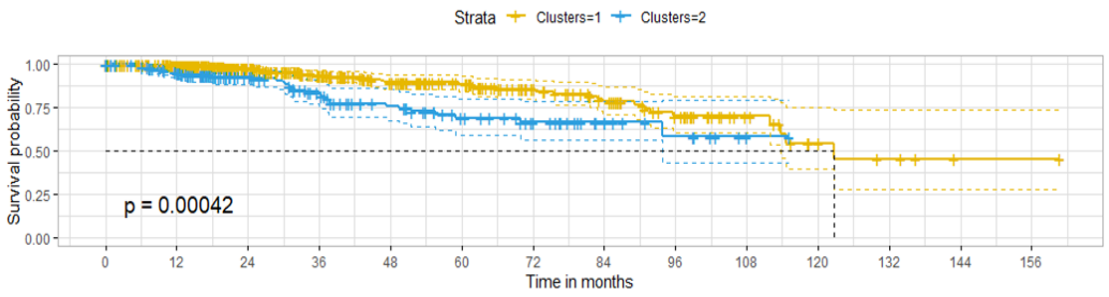

Figure 47: Summary of Kaplan-Meier Survival Curves grouped by Clusters in TCGA-BRCA dataset

After applying Kaplan-Meier analysis on the TCGA-BRCA survival data depending on clusters' membership; we obtained two distinct survival curves: the yellow curve shows patients belonging to C1‘s survival, while the blue one shows C2 patients' survival. The difference between both survival curves is statistically significant (p =0.00042).

Stratifying the same observations by Ki-67 alone reported a significant p-value of 0.002 according to the log-rank test. The average survival is approximately 10 years for C1. That of C2 could not be assessed, but it seems to have a poor survival compared to C1 all along the survival follow-up period (Figure 47).

The visualization of the OS Kaplan-Meier rates depending on the clusters and molecular subgroups belonging is presented in figure 48, where we see that there is a significant difference in survival between the different clusters within the same molecular subgroup. For example, luminalB HER2+ C1 subgroups have better survival than C2 of the same subgroup. Similarly, C2 of the LuminalB HER2- subgroup has an OS rate more defavorable than that of C1. Additionally, both TN belonging clusters have a remarkable difference in survival. This suggests that individuals routinely classified in the same molecular subgroup should be further subdivided into two other subdivisions, since their OS rate differs remarkably, so this allows to furthermore refine the prognosis.

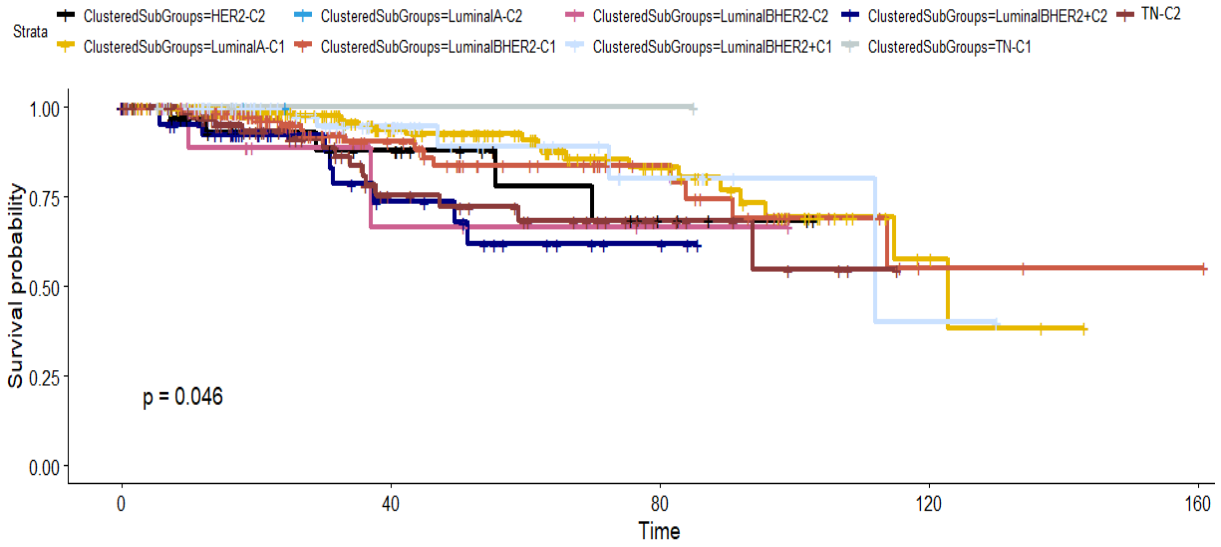

Figure 48: Summary of Kaplan-Meier Survival Curves grouped by Clusters and molecular subgroups in TCGA-BRCA dataset

In this section, we computed survival curves using the combination of both factors "clusters membership" and "Molecular subgroups membership". We visualized the output using the survminer package in figure 49. The plot shows survival curves by the "clusters membership" variable faceted according to the categories of Molecular subgroups membership. The cumulative event (reported on Y-axis) represents the cumulative incidence that estimates the risk that the BC patient will experience death during a specific time (X-axis).

First, the curves of the C2 belonging records regardless of their molecular subgroup are much less developed than those belonging to C1. This is mainly due to the event occurring at an early stage which causes the curves to be narrowed and therefore not exceed 10 years of survival at the latest. On the other hand, those of C1 have extended curves on the time axis.

The very remarkable shrinkage and flattening of the TN C1 and LumA C2 curves is due to the rare records belonging to these two subdivisions respectively, where there was no death event which caused these two curves to remain flat.

Similarly, it should also be noted that the C2 curves within each molecular subgroup show a force of mortality increase from the first months of monitoring. Unlike the C2 curves, we see that at the start of the study there is a certain curve flattening which reveals a very reduced force of mortality and which begins to rise over the follow-up period. This suggests that C2 classified patients,

regardless of the molecular subgroup, will be more likely to experience a death event early compared to those classified in C1.

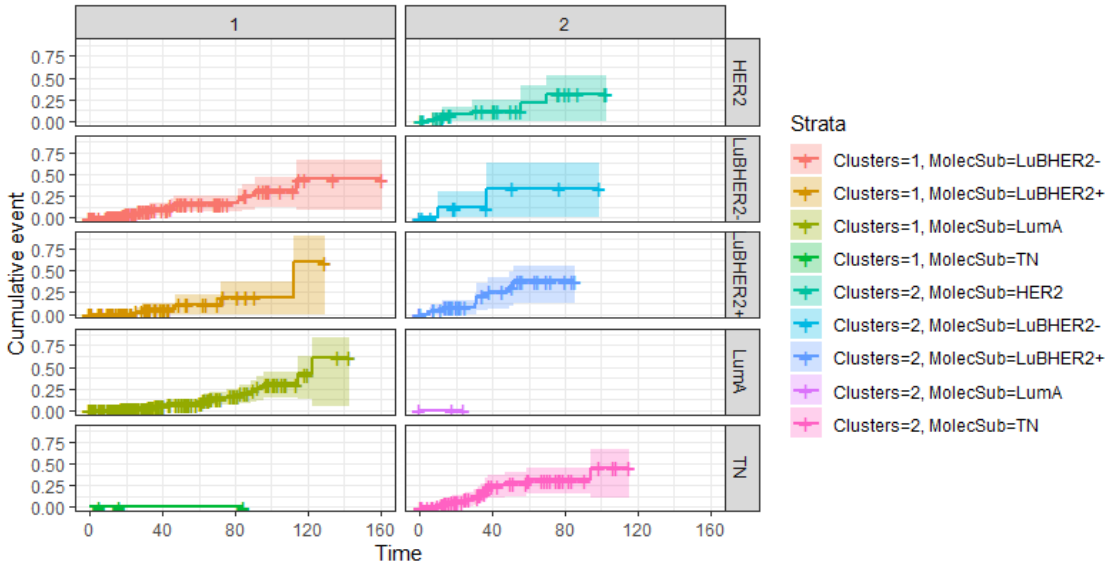

Figure 49: Clusters cumulative force of mortality depending on molecular subgroups membership for 10years of follow-Up

(Left column 1: Cluster1 membership; right column 2: cluster 2 membership)

c) Important variables selection:

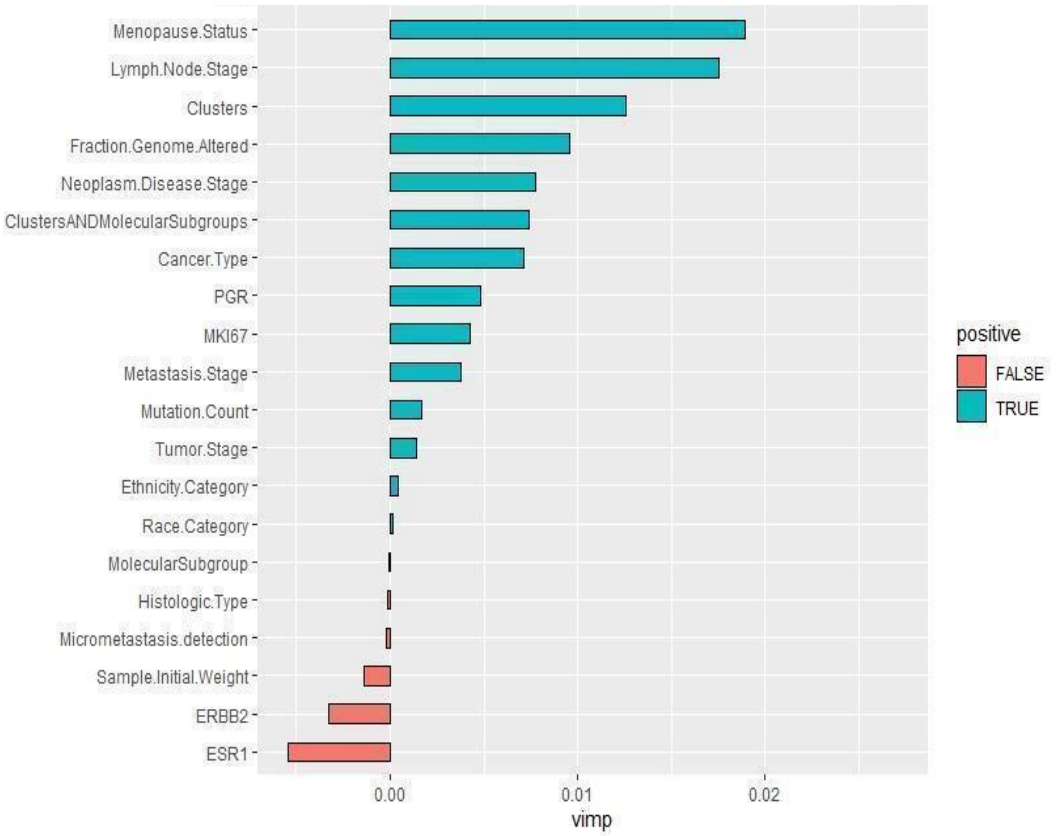

Figure 50: Important variables selection by Random forest VIMP algorithm on TCGA-BRCA dataset

(Blue bars indicate positive VIMP, red indicates negative VIMP. Importance is relative to the positive length of bars.)

The VIMP algorithm as previously applied to the Moroccan database was reproduced on this TCGA-BRCA dataset, taking into consideration the whole panoply of available variables.

Figure 50 shows graphically that histologic type; micrometastasis detection; sample initial weight; ERBB2 and ESR1 are colored by red bars, which means that they therefore generate false positive records of the RF predictive model, thus decreasing its accuracy. Their retrograded ranking makes them the least important features. On the other hand, 2-clusters partition and the partition based on clusters and molecular subgroups simultaneously, have high importance according to the RF model, contrary to the molecular subdivision which is the least informative among the positively important features. This outcome validates the importance of the two partitions we have deployed (either based on 2-clusters exclusively or with molecular classification simultaneously). These two partitions are well ahead of the standard molecular classification taken alone to predict the TCGA-BRCA records survival.

2) METABRIC dataset

Regarding the METABRIC dataset analysis; one of the main encountered limitations regards the variables typology. ER, PgR and HER2 are categorical while on the other hand, Ki-67 is the only numeric one. Clustering methods are mainly adapted for numerical variables but EM is one of the few algorithms that supports this limitation. Since it is not primarily based on a numeric matrix, as is the case with the K-Means algorithm, we opted for EM clustering for this section.

A) EM clustering on METABRIC dataset

In order to further explore the existence of an intra-molecular subgroup heterogeneity within the 1885 MEATBRIC composing records, the EM Clustering method was used. This hypothesis was tested in an unsupervised manner. During the EM clustering, the "v-fold cross-validation" algorithm was used to automatically determine the appropriate number of clusters. We found that each phenotype was statistically divided into two further subdivisions as shown in the table 19 below:

Table 19: Ki-67 distribution within Cluster1 and Cluster2 in METABRIC dataset

	Cluster1	Cluster2	Overall

Minimum	-2.42490	1.602300	-2.42490
Maximum	1.89540	4.775600	4.77560
Mean	-0.15744	2.401081	-0.00270
Standard deviation	0.80468	0.568404	0.99992

- C1 (for Cluster 1): includes patients with a low Ki-67 proliferation index (-0.15 ± 0.8 as mean Z-Score).
- C2 (for Cluster 2): includes patients with a high Ki-67 proliferation index (2.4 ± 0.5 as mean Z-Score).

The G-test application revealed a p-value = 0.0001, which testifies to a significant difference in patients' distribution within clusters.

B) Overall Survival analysis according to Clusters membership

For OS analysis, the data processing was performed using the survival data for the same 1885 clustered records, and the associated histo-prognostic features were considered predictors of survival rate. It was mainly carried out by grouping the patients by their cluster's membership, without taking in consideration their molecular subgroups. The median survival for C1's patients was found to be 6.21 years while it is 4years for C2. The difference between the latter's significant (The log rank test p-value= 0.0085) as shown in figure 51 below:

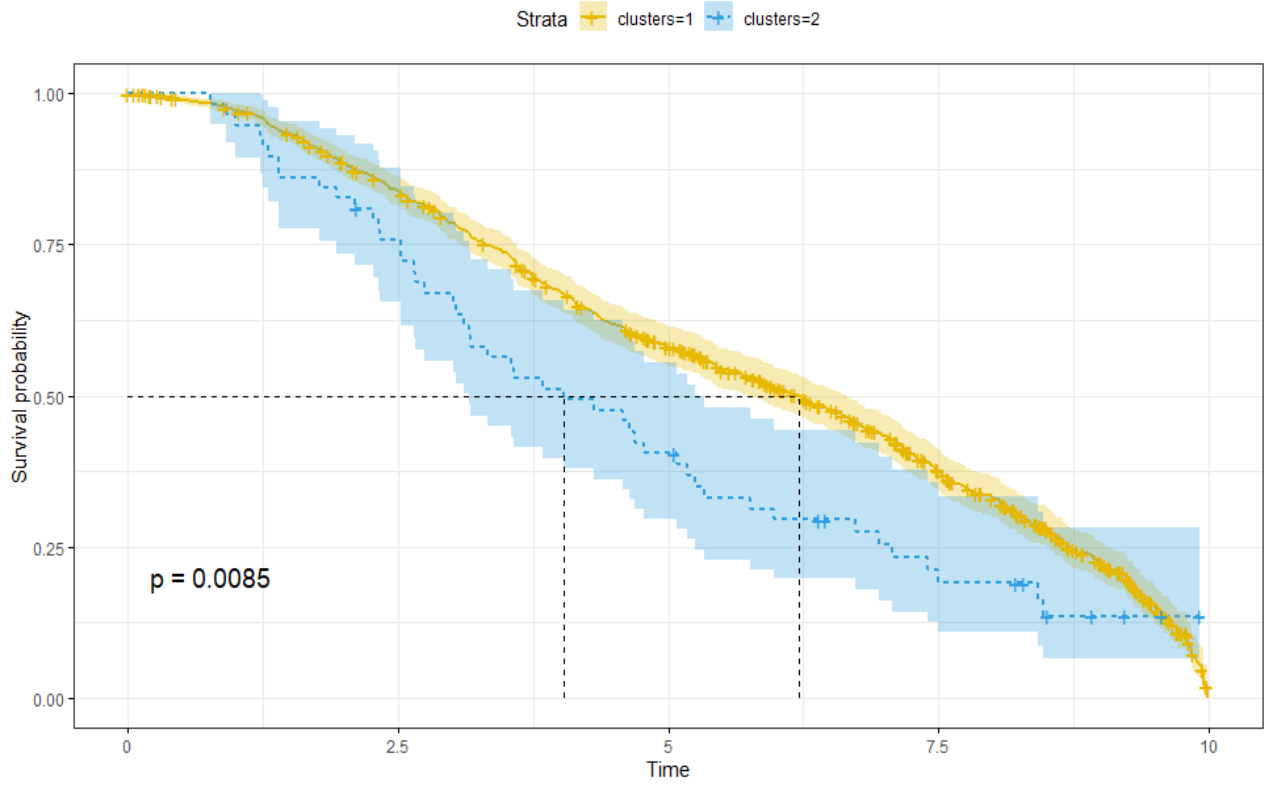

Figure 51: Overall survival analysis graph: Kaplan-Meier curves grouped by Clusters membership

C) Overall Survival analysis according to molecular subgroups and clusters belonging:

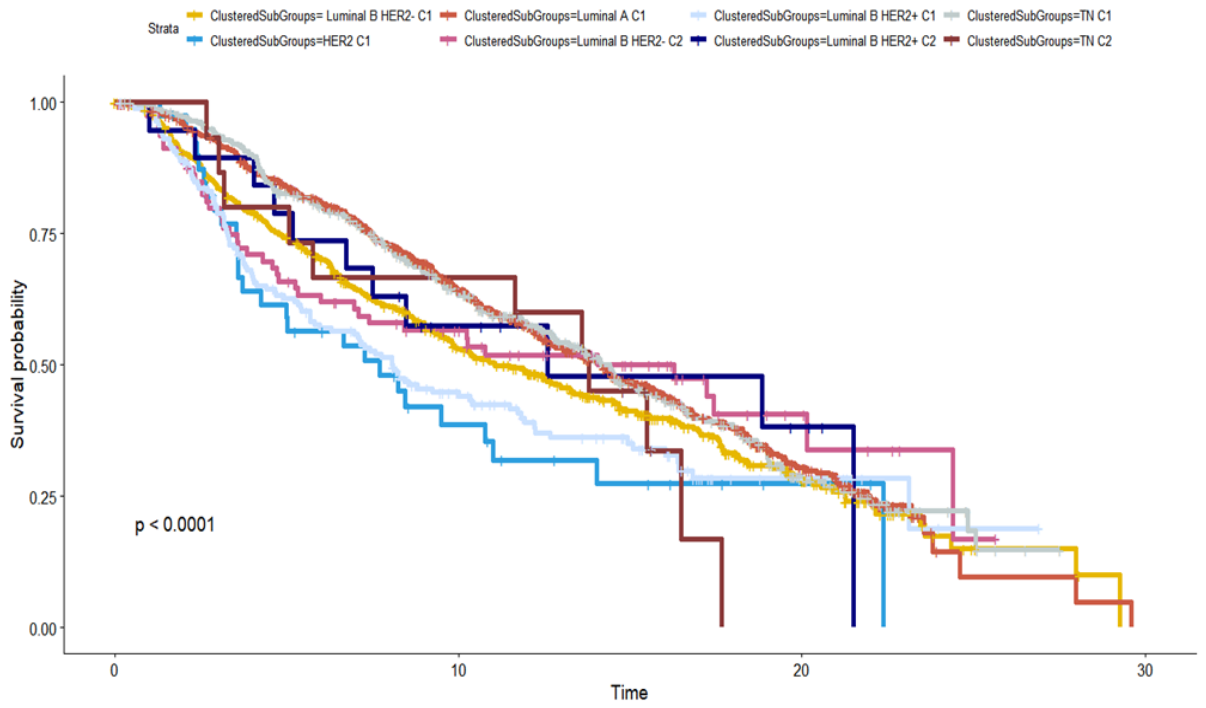

Figure 52: Overall survival analysis graph: Kaplan-Meier curves grouped by clusters and molecular subgroups

(X-axis: Survival probability, Y-axis: Time in years)

The plotted summary of the 30-year OS rates presented in the figure 52 above shows the survival of the same patients, but based on their belonging to the clusters within each molecular subgroup. Within this METABRIC survival dataset, it has to be highlighted that there was no tumor classified in the HER2 subgroup and simultaneously belonging to C2, we remark the same thing with the Luminal A subgroup and classified as C2.

The worst survival goes to records classified in C1 of HER2 subgroup by reaching 50% of OS within 8 years of followUp. Then records belonging to C1 of Luminal B HER2 +, Luminal B HER2- Cluster1, Luminal B HER2 + Cluster2, TN Cluster2 closely follows TN Cluster1, Luminal B HER2- Cluster2 and finally Luminal A Cluster1 which has the best OS.

Generally, the partition of survival curves depending on molecular subgroups and clusters belonging is more precise and significant than just the molecular classification partition, especially for the luminal B HER2+ Cluster1/Cluster2 and luminal B HER2- Cluster1/Cluster2. TN clusters,

on the other hand, don't show a real difference in terms of survival outcome, since their respective curves are almost superimposed.

After having found that the METABRIC dataset has to be optimally further subdivided in 2 clusters, we would like to assess their importance in terms of the survival outcome prediction, compared to the other histopronostic parameters recorded in the same dataset. The variable importance selection algorithm was therefore addressed, which helps us determine the most important explanatory variables that govern the survival of the patients. We consider the "ClusteredSubGroups" variable, which refers to "molecular subgroups and clusters belonging" for each METABRIC BC record, as a new variable to detect its importance ranking and therefore its influence on the survival probability prediction of those BC patients, using two algorithms of important variables selection: VIMP and Minimal Depth algorithms.

D) Important variables selection

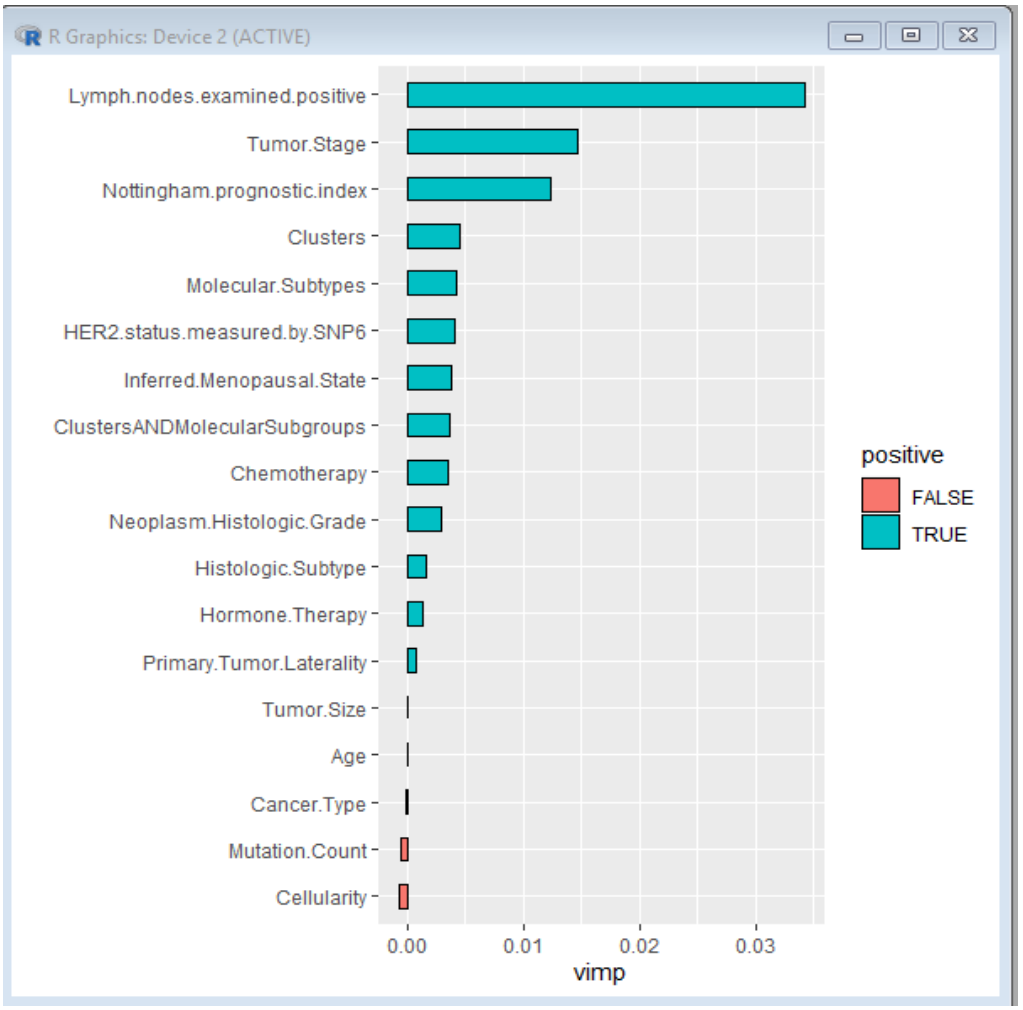

Figure 53: Important variables selection by Random forest VIMP algorithm on METABRIC dataset

According to figure 53, RF's VIMP algorithm shows that "2-clusters partition" is ranked as the 4th most important feature. The latter is followed by the molecular classification-based partition as the 5th rank. Finally, the partition combining both simultaneously under the "ClustersANDMolecularSubgroups" feature is ranked in the 8th position of the most influential variables on the MEATBRIC patients' survival. We can conclude that among all the predictive variables, both our proposed partitions are classified within the top 8 most influential ones. Therefore, they contribute the most to the predictive accuracy of the RF.

Section 2 main findings:

- -The partition of both external datasets, TCGA and METABRIC, showed that they should optimally be subdivided into two other internal classes strictly associated with the Ki-67 proliferation index, denoted C1 and C2. And this, according to several calculation methods evaluated by several measurement indices.
- -C1 is essentially characterized by a low average of Ki-67 compared to that of C2, whether measured by IHC (% of staining) or gene expression methods (Z score). The average of C1 is always much lower than that of C2.
- -Clinically known molecular subgroups with poor prognosis such as Pure HER2 and TN are more overrepresented in C2, unlike luminal subgroups which mostly belong to C1.
- -The establishment of OS rates analysis by assessing Kaplan-Meier curves made it possible to make a distinction in BC patients' survival, depending on their clusters' belonging.
- -Mostly, the average survival of C1's patients remains significantly more favorable than that of patients belonging to C2, thereby confirming our first main findings on the Moroccan dataset.
- -Overall, this new refinement of the molecular classification is important as it is voted by two different algorithms in survival prediction, compared to several other prognostic, genetic and molecular criteria. All the three analyses (Moroccan, TCGA and METABRIC) converge on this point.
- -We have been able to reduce the whole panoply of histopathological factors to a few predictors, which can predict cluster's membership and survival with better accuracy.

- -In countries with limited means and limited resources which cannot use genomic signatures, it has been shown that the cluster classification can give us an idea of the patients' prognosis, and this, by the use of few predictive factors taken into consideration routinely in all the pathology laboratories.

Discussion

This study explores the possibility of partitioning BC molecular subgroups in order to better define patient survival. The partitioning approach was applied starting from 1128 Moroccan BC records, and then tested on two external independent datasets, also used to further validate the clinical significance of the newfound subdivisions, in terms of survival prediction.

We found that the routinely established BC molecular classification could be further refined by only using Ki-67, ER, PgR and HER2 variables. Each subtype can be subdivided in two distinct clusters with significantly different Ki-67 distribution and survival outcome. Tumors belonging to the cluster with low mitotic activity (C1) are overrepresented in the luminal subtype, while HER2 and TN subtypes are enriched in tumors belonging to the cluster with high mitotic activity (C2). This partitioning is also associated with OS and is equally or even more important than tumor size in predicting outcome. Indeed, Marwah et al. (Marwah et al. 2018)revealed that a higher Ki-67 was found with a size greater than 5cm while tumors smaller than 2cm showed a lower rate. This finding was also confirmed by Querzoli & al(Querzoli et al. 1996). Similarly, several studies showed a positive correlation between Ki-67 and the tumor histological grade. However, we can point out that our analysis on the Moroccan dataset did not rank histological grade, unlike cluster membership and tumor size, among the most important variables able to predict survival.

Another interesting result is the presence of C1 samples within TN tumors. Histological subtype could in part explain this result, with cystic adenoid carcinoma as a typical example. Although it does not express hormone receptors and does not overexpress HER2, it shows a low proliferation index (Bouzubar et al. 1989) (D.-Y. Wang et al. 2019). Ki-67 is also reported to be higher in TNBC of no special type compared to TNBC of other histological subtypes (Tan et al. 2004). It has been

suggested by some authors as a prognostic and predictive marker for TNBC (Zhu et al. 2020). Keam's study proposed a 10% Ki-67 cut-off to define two different prognostic subgroups: the first one with high Ki-67, which despite a better response to chemotherapy was more aggressive, and the second one with low Ki-67, which showed a lower aggressiveness but also a lower response to chemotherapy (Keam et al. 2011; Bartlett et al. 2016)). Our results confirm these findings, with TNBCs partitioned into 2 further subdivisions: C1 and C2, with mean Ki-67 of 16.4±13% and 73±15.4%, respectively.

On the other hand, Bartlett & al. recently reported that, despite little concordance at the single tumor level, heterogeneity within ER+ tumors in terms of prognosis is detectable and confirmed by different BC multiparameter tests (Bartlett et al. 2016). In addition, Aleskandarany & al. confirmed that within the Luminal B / HER2- subgroup, the group with high proliferation index had worse evolution and prognosis than the group with low Ki-67. This supports our results, with some cases within the luminal B / HER2- molecular subgroup clustered in C2 and showing worse survival in all the three analyzed datasets. Therefore, BC molecular subgroups should be considered as a spectrum of diseases (Bartlett et al. 2016).

In terms of survival, we showed that subtype partitioning helps to refine the prognosis. Kyung Lee and al. demonstrated that the combination of p53 and Ki-67 has the best predictive power, especially for long-term OS in the Luminal A subgroup (Lee et al. 2015). Interestingly, in our study we were able to highlight that not only Luminal A but also the other molecular subtypes, Luminal B in particular, could benefit from a further refinement based on Ki-67. It should be noted that luminal tumors would represent a genotypically heterogeneous group with some tumors exhibiting chromosomal instability with aneuploidy and others without genomic instability and diploids (Yanagawa et al. 2012). On the other hand, no significant prognostic difference could be established for HER2 and TN tumors, since for our Moroccan cohort survival information was complete only for a small subset of patients. Therefore we repeated the analyses on TCGA-BRCA and METABRIC, which allowed us not only to extend and validate our clustering results to a broader context, but also to confirm that the intrinsic subdivisions proposed within the molecular subgroups

have a clinical prognostic impact. TCGA-BRCA and METABRIC contain a very wide panoply of prognostic features and cover even genomic data for each patient, therefore they could be further exploited to identify additional biomarkers translatable to the clinics.

Conclusion

The efforts made over the past 10 years to better understand the histopathology of IBC have led to an important conclusion: the latter covers different cancers. Some of them have a completely defined “molecular portrait”, which can be identified by genomic methods. However, if the path to the integration into clinical practice of a molecular and morphological portrait is still long, it will nevertheless have to be done, in order to offer a more successful diagnosis for patients and physicians. Genomic profiling can be very time-consuming and expensive, so we tried to define a new superposition between the histopathological and bioinformatic analysis; to define a novel refinement of BC molecular classification which also turned out to be one of the most important variables in terms of survival prediction.

Rapport récapitulatif en français:

Introduction :

Le carcinome mammaire ou cancer du sein (CS) est une pathologie aux caractéristiques cliniques, histopathologiques et moléculaires bien définies. La morphologie des tumeurs a été l'étalon-or pour classer les tumeurs mammaires en entités au pronostic défini.

Cependant, la classification morphologique traditionnelle présente des limites, ce qui laisse la place à de nouvelles méthodes moléculaires censées améliorer la stratification des patients et la prédiction de leur pronostic.

Cette étude est un travail de collaboration entre l'Italie et le Maroc ; par conséquent, le présent manuscrit est conventionnellement divisé en deux chapitres principaux et distincts, mais évoquant des axes complémentaires concernant le raffinement de la classification du CS. Le premier chapitre évoque un thème élaboré au Laboratoire de génomique du cancer, Fondazione Edo ed Elvo Tempia, Biella (Italie), dans le cadre de l'école doctorale « Complex Systems for Quantitative Medicine » de l'université de Turin.

Le second chapitre évoque un autre thème élaboré principalement dans le service d’anatomopathologie de l'hôpital universitaire ibn Rochd de Casablanca/Laboratoire de pathologie moléculaire et de génétique ; Faculté de médecine, Université Hassan II de Casablanca-Maroc.

La principale question abordée dans le premier chapitre concerne plus spécifiquement le sous-typage des cancers du sein triple négatif (CSTN), selon la combinaison de plusieurs techniques bio-informatiques et génomiques. L'intérêt principal est donc d'explorer l'hétérogénéité de la

maladie et de définir des sous-types de cancer du sein ainsi qu'un répertoire de plus en plus précis d'altérations génétiques, qui permettront de différencier de manière distinctive un sous- groupe triple négatif d'un autre. Ainsi, afin de garantir une certaine reproductibilité généralisée des résultats que nous proposons, nous avons également eu recours à une validation externe basée sur les ensembles de données CSTN-TCGA et italiens.

D'autre part, la question principale abordée dans le deuxième chapitre évoque si la classification moléculaire actuellement acceptée comme référence pour la détermination des sous-types de CS peut être affinée par des méthodes de partitionnement statistique. Ceci est particulièrement utile dans les pays à faible revenu comme le Maroc, où les laboratoires n'ont pas nécessairement accès aux nouvelles méthodes de tests moléculaires qui s'avèrent très coûteuses comme les tests de profilage d'expression génétique par exemple.

Par ailleurs, notre travail vise également à évaluer le degré de prédiction de ces approches statistiques et la précision de leur classification. D'où l'intérêt d'avoir utilisé principalement une grande base de données marocaine, qui a été comparée à deux autres bases différentes : METABRIC et TCGA-BRCA qui ont servi de validation externe pour notre étude comparative de cohorte nord-africaine.

Chapitre 1 : Identification d'un nombre minimal de gènes pour prédire les sous-groupes de CSTN à partir des profils d'expression géniques

Contexte :

Les cancers du sein triple-négatifs (CSTN) touchent environ 15 % des femmes présentant des tumeurs mammaires. La dénomination CSTN est une définition immunohistochimique correspondant à l'absence d'expression des récepteurs aux œstrogènes (ER) et à la progestérone (PgR), et de l'amplification du récepteur 2 du facteur de croissance épidermique humain (HER2). Les seuils de négativité retenus par les directives de l'American Society of Clinical Oncology sont moins de 1 % de cellules marquées pour les récepteurs hormonaux (Hammond et al. 2010), et des scores de 0+, 1+ ou 2+ pour le marquage de HER2 mais sans amplification par hybridation in situ en fluorescence (FISH) pour ERBB2 (Hwang et Gown 2016).

Les CSTN sont des carcinomes canalaires de grande taille, de haut grade, avec un indice mitotique Ki67 élevé et de nombreuses atypies nucléaires à l'examen anatomo-pathologique (Carey et al. 2007). Ces cancers sont souvent apparentés au sous-type basal, introduit pour la première fois par et (Perou et al. 2000 ; Podo et al. 2010) dans leur travail princeps, et présentent des similitudes avec les cancers développés sur mutation germinale *BRCA*. Le sous-type basal-like (BL) est caractérisé par une surexpression du gène de la cytokératine basale et l'absence d'expression des gènes codant pour les œstrogènes, la progestérone et HER2. Des mutations du gène *BRCA1/2* sont trouvées dans environ 30 % des cas (Matros et al. 2005). Les CSTN sont généralement de grandes tumeurs de haut grade associées à un âge plus jeune au moment du diagnostic, avec un profil agressif et des taux élevés de mutations du gène *p53*, accompagnés d'une forte détection immunohistochimique de *p53* (Dentet al. 2007). Ils présentent donc un risque élevé de rechute, malgré une plus grande

sensibilité à la chimiothérapie,et de récidive métastatique dans les trois premières années après le diagnostic. Elles ne sont pas éligibles aux traitements ciblant les récepteurs hormonaux ou HER2. Cependant, en plus de la chimiothérapie, ces cancers peuvent bénéficier de nouvelles options thérapeutiques, selon la nature de la tumeur. Depuis 2005, le développement intensif des technologies à haut débit pour analyser le statut et/ou l'expression des mutations génétiques, a permis d'accroître les connaissances sur le profil génotypique et phénotypique des CSTNs (Geyer etal. 2009). Premièrement, plusieurs sous-catégories peuvent être identifiées en analysant leur morphologie et certaines ont soit un pronostic particulier, soit une réponse thérapeutique spécifique. Deuxièmement, les technologies à haut débit, grâce à l'analyse de milliers de gènes, ont commencé à mettre en évidence des sous-classes moléculaires de CSTN, présentant des anomalies moléculaires spécifiques associées à la réponse au traitement et/ou à la survie. Troisièmement, les preuves se sont accumulées, montrant que le microenvironnement du CSTN, c'est-à-dire les cellules et les molécules présentes dans le stroma tumoral, joue un rôle important dans la progression de la maladie. Ainsi, les caractéristiques du microenvironnement peuvent servir de nouvelle base de sous-classification du CSTN avec un impact thérapeutique potentiel (H. Zheng et al. 2021). En 2011, un groupe de chercheurs dirigé par (Lehmann et al. 2011) à l'Université Vanderbilt, a évalué une nouvelle classification, appelée CSTNtype-6, par laquelle ils ont procédé à l'identification de six sous-types de CSTN, sur la base du profilage de l'expression génétique de plusieurs centaines d'échantillons de CSTN. Diverses anomalies d'expression liées aux gènes régulateurs du cycle cellulaire, tels que les gènes de réparation de l'ADN *BRCA2* et *TP53*, ont été détectées dans le sous-type BL1 (basal-like type 1). Le deuxième sous-type de type basal (BL2) était davantage associé à une activation anormale d'autres voies de signalisation, telles que EGFR, MET, la migration cellulaire, l'interaction matrice extracellulaire-récepteur et la différenciation. À l'inverse, le sous-type MSL (cellules souches mésenchymateuses) était davantage associé à une sous-expression de la prolifération cellulaire et à une surexpression des gènes liés aux cellules souches mésenchymateuses. Le sous-type IM (immuno-modulateur) était principalement reconnu par les voies de transduction du signal immunitaire, telles que celles des cellules NK, B,

dendritiques et T. Le sous-type M, quant à lui, était enrichi par les voies de signalisation liées à la migration cellulaire ainsi que par les voies d'interaction matrice extracellulaire-récepteur et de différenciation. Le sous- type LAR (luminal androgen receptor) était très différent de tous les autres : bien que négatif pour le récepteur ER, il exprimait le récepteur des androgènes (AR) et/ou ses effecteurs en aval, et était fortement associé aux voies de signalisation liées aux hormones, comme la synthèse des stéroïdes et le métabolisme des androgènes/œstrogènes.

En 2016, le même groupe de chercheurs a affiné la classification susmentionnée car il a observé une présence significative de lymphocytes infiltrant la tumeur (TIL) et de cellules stromales dans les sous-types IM et MSL, respectivement.

Ainsi, les précédents sous-types de CSTN ont été affinés en BL1, BL2, M et LAR, ce qui a donné lieu à la classification CSTNtype-4 (Lehmann et al. 2016). Par la suite, Burstein a supervisé une autre étude visant à identifier les marqueurs distincts qui caractérisent chaque sous-type de CSTN. Il s'est avéré qu'en plus de l'analyse des variations du nombre de copies (CNV), les techniques de profilage génomique peuvent être utilisées pour stratifier davantage les tumeurs mammaires triple négatif. En conséquence, quatre sous-types différents ont été trouvés, avec un pronostic distinct et stable, étiquetés comme suit : LAR, Mesenchymal (MES), Immunosuppressed Basal Type (BLIS) et Immune Activated Basal Type (BLIA) (Burstein et al. 2015). D'autre part, dans une étude plus récente de Jézéquel et al. par des techniques de profilage transcriptomique, trois sous-types distincts ont été mis en évidence. Le premier est reconnu par un phénotype moléculaire apocrine présentant un pronostic favorable, les deux autres groupes avaient des propriétés plus basales : tandis que l'un était plus agressif et couplé à un phénotype immunosuppresseur, le troisième présentait une réponse immunitaire adaptative (Jézéquel et al. 2019 ; Ensenyat-Mendez et al. 2021). Enfin, une autre étude développée par Liu et al. et basée essentiellement sur les ARN longs non codants (lncRNA) pour classer les tumeurs CSTN, a abouti à l'élaboration du système de classification de l'Université Fudan de Shanghai (FUSCC). Quatre sous- types ont été reconnus : IM, LAR, MES et BLIS, présentant une régulation élevée des voies prolifératives et la pire survie globale. (Y.-R. Liu et al. 2016). Cependant, les événements moléculaires moteurs potentiels au sein de chaque sous-type de

CSTN, ainsi que leur réponse au traitement, restent rarement explorés. Il est donc nécessaire d'approfondir les connaissances sur les altérations génomiques sous-jacentes, ainsi que vers une sous-classification standardisée et facilement applicable. Les efforts déployés ces dix dernières années pour mieux comprendre la biologie du CSTN ont abouti à une conclusion importante : le terme "triple négatif" recouvre différents cancers. Certains d'entre eux ont un "portrait moléculaire" complètement défini, qui peut être identifié par des méthodes génomiques. Cependant, si le chemin vers l'intégration dans la pratique clinique d'un portrait moléculaire et morphologique est encore long, il devra néanmoins se faire pour offrir un diagnostic plus précis ainsi qu'un traitement plus personnalisé aux patients. Dans cette même perspective et en partant principalement de la classification de Lehman, nous avons cherché à identifier un nombre limité de gènes pouvant servir de signature génétique pour la prédiction des différents sous-types de CSTN.

Matériels et méthodes

- Description des bases de données:

Deux ensembles de données CSTN ont été téléchargés à partir de plateformes à accès public. Le premier a été extrait du Gene Expression Omnibus (GEO) et fait référence au séquençage de l'ARN du transcriptome entier (RNAseq) effectué sur des biopsies de recherche avant traitement de l'étude de phase III de BrighTNess (AFT-04). Cet ensemble de données (GSE164458) est constitué de valeurs d'expression RNA-seq log-normalisées de tumeurs de stades cliniques II à III. Il sera appelé GEO-TN.

Le second a été récupéré sur le portail de données Genomic Data Commons (GDC) du National Cancer Institute et réfère au projet d'atlas du génome du cancer (TCGA) : seuls les échantillons CSTN ont été sélectionnés, sur la base de leur statut immuno-histochimique négatif pour ER, PgR et HER2, pour un total de 63 enregistrements CSTN sur 1093 enregistrements de CS invasif. Cet ensemble de données contient des valeurs d'expression RNA-seq log-normalisées et des données cliniques. Il sera appelé TCGA-TN.

Le troisième ensemble de données a été téléchargé dans le dépôt public sous le numéro d'accès GEO GSE206912, et réfère à 72 CSTN de patients italiens traités chirurgicalement à l'hôpital de Biella et au Policlinico Gemelli à Rome, qui ont subi un profilage de l'expression génique au laboratoire de génomique de la Fondazione Edo ed Elvo Tempia, Biella (Italie). Il sera appelé Italian-TN. Le prélèvement des échantillons a été approuvé par les comités d'éthique de Novara et du Policlinico Gemelli (Prot. 861 CE 149/19 et Prot. 3559, respectivement). Après sélection de la zone tumorale, l'ARN total a été isolé à partir de coupes de macro dissections à l'aide du kit FFPE Absolutely RNA d'Agilent ; il a été transcrit de manière inverse en ADNc correspondant et transcrit in-vitro avec le kit d'amplification du transcriptome entier TransPlex de Sigma; l'ADNc a été amplifié et marqué avec le kit de marquage d'ADN SureTag d'Agilent; hybridation à l'aide du kit d'hybridation d'expression génétique Agilent sur des microarrays SurePrint G3 Human GE 8x60K V3 du génome entier contenant des sondes pour 26 803 ARN codants et 30 606 lncRNA ; les lames ont été lavées à l'aide du kit de tampon de lavage d'expression génétique, puis scannées avec le scanner Agilent version C. Tous les protocoles et kits ont été achetés auprès d'Agilent technologies. Après la numérisation, l'analyse de l'image de la matrice a été effectuée à l'aide du logiciel Agilent Feature Extraction v12.1, puis les données d'expression brutes ont été traitées par soustraction du bruit de fond suivie d'une normalisation du quantile entre les matrices, à l'aide du paquet LIMMA (LInear Models for Microarray Analysis) du logiciel R. Cet ensemble de données contient des intensités normalisées en logarithme.

- Prédiction des sous-types CSTN:

Avant la prédiction du sous-type, l'extension "*dplyr*" sur R a été utilisée pour éliminer les gènes non exprimés dans tous les échantillons (avec valeur d'expression nulle). Les données prétraitées des ensembles de données GEO-TN, TCGA-TN et Italian-TN ont ensuite été téléchargées dans l'outil en ligne TNBCtype, qui vérifie d'abord la présence de tout échantillon positif pour les récepteurs hormonaux et le supprime. Il calcule ensuite la corrélation de Spearman (et sa significativité) entre chaque échantillon et les six centroïdes des sous-types CSTNs précédemment déterminés et affecte les échantillons au sous-type le plus corrélé. UNS est attribué aux échantillons instables, avec une

corrélation très faible et non statistiquement significative avec n'importe quel sous-type. Les échantillons UNS ont été exclus des analyses en aval.

- Détermination de la signature génétique :

Cette étape a été élaborée par "R software for Statistics v.4.1.0" et basée sur le calcul des gènes différentiellement exprimés (DEG) spécifiques à chaque sous-type de CSTN, par opposition aux autres. Deux méthodes différentes ont été sélectionnées pour obtenir le meilleur choix de DEGs. La première était une comparaison de classe utilisant le paquet LIMMA, où les gènes exprimés de manière différentielle entre chaque sous-groupe CSTN prédit et les échantillons restants ont été obtenus. La détection de l'expression différentielle des gènes a été effectuée en appliquant un seuil aux valeurs p ajustées de Benjamini & Hochberg (<0,01). La deuxième méthode utilisée a été la différence de moyenne basée sur le test U de Mann-Whitney (MWU), en utilisant la même méthode pour ajuster les valeurs de p pour les comparaisons de tests multiples. La détection de l'expression différentielle des gènes a été effectuée en appliquant un seuil aux valeurs p ajustées (<0,01) et à la difference d'expression m"diane entre les sous-groupes (LogFC>1 et <1) pour les genes régulés à la hausse et à la baisse, respectivement. Les résultats des deux méthodes ont été combinés par la fonction "*merge*" de l'extension "*dplyr*" sur R pour une analyse beaucoup plus approfondie.

- Analyse du réseau des sous-types de CSTN et identification de cibles médicamenteuses

L'analyse fonctionnelle des gènes différentiellement exprimés a été réalisée à l'aide de l'outil en ligne MetaCore™ version 22.1 de la suite logicielle (Clarivate Analytics, Philadelphie, PA, États-Unis). L'analyse du réseau de gènes a été réalisée à l'aide de l'algorithme du plus court chemin de Dijkstra pour trouver le chemin le plus court entre les paires de gènes (ou de produits géniques), dans chaque direction, en tenant compte d'une étape (directe) ou de deux étapes (un objet de réseau supplémentaire comme interaction intermédiaire).

En ce qui concerne l'analyse des cibles médicamenteuses, nous avons recherché les interactions médicament-cible thérapeutiques, c'est-à-dire celles qui sont validées expérimentalement, et les interactions médicament-cible secondaires, qui sont simplement prédites sur la base de similitudes dans les structures.

- Prédiction de sous-types selon la signature génétique

Cette étape a été évaluée par "Weka v3.9.3 software for data mining". L'appartenance à un sous-type a été considérée comme la variable d'intérêt, tandis que tous les autres attributs (gènes sélectionnés) ont été utilisés comme variables prédictives. Des algorithmes d'apprentissage automatique pertinents ont donc été sélectionnés pour comparer et évaluer les performances des modèles. Les modèles suivants ont été utilisés : Naive Bayes (NB); régression logistique (LR) ; arbre de décision (DT) ; forêt aléatoire (RF) ; machine à vecteur de support (SVM) ; classificateur K-voisins les plus proches (KNN) ; perceptron multicouche (MP).

L'analyse comprend une ingénierie automatique des caractéristiques, basée sur une validation croisée à k volets, où l'échantillon original est divisé en k sous-ensembles. Le modèle a été entraîné sur tous les sous-ensembles sauf un (k-1), puis évalué sur le sous-ensemble qui n'a pas été utilisé pour l'entraînement. Ce processus de validation croisée a été systématiquement répété k fois, où chacun des k sous-ensembles a été utilisé exactement une fois comme données de validation (et exclu de la formation) à chaque fois. Les résultats des k itérations ont ensuite été moyennés (ou combinés d'une autre manière) pour produire une seule estimation finale. K a été fixé à 10.

- Métriques d'évaluation des prédictions

Chaque modèle de prédiction a été évalué par dix mesures différentes, à savoir: Taux de vrais positifs (TP) ; Taux de faux positifs (FP) ; Exactitude ; Kappa de Cohen ; Précision ; Rappel ; F-mesure ; Coefficient de corrélation de Matthews (MCC) ; Courbe (ROC) ; Aire de la courbe précision-rappel (PRC).

- sélection des meilleurs attributs:

Cette étape était utile pour choisir un petit sous-ensemble d’attributs (gènes) qui s’avère suffisant pour classer efficacement la classe cible (sous-type CSTN), en réduisant le coût de calcul et en améliorant la précision. En conséquence, la qualité de la prédiction de chaque gène de l'ensemble de données d'entraînement a été évaluée et les gènes qui ont fourni moins de valeur (votés par la règle de la majorité des différents algorithmes de sélection d'attributs) ont été écartés. Sept algorithmes de

sélection d'attributs différents ont été utilisés par le logiciel Weka : Corrélation de Pearson ; Gain d'information ; Incertitude symétrique ; Sous-ensemble Cf ; Rapport de gain ; Relief F ; One R. Leur hypothèse centrale est que les ensembles d'attributs importants sont fortement corrélés avec la classe cible, et que les attributs non corrélés sont moins importants. La liste finale des méthodes de sélection des attributs rassemble les résultats du classement de tous les attributs, du plus important au moins important. Seuls les gènes classés comme non importants par au moins quatre des sept algorithmes ont ensuite été mis en évidence comme étant les attributs les moins importants.

Résultats

1. Prédiction des sous-types de CSTN et détermination des signatures génétiques

Les trois ensembles de données CSTN ont été sous-typés à l'aide de l'outil en ligne CSTNtype. Pour l'ensemble de données GEO-TN, il y avait 23 échantillons ER+ détectés et 64 échantillons prédits UNS, qui ont été écartés. Par conséquent, le nombre final d'échantillons obtenus était de 395. Cet ensemble de données est de loin le plus important et a été utilisé comme ensemble d'entraînement. L'ensemble de données TCGA-TN était initialement composé de 63 enregistrements, dont 13 instables ont été écartés, ce qui a donné 50 échantillons CSTN. 17 échantillons ont été prédits comme UNS et ont donc été automatiquement éliminés de l'ensemble de données Italian-TN, ce qui a donné un nombre final de 55 échantillons. Ces deux derniers ont été utilisés comme ensembles de validation. Les sous-types IM et M sont les plus répandus, tandis que BL2 et LAR sont les moins fréquents, ce qui peut nous donner une idée du déséquilibre des sous-groupes.

Les deux tests utilisés pour déterminer les gènes différentiellement exprimés ont convergé vers les gènes les plus significatifs au sein de chaque sous-groupe, contrairement aux autres. Par la suite, deux listes de gènes ont été générées, la première avec les 120 gènes les plus sur-exprimés et la seconde avec les 81 gènes les plus sous-exprimés.

2. Analyse des réseaux des sous-types CSTN

Il est très intéressant de rechercher des interactions génétiques au sein des quelques gènes caractéristiques des sous- groupes CSTN. Cela peut conduire à une meilleure compréhension des

phénotypes spécifiques des sous- groupes CSTN qu'en considérant simplement les effets d'un seul gène. Afin d'identifier les voies complexes qui contrôlent les fonctions essentielles dans la cancérogenèse spécifique des sous- groupes CSTN, nous avons analysé les réseaux de gènes en utilisant la fonction " chemins les plus courts " de la suite d'analyse Metacore, en autorisant un maximum de 2 étapes (un élément supplémentaire comme intermédiaire) pour connecter les gènes dans le chemin. Nous avons trouvé des interactions entre chaque gène spécifique du sous-type (ou son produit) et d'autres entités telles que des protéines de liaison, des enzymes, des facteurs de transcription, des protéines kinases et des récepteurs à activité enzymatique, par le biais de différents mécanismes de régulation. Tous les gènes sur-exprimés dans BL1, à l'exception de *KLRG2*, sont connectés via un ou deux facteurs de transcription, *ELF5*, *PADI2*, Matrilysine (*MMP-7*), *COBL* et *CLSP* étant les gènes de signature les plus interconnectés et HNF3-alpha, les récepteurs d'androgènes et d'œstrogènes les facteurs de transcription intermédiaires les plus interconnectés. Parmi les gènes sous-exprimés dans BL1, seuls *IGF-2* et *PRSS11* (HtrA1) sont connectés via la Vitronectine ou l'IBP et la localisation de ces quatre protéines est extracellulaire. En ce qui concerne les gènes sur-exprimés dans BL2, la plupart d'entre eux codent pour des protéines cytoplasmiques régulées par quelques facteurs de transcription intermédiaires (p53, STAT3, RAR-alpha, récepteur des androgènes, FKHR), à l'exception de la Calgranuline A cytoplasmique qui est directement liée à la Calgranuline B extracellulaire via une boucle autorégulatrice (activation mutuelle par liaison). La S100-A16 n'est liée à aucun autre gene sur-exprimé, tandis que le seul autre produit extracellulaire, la Stromelysine-1, est régulé transcriptionnellement par plusieurs facteurs de transcription intermédiaires et constitue également une cible thérapeutique médicamenteuse. Le seul produit nucléaire est la SFN et il existe six protéines membranaires, toutes contrôlées par quelques facteurs de transcription intermédiaires. Parmi les gènes sous-exprimés dans BL2, les protéines les plus interconnectées sont *NDRG2* et *COBL*, toutes deux cytoplasmiques, BAMBI et MBOAT1, toutes deux situées sur la membrane cellulaire, et EHZF qui est située dans le noyau.

Douze des vingt produits génétiques sur-exprimés dans LAR sont directement régulés par le récepteur d'androgène, c'est-à-dire dans la signature LAR elle-même. Il s'agit de : la protéine extracellulaire Amphiregulin ; quatre protéines membranaires (alpha-ENaC, CD166, TSPAN1, STEAP4) ; sept protéines cytoplasmiques (ALOX15B, FLJ20184, KIAA1324, ATAD4, CRAT, FASN, CYP19).

Trente et une des trente-cinq protéines codées par les gènes sous-exprimés dans LAR sont directement connectées sans aucun intermédiaire, les facteurs de transcription LBP9, c-Myc et CXXC1 contrôlant la plupart des gènes signatures.

Aucune des protéines codées par les gènes du sous-type sur-exprimés dans M n'est directement connectée à une autre, mais elles sont toutes connectées si l'on ajoute un intermédiaire, SOX6 et ID4 (nucléaire), MDFI et Desmocollin 3 (cytoplasmique), et la glycoprotéine transmembranaire BAMBI étant les plaques tournantes du réseau les plus interconnectées. Le réseau impliquant les protéines codées par les gènes sous-exprimés dans M n'est pas facilement interprétable.

Comme pour le sous-type IM, les deux seuls gènes sur-exprimés codent pour deux facteurs de transcription, SPI- B et Aiolos, qui sont parmi les plus interconnectés au sein du réseau lorsqu'un intermédiaire est inclus. La majorité des intermédiaires convergent vers IP-10, MIG ou I-TAC, trois extracellulaires, ou vers CD38, une glycoprotéine transmembranaire de type II, toutes surexprimées dans le sous-type IM. Un autre nœud central du réseau IM est le granzyme B, une protéase sécrétée par les cellules tueuses naturelles et les lymphocytes T cytotoxiques. Les gènes sous-exprimés dans IM sont *ID4*, *MDFI*, *KRT81*. Seules les protéines codées par les deux premiers sont connectées, via soit le facteur de transcription p53, soit la déméthylase JMJD2A.

Enfin, le gène non codant *MEG3* est l'élément central du réseau résultant des gènes sur-exprimés dans MSL et est lié à l'IGF-I et l'IGF-II via l'inhibition de plusieurs microARN (miR-218-3p, miR-96-5p, miR-19-3p, miR-493-5p, miR- 665-3p, miR-129-5p, miR-18a-5p, miR-129-3p, miR-181a-5p) ciblant les deux facteurs de croissance extracellulaires.

D'autre part, les éléments de contrôle du cycle cellulaire tels que CDK1 et CDKN2A jouent un rôle central dans les gènes sous-exprimés dans MSL.

3. Identification des cibles médicamenteuses

Les gènes différentiellement exprimés dans chaque sous-type ont ensuite été analysés avec Metacore, afin de rechercher toute cible médicamenteuse. La cible médicamenteuse la plus surexprimée de BL1 est la Matrilysine, codée par *MMP7* et ciblée par plusieurs inhibiteurs thérapeutiques, tels que : Batimastat ; Marimastat, et Rebimastat.

Quant au sous-groupe BL2, la principale interaction thérapeutique inhibitrice entre le médicament et la cible concerne la Stromelysin-1 codée par *MMP3* et ciblée par la Doxycycline, et le Tanomastat.

D'autre part, l'une des cibles médicamenteuses de LAR les plus récurrentes et potentiellement importantes est le récepteur d'androgène codé par *AR* et inhibé par Bicalutamide, Diethylstilbestrol, Drospirenone, Finasteride, Flutamide, Metandienone, RU58841, Silibinin, Zanoterone. Le second est le CYP19 codé par le *CYP19A1* et ciblé par plusieurs inhibiteurs de l'aromatase, comme l'Aminoglutéthimide, l'Anastrozole, l'Exémestane, le Létrozole et la Testolactone. Ensuite, GGT1, ciblé par l'Acivicine et par l'Oxiglutathione ; GGTF-I-beta, codé par *PGGT1B* et ciblé par le L-778,123 ; ALDR, codé par *AKR1B1* et ciblé par le Tolrestat ; alpha-ENaC, codé par *SCNN1A* et ciblé par l'Amiloride. Quant aux sous-types M, IM et MSL, aucune interaction thérapeutique spécifique médicament-cible n'a été repérée. En revanche, plusieurs interactions médicament-cible secondaires d'inhibition pour les gènes sur-exprimés, prédites sur la base de similitudes dans les structures, ont été trouvées. Le récepteur Ephrin-B 3, codé par *EPHB3* et sur-exprimés dans le sous-groupe M, est une cible prédite de plusieurs médicaments inhibiteurs tels que CC-223, Dovitinib, Nazartinib, Nilotinib et Ponatinib ; CD38 dans le sous-groupe IM est une cible prédite de Ca ('2+), du propionate de fluticasone et de la quercétine ; SR-B codé par *SCARB1* et surexprimé dans le groupe LAR est une cible prédite de la bêta-cyclodextrine, de l'acide docosahexaénoïque et de l'ITX-5061.

Réciproquement, aucune interaction thérapeutique activatrice entre le médicament et la cible pour les gènes sous-exprimés n'a été repérée dans les six sous-groupes de CSTN.

4. Prédiction des sous-types CSTN

Il est très important dans toute étude biologique d'identifier les informations les plus significatives à partir de données biologiques complexes. On sait que les changements physiologiques et pathologiques du phénotype tumoral et sa sensibilité à des traitements spécifiques sont généralement déterminés par des interactions moléculaires. Nous avons donc évalué si les signatures génétiques spécifiques aux sous-types décrites précédemment étaient également capables de prédire les classes d'échantillons.

En conséquence, sept modèles de prédiction différents ont été appliqués à l'ensemble de données GEO-TN, à partir des listes de gènes sous et sur-exprimés obtenues précédemment. Pour les deux listes, une validation croisée 10 fois a été utilisée car elle donne aux modèles la possibilité de s'entraîner sur de multiples fractionnements entraînement-test, ce qui donne une meilleure indication de la performance des modèles sur des données non vues. La variable à prédire était le "sous-type CSTN" et les caractéristiques explicatives étaient les gènes régulés à la hausse ou à la baisse.

Le modèle MP suivi du modèle SVM se distinguent par les meilleurs scores; en revanche, LR et DT semblent être les moins performants parmi tous les modèles, pour les deux listes. Par conséquent, le modèle MP a été choisi pour une utilisation ultérieure en validation externe sur les ensembles de données TCGA-TN et Italian-TN. Par conséquent, afin de savoir si l'un des gènes avait un faible poids prédictif selon le meilleur modèle prédictif (MP), sept méthodes différentes de sélection d'attributs ont été élaborées, qui ont voté pour des gènes de classement légèrement différent. Les gènes jugés non importants par la majorité des algorithmes ont été supprimés. Après le raffinement des deux listes de gènes, une comparaison ROC par sous-groupe a été effectuée, avant et après la sélection des attributs, pour évaluer si l'élimination des gènes susmentionnée modifiait les performances de prédiction du même modèle. Les prédictions ont d'abord été mesurées sur l'ensemble d'entraînement avec l'option de validation croisée 10-fois, puis sur les deux ensembles de validation. Des scores ROC très stables ont été obtenus, même après la suppression des gènes les moins importants. En ce qui concerne les gènes sur-exprimés, malgré la suppression de 17 gènes, le

score ROC s'est amélioré dans les ensembles de données d'entraînement et de validation, dans la majorité des cas.

Conclusion

Notre étude a tiré pleinement parti des ensembles de données CSTN disponibles pour stratifier les échantillons et les gènes en sous-types distincts, en fonction des profils d'expression génique. Le développement d'une approche d'exploration de données pour acquérir une grande quantité d'informations à partir de plusieurs ensembles de données, nous a permis d'identifier un nombre bien déterminé de gènes qui peuvent aider à la reconnaissance des sous-types de CSTN. Ce petit nombre de gènes peut être testé en clinique sans avoir recours à des approches transcriptomiques complètes. La plupart des gènes de signature ont déjà été associés au cancer du sein et ont le potentiel pour devenir de nouveaux marqueurs de diagnostic et/ou des cibles thérapeutiques pour des sous-classes spécifiques de cancer du sein triple négatif.

Implications potentielles

Dans l'ensemble, nos signatures génétiques affinées pour chaque sous-type de CSTN peuvent fournir un outil clinique simple, accessible à la plupart des services de pathologie, qui pourrait contribuer à explorer l'hétérogénéité du CSTN et à identifier le traitement approprié pour chaque patient sur la base des cibles médicamenteuses spécifiques au sous-type. De nouveaux essais cliniques prenant en compte le portrait moléculaire de la tumeur sont en fait en cours de développement, pour le cancer du sein transgénique également.

Chapitre 2 : L'indice de prolifération Ki-67 pour stratifier davantage les sous-types moléculaires du cancer du sein invasif : Étude de cohorte comparative nord-africaine avec validation externe TCGA-BRCA et METABRIC.

Introduction:

Au niveau mondial, le CS est le cancer le plus fréquent chez les femmes, avec environ 2,2 million de nouveaux cas diagnostiqués en 2020 (11,7 % de tous les cancers, pour les deux sexes, tout âge confondu). Son taux d'incidence varie considérablement entre les régions du monde, avec un pic en Asie, suivi par les parties centrale, orientale et occidentale de l'Europe, l'Amérique du Nord et latine et l'Afrique (The Global Cancer Observatory. Globocan 2020). La mortalité au sein du continent africain varie de 5090 événements en Afrique australe, région la moins concernée, à 25626 en Afrique occidentale, ce qui en fait la région d'Afrique la plus concernée par ce type de cancer.
En 2020, 11747 nouveaux cas de CS ont été enregistrés au Maroc, représentant 19,8% de l'ensemble des cancers chez la femme et le premier cancer diagnostiqué. C'est le premier également en termes de mortalité (3695 décès estimés) et de prévalence (31420 cas pour une prévalence de 5 ans) (La Fondation Lalla Salma Prévention et traitement du cancer. GUIDE DE DÉTECTION des CANCERS PRÉCOCE du sein et du col de l'utérus. Edition 2011. 2011).
Le registre du cancer de la région du Grand Casablanca (2016-2020), selon le dernier rapport élaboré par la Direction de l'épidémiologie et de la lutte contre les maladies du ministère de la Santé, estime la fréquence du CS à 35,8%, avec un pic enregistré entre 55 et 59 ans (RCRGC. CANCER REGISTER).
Il est donc clair que le CS est le premier cancer féminin, ce qui en fait un problème de santé publique au Maroc, ainsi que dans le monde.
On estime actuellement qu'une femme sur 9 développera un CS au cours de sa vie et qu'une sur 27 en mourra, soulignant l'importance de cette maladie en termes de santé publique. Il est à noter que les hommes peuvent également développer un CS. Ces cas sont toutefois rares, puisqu'ils ne représentent que 1% des carcinomes mammaires (Yalaza, İnan, et Bozer 2016).
Actuellement, la mammographie est le meilleur moyen de détecter le CS à un stade précoce. En moyenne, la tumeur peut être détectée 1,7 an avant qu'une femme ne ressente une grosseur. Aux premiers stades d'une tumeur localisée, les chances de survie à 5 ans sont de 95%. Ces chances

diminuent aux stades tardifs: elles sont inférieures à 50% lorsque la tumeur s'est disséminée dans les ganglions lymphatiques et inférieures à 20% lorsqu'elle s'est disséminée dans des organes distants. La détection précoce des lésions cancéreuses, la chirurgie, avec ablation sélective de la tumeur, et les différentes thérapies (chimiothérapie, radiothérapie, hormonothérapie ou autres thérapies ciblées) ont contribué à réduire considérablement la mortalité due au CS. Cependant, malgré ces progrès, certains types de cancer agressifs et métastatiques sont difficiles à traiter et restent encore incurables.

Il est donc très important de définir la panoplie complète des biomarqueurs influençant la survie des patients atteints de CS. Les méthodes statistiques peuvent aider à sélectionner la meilleure combinaison de biomarqueurs à utiliser afin de prédire la survie et le pronostic (Vickers et Cronin 2010). Plusieurs études ont été réalisées précédemment en utilisant des techniques statistiques conventionnelles qui sont limitées en termes de génération de visualisations claires et créatives des résultats obtenus par l'analyse de ces facteurs (Rajula et al. 2020).

Les limites de ces techniques statistiques ont peut-être permis aux cliniciens d'utiliser d'autres techniques d'apprentissage automatique plus robustes et plus profondes, telles que les arbres de décision (DT), Naive Bayes (NB), le modèle linéaire généralisé (GLM), Random Forest (RF), Fast Large Margin (FLM), Deep Learning (DL), Logistic Regression (LR), Gradient Boosted Trees (GBT), Support Vector Machine (SVM), K-nearest neighbours (KNN) et le Multilayer Perceptron (MP) (Rajula et al. 2020 ; Dubey, Gupta et Jain 2015).

Nous avons évalué les mêmes techniques de prédiction mentionnées ci-dessus sur un ensemble de données de patients marocains avec 1266 dossiers de CS dans cette étude rétrospective de 5 ans de suivi.

Dans ce travail, nous avons appliqué ces techniques d'apprentissage automatique sur une grande cohorte de patients pour explorer si la classification moléculaire acceptée comme référence pour la détermination des sous-types de CS peut être affinée par des méthodes de partitionnement statistique. Ceci est particulièrement utile dans les pays à revenu faible et moyen comme le Maroc, où les laboratoires n'ont pas nécessairement un large accès aux nouvelles méthodes moléculaires qui

s'avèrent très coûteuses comme les tests de profilage d'expression génique. En outre, notre travail consiste également à savoir si ces résultats peuvent être reproductibles avec un certain degré de pertinence.

Cependant, il est nécessaire de souligner que des études sur le cancer du sein utilisant des techniques d'apprentissage automatique ont déjà été développées auparavant par plusieurs auteurs, mais les facteurs étudiés varient d'une étude à l'autre, en fonction de la population cible, de sa géolocalisation, de son mode de vie, des bases de données disponibles et même de l'objectif de l'étude.

Nous avons donc conclu qu'il est nécessaire de développer un modèle pour le contexte africain, modèle qui n'a jamais été étudié auparavant et plus précisément au Maroc, afin d'étudier les variables pouvant régir le taux de survie des patientes marocaines atteintes de cancer du sein à travers les indicateurs histo-prognostiques habituellement analysés en routine dans tous les laboratoires d'anatomopathologie. Nous nous intéresserons également par la suite à la technique de sélection des variables les plus pertinentes en utilisant ces mêmes techniques d'apprentissage automatique dans le domaine médical.

Matériels et méthodes :

- Conception et contexte de l'étude :

Il s'agit d'une étude de cohorte rétrospective comparative incluant des patients marocains atteints de cancer du sein avec un suivi de 5 ans, les bases de données TCGA-BRCA et METABRIC avec un suivi de 13 et 30 ans, respectivement. Tous les carcinomes mammaires invasifs enregistrés dans les bases de données mentionnées ont été inclus. En revanche, les tumeurs bénignes ou de malignité incertaine, les récidives tumorales, les cancers du sein chez l'homme et les patients dont le statut immuno-histochimique était incomplet ou équivoque ont tous été écartés.

- Bases de donnés collectées:

a) TCGA-BRCA

Il a été récupéré à partir de "https://portal.gdc.cancer.gov/" et était initialement composé de 963 enregistrements de CS invasifs. L'ensemble de données contient les caractéristiques suivantes : ESR1, PGR, ERBB2 et MKi-67 (Z-scores d'expression des gènes), invasion ganglionnaire, statut de la ménopause, stade de la tumeur, fraction du génome altéré, stade métastatique, type de cancer, poids initial de l'échantillon, nombre de mutations, détection des micro-métastases, ethnie, type histologique, race, période de survie globale (sur 13 ans), stade de diagnostic. Après le filtrage des valeurs manquantes, le nombre final d'enregistrements CS restants était de 624. 24 patients ont été enregistrés comme étant hispaniques ou latinos ; 458 comme étant non hispaniques ou latinos et le statut était manquant pour 143 patients. Comme catégorie de race: 428, 70, 38,1 ont été enregistrés comme blancs, noirs, ou afro-américains, asiatiques respectivement et les données sur la race étaient manquantes pour les 88 patients restants.

b) METABRIC

Il contient 1885 enregistrements CS et a été initialement récupéré à partir de: https://www.cbioportal.org/. Elle contient les caractéristiques histopronostiques suivantes : MKi-67 (Z-score d'expression des gènes) ; ER/PgR/HER2 (Sur-exprimé / Sous-exprimé) ; Age au diagnostic; Type de cancer ; Cellularité ; Chimiothérapie ; Grade histologique du néoplasme ; Statut HER2 mesuré par SNP6 ; Sous-type histologique ; Hormonothérapie ; État ménopausique présumé ; latéralité de la tumeur primaire ; indice pronostique de Nottingham ; radiothérapie ; taille de la tumeur ; statut vital du patient ; ganglions lymphatiques examinés positifs ; nombre de mutations; durée de survie globale (sur 30 ans de suivi); statut de survie (censuré/mort). Il n'y avait aucune valeur manquante dans cet ensemble de données. Aucune caractéristique sociale ou démographique n'était présente dans cet ensemble de données.

c) Base de données marocaine :

Les données générales et cliniques de tous les carcinomes invasifs enregistrés du 1er janvier 2013 au 30 mars 2018 au service d’anatomo-pathologie de l'hôpital universitaire Ibn Rochd de Casablanca ont été récupérées, ce qui a conduit à 1266 patientes marocaines avec un CS invasif et 165 d'entre eux ont été suivies au Centre national de traitement des cancers du Roi Mohammed VI,

où leurs données de survie après 5 ans ont été collectées (à partir de leurs dossiers médicaux correspondants des registres nationaux). Ce centre est considéré comme le plus grand hôpital public du Maroc avec le plus grand registre du cancer. Le patient était considéré comme mort si le décès était confirmé à une date définie. Ou confirmé vivant, par son médecin traitant à la dernière date de suivi. Aucune information démographique/sociale n'a été trouvée dans le registre national. Les caractéristiques histopronostiques collectées sont : l'âge au moment du diagnostic, la taille de la tumeur (TS), l'infiltration des ganglions lymphatiques (NI), le grade SBR, les récepteurs aux œstrogènes (ER), les récepteurs à la progestérone (PgR), l'indice de prolifération Ki-67 et le statut des récepteurs HER2 par immunohistochimie, l'absence/la présence d'emboles vasculaires (VE), le type histologique, la classification TNM, la première et la dernière date de suivi, l'état à la dernière date de suivi. Ces ensembles de données ont été spécifiquement choisis en raison du grand nombre d'échantillons de CS qu'ils contiennent. Le jeu de données marocain a été utilisé principalement comme un jeu de données interne sur lequel l'analyse se concentre. Les ensembles de données METABRIC et TCGA-BRCA ont été utilisés pour servir d'ensembles de validation externes accessibles publiquement.

- Classification moléculaire du CS :

Pour tous les ensembles de données, les variables Ki-67, ER, PgR et HER2 ont été extraites pour une analyse plus approfondie. Par la suite, le CS a été systématiquement classé en cinq sous-groupes intrinsèques, comme suit :

- LuminalA (LumA) : ER+ et/ou PgR+ ; HER2- ; Ki-67 faible
- LuminalB HER2+ (LumB HER2+) : ER+ et/ou PgR+ ; HER2+ ; Ki-67 élevé
- LuminalB HER2- (LumB HER2-) : ER+ et/ou PgR+ ; HER2- ; Ki-67 élevé
- HER2 pur : ER- et PgR- ; HER2+ ; indépendamment du Ki-67
- Triple négatif (TN) : ER- et PgR- ; HER2- ; indépendamment du Ki-67.

- Pré-traitement informatique

Nous avons d'abord évalué une étape de nettoyage, consistant à normaliser toutes les variables numériques pour améliorer les performances des algorithmes qui utilisent des entrées pondérées ou des mesures de distance et les rendre comparables.

En ce qui concerne les ensembles de données METABRIC et TCGA-BRCA qui contiennent des variables z-score, les variables positives ont été considérées comme ayant une forte expression de MKi-67, ER, PgR et HER2; en revanche, les variables ayant des valeurs z-score négatives ont été considérées comme sous-exprimées. Toutes les données manquantes ont été exclues, et les lignes correspondantes ont été supprimées des ensembles de données. D'où le filtrage de tout patient présentant au moins une valeur manquante pour l'une des quatre variables suivantes: ER, PgR, Ki-67 et HER2 qui sont les principales caractéristiques qui nous intéressent.

- Partitionnement statistique:

-Clustering par Estimation-Maximisation (EM)

Principalement évalué par le logiciel "STATISTICA, v10". Il consiste à classer automatiquement chaque patiente dans le cluster auquel elle a le plus de chance d'appartenir (probabilité la plus élevée). Initialement (développé par Dempster et al.1977) consiste à trouver les solutions de classification qui maximiseront la probabilité globale des données.

-Le PAM clustering

Il définit des k-objets représentatifs des classes, appelés médoïdes, situés au centre des classes.

-K-means Clustering

Après avoir initialisé ses centroïdes en prenant des données aléatoires dans le jeu de données, K-means alterne plusieurs fois ces deux étapes pour optimiser les centroïdes et leurs groupes:

- regrouper chaque objet autour du centroïde le plus proche.
- Remplacer chaque centroïde en fonction de la moyenne des descripteurs de son groupe.
- Après quelques itérations, l'algorithme trouve une division stable du jeu de données avec K groupes.

-Le clustering hiérarchique

Il commence par considérer que chaque point est un cluster à lui seul. Ensuite, il trouve les deux clusters les plus proches, et les agrège en un seul cluster. Cette étape est répétée jusqu'à ce que tous les points appartiennent à un seul cluster, constitué de l'agglomération de tous les clusters initiaux.

- Détermination du nombre optimal de clusters:

-Indices de qualité :

Évalués par l'extension NbClust sous le logiciel R. Cette bibliothèque permet de déterminer le bon nombre de classes dans les deux ensembles de données. Trente indices de validation proposés, afin de déterminer de la manière la plus impartiale et objective le nombre optimal de clusters.

-Mesures de validation :

Les mesures de validation ont été évaluées par le "package clValid" en R, qui aide à sélectionner simultanément plusieurs algorithmes de clustering, des métriques de validation, et le nombre de clusters en un seul appel de fonction, afin de déterminer la méthode la plus appropriée et le nombre optimal de clusters.

-Les mesures de stabilité des clusters comprennent : Le chiffre de mérite (FOM) ; La distance moyenne entre les médoïdes (ADM) ; La proportion moyenne de non-chevauchement (APN) ; La distance moyenne (AD).

-Mesures de validation interne: la compacité, la séparation ; avec la connectivité, constituent les trois mesures internes les plus importantes.

- Sélection des variables importantes :

-Algorithme VIMP

Se base principalement sur le modèle RF pour tester la prédiction et classe les variables les plus importantes en fonction de leur impact sur la capacité de prédiction de la forêt. Une valeur VIMP égale ou proche de zéro indique que la variable ne contribue pas à la précision prédictive ; en revanche, des valeurs négatives indiquent que la précision prédictive s'améliore lorsque la variable est mal spécifiée. Par conséquent, nous avons utilisé la fonction "gg_vimp" associée à l'extension "ggRandomForests" du logiciel R, que nous avons utilisée pour extraire essentiellement des mesures VIMP.

-Profondeur minimale

Elle suppose que les variables ayant un impact élevé sur la prédiction sont celles qui divisent le plus fréquemment les nœuds les plus proches du nœud racine, donc les plus grands échantillons de la population. Les niveaux de nœuds dans chaque arbre sont numérotés en fonction de leur distance relative à la racine de l'arbre (qui est indiquée par le niveau 0). L'hypothèse est que les plus petites valeurs de profondeur minimale indiquent que la variable a un grand impact sur la prédiction de la forêt. Cette dernière a été élaborée avec l'extension "randomForestSRC".

Section 1 : principaux résultats sur la base de données marocaine :

-La distinction d'une nouvelle subdivision intrinsèque à la classification moléculaire développée en routine dans la clinique, notée Cluster1 et Cluster2.

-Le Cluster1 inclut modérément tous les patients CS avec un Ki-67 bas et un profil tumoral moins agressif et prolifératif que ceux appartenant au Cluster2.

La distinction entre deux patients atteints de cancer du sein appartenant au même sous-groupe moléculaire semble donc essentielle car il existe un certain degré d'hétérogénéité au sein d'un même sous-groupe moléculaire, qui peut être lié à un pronostic différent.

Parmi toutes les caractéristiques histopronostiques, les récepteurs hormonaux soutiennent l'appartenance au cluster 1, tandis que le Ki-67 ainsi que le HER2, l'invasion des ganglions lymphatiques, la présence d'emboles vasculaires et le grade SBR soutiennent l'appartenance au cluster 2, car ils confèrent aux tumeurs un pouvoir prolifératif et agressif qui converge avec le profil trouvé dans les tumeurs du cluster 2.

-La survie globale moyenne des patientes appartenant au Cluster1 est beaucoup plus favorable par rapport à celle du Cluster2.

-La survie globale des patients appartenant au Cluster1 de LuminalB HER2- est beaucoup plus proche de celle des patients appartenant au Cluster1 de luminal B HER2+ que de celle des patients du Cluster2 de luminal B HER2-. Il en va de même pour les autres sous-groupes clustérisés, d'où la

nécessité de diviser chaque sous- groupe en deux subdivisions Cluster1 et Cluster2 qui affinent leur pronostic et leur prédiction de survie.

-Nous avons évalué plusieurs modèles de prédiction de survie et les avons évalués avec différentes métriques, en appliquant l'option de validation croisée. Les résultats montrent que la survie peut être prédite efficacement par tous les modèles qui ont présenté de bons scores de précision, mais surtout par le modèle Random Forest qui les a précédés et a prédit avec succès la survie pour chaque cluster dans chaque sous-groupe moléculaire.

Section 2 : principaux résultats sur la base d'une étude comparative avec la validation externe de TCGA-BRCA et METABRIC :

Dans cette section, l'objectif principal est de valider en externe la même " partition à 2 clusters " au sein d'autres bases de données indépendantes de CS, afin de confirmer / infirmer nos résultats préliminaires obtenus précédemment sur la population marocaine. Pour ce faire, il sera nécessaire d'explorer le nombre optimal de clusters cachés dans de nouvelles bases de données avec différents nombres de cas et différentes caractéristiques histopathologiques, transcriptomiques et/ou génétiques, mais de la manière la plus impartiale et objective possible.

C'est la raison pour laquelle des techniques d'apprentissage automatique non supervisées sont utilisées. Ces dernières ont l'avantage de trouver par elles-mêmes le meilleur nombre de clusters afin de ne pas l'emporter sur le nombre de clusters (k = 2) obtenu dans la base de données interne marocaine.

Dans ce but, le jeu de données TCGA a été utilisé, qui est librement accessible et récupéré sur le CBioPortal For Cancer Genomics. Il contient une grande richesse d'informations : une combinaison de données histopathologiques, RNA-seq et de survie pour 625 patients.

-La partition des deux ensembles de données externes, TCGA et METABRIC, a montré qu'ils devraient de manière optimale être subdivisés en deux autres classes internes strictement associées à l'indice de prolifération Ki-67, dénommées Cluster1 et Cluster2. Et ce, selon plusieurs méthodes de calcul évaluées par plusieurs indices de mesure.

-Le Cluster1 est essentiellement caractérisé par une faible moyenne de Ki-67 par rapport à celle du Cluster2, qu'elle soit mesurée par IHC (% de coloration) ou par des méthodes d'expression génique (Z score). La moyenne du Cluster1 est toujours beaucoup plus faible que celle du Cluster2.
Cette différence de moyenne s'est avérée être significativement différente, également, au sein de chaque sous-groupe moléculaire.
Les sous-groupes moléculaires cliniquement connus et de mauvais pronostic, tels que HER2 pur et TN, sont davantage surreprésentés dans le Cluster2, contrairement aux sous-groupes luminaux qui appartiennent principalement au Cluster1.
-L'établissement de l'analyse des taux de survie globale par l'évaluation des courbes de Kaplan-Meier a permis de faire une distinction dans la survie des patients, en fonction de leur appartenance aux clusters.
-La survie moyenne des patients du Cluster 1 reste toujours significativement plus favorable que celle des patients du Cluster 2, confirmant ainsi nos premiers résultats principaux sur l'ensemble des données marocaines.
-Dans l'ensemble, ce nouveau raffinement de la classification moléculaire est important car il est voté par deux algorithmes différents dans la prédiction de la survie, par rapport à plusieurs autres critères pronostiques, génétiques et moléculaires. Les trois analyses (marocaine, TCGA et METABRIC) convergent sur ce point.
-Nous avons pu réduire toute la panoplie des facteurs histopathologiques à quelques prédicteurs, qui peuvent prédire l'appartenance à un groupe et la survie avec une meilleure précision.
-Dans les pays aux moyens et ressources limités qui ne peuvent pas utiliser les signatures génomiques, il a été démontré que la classification en clusters peut nous donner une idée du pronostic des patients, et ce, par l'utilisation de quelques facteurs prédictifs pris en compte en routine dans tous les laboratoires de pathologie.

Discussion

Cette étude explore la possibilité de partitionner les sous-groupes moléculaires du cancer du sein afin de mieux définir la survie des patients. L'approche de partitionnement a été appliquée à partir de 1128 dossiers marocains de cancer du col de l'utérus, puis testée sur deux ensembles de données externes indépendants, également utilisés pour valider la signification clinique des nouvelles subdivisions, en termes de prédiction de survie.

Nous avons constaté que la classification moléculaire du CS établie de manière routinière pouvait être affinée en utilisant uniquement les variables Ki-67, ER, PgR et HER2. Chaque sous-type peut être divisé en deux groupes distincts avec une distribution de Ki-67 et des résultats de survie significativement différents. Les tumeurs appartenant au cluster à faible activité mitotique (C1) sont surreprésentées dans le sous-type luminal, tandis que les sous-types HER2 et TN sont enrichis dans les tumeurs appartenant au cluster à forte activité mitotique (C2). Ce cloisonnement est également associé à la survie globale (OS) et est aussi important, voire plus, que la taille de la tumeur pour prédire le résultat. En effet, Marwah et al. (Marwah et al. 2018) ont révélé qu'un Ki-67 plus élevé était retrouvé avec une taille supérieure à 5cm alors que les tumeurs inférieures à 2cm présentaient un taux plus faible. Ce résultat a également été confirmé par Querzoli & al(Querzoli et al. 1996). De même, plusieurs études ont montré une corrélation positive entre le Ki-67 et le grade histologique de la tumeur. Cependant, nous pouvons souligner que notre analyse sur le jeu de données marocain n'a pas classé le grade histologique, contrairement à l'appartenance à un groupe et à la taille de la tumeur, parmi les variables les plus importantes capables de prédire la survie.

Un autre résultat intéressant est la présence d'échantillons C1 au sein des tumeurs TN. Le sous-type histologique pourrait en partie expliquer ce résultat, le carcinome adénoïde kystique étant un exemple typique. Bien qu'il n'exprime pas de récepteurs hormonaux et ne surexprime pas HER2, il présente un faible indice de prolifération (Bouzubar et al. 1989) (D.-Y. Wang et al. 2019). Le Ki-67 est également signalé comme étant plus élevé dans le CSTN sans type particulier par rapport au CSTN d'autres sous-types histologiques (Tan et al. 2004). Il a été suggéré par certains auteurs

comme un marqueur pronostique et prédictif du CSTN (X. Zhu et al. 2020). L'étude de Keam a proposé un seuil de 10 % pour le Ki-67 afin de définir deux sous-groupes pronostiques différents : le premier avec un Ki-67 élevé, qui malgré une meilleure réponse à la chimiothérapie était plus agressif, et le second avec un Ki-67 faible, qui montrait une moindre agressivité mais aussi une moindre réponse à la chimiothérapie (Keam et al. 2011 ; Bartlett et al. 2016)). Nos résultats confirment ces conclusions, avec des CSTNs partitionnés en 2 autres subdivisions : C1 et C2, avec des Ki-67 moyens de 16,4±13% et 73±15,4%, respectivement.

D'autre part, Bartlett & al. ont récemment rapporté que, malgré une faible concordance au niveau de la tumeur unique, l'hétérogénéité au sein des tumeurs ER+ en termes de pronostic est détectable et confirmée par différents tests multiparamètres (Bartlett et al. 2016). De plus, Aleskandarany & al. ont confirmé qu'au sein du sous-groupe Luminal B / HER2-, le groupe avec un indice de prolifération élevé avait une évolution et un pronostic plus mauvais que le groupe avec un Ki-67 bas. Cela confirme nos résultats, avec certains cas au sein du sous-groupe moléculaire Luminal B / HER2- regroupés en C2 et présentant une survie plus faible dans les trois ensembles de données analysés. Par conséquent, les sous-groupes moléculaires du cancer du sein doivent être considérés comme un spectre de maladies (Bartlett et al. 2016).

En termes de survie, nous avons montré que le partitionnement des sous-types permet d'affiner le pronostic. Kyung Lee et al. ont démontré que la combinaison de p53 et Ki-67 a le meilleur pouvoir prédictif, notamment pour la survie globale à long terme dans le sous-groupe Luminal A (Lee et al. 2015). De manière intéressante, dans notre étude, nous avons pu mettre en évidence que non seulement le luminal A mais aussi les autres sous-types moléculaires, le luminal B en particulier, pourraient bénéficier d'un affinement supplémentaire basé sur le Ki-67. Il faut noter que les tumeurs luminales représenteraient un groupe hétérogène sur le plan génotypique avec certaines tumeurs présentant une instabilité chromosomique avec aneuploïdie et d'autres sans instabilité génomique et diploïdes (Yanagawa et al. 2012). D'autre part, aucune différence pronostique significative n'a pu être établie pour les tumeurs HER2 et TN, puisque pour notre cohorte marocaine, les informations de survie n'étaient complètes que pour un petit sous-ensemble de patients. Nous avons donc répété

les analyses sur TCGA-BRCA et METABRIC, ce qui nous a permis non seulement d'étendre et de valider nos résultats de clustering à un contexte plus large, mais aussi de confirmer que les subdivisions intrinsèques proposées au sein des sous-groupes moléculaires ont un impact pronostique clinique. TCGA-BRCA et METABRIC contiennent une très large panoplie de caractéristiques pronostiques et couvrent même les données génomiques de chaque patient, ils pourraient donc être exploités davantage pour identifier des biomarqueurs supplémentaires traduisibles en clinique.

Conclusion

Les efforts déployés ces dix dernières années pour mieux comprendre l'histopathologie du cancer du sein invasif ont abouti à une conclusion importante : ce dernier recouvre des cancers différents. Certains d'entre eux ont un "portrait moléculaire" complètement défini, qui peut être identifié par des méthodes génomiques. Cependant, si le chemin vers l'intégration dans la pratique clinique d'un portrait moléculaire et morphologique est encore long, il devra néanmoins se faire, afin d'offrir un diagnostic plus abouti aux patients et aux médecins. Le profilage génomique peut être très long et coûteux, c'est pourquoi nous avons essayé de définir une nouvelle superposition entre l'analyse histopathologique et bioinformatique et de définir un nouveau raffinement de la classification moléculaire du CS qui s'est également avéré être l'une des variables les plus importantes en termes de prédiction de survie.

BIBLIOGRAPHY:

Abdul Aziz, Ahmad Aizat, Md Salzihan Md Salleh, Maya Mazuwin Yahya, Andee Dzulkarnaen Zakaria, and Ravindran Ankathil. 2021. "Genetic Association of CYP1B1 4326 C>G Polymorphism with Disease-Free Survival in TNBC Patients Undergoing TAC Chemotherapy Regimen." *Asian Pacific Journal of Cancer Prevention: APJCP* 22 (4): 1319–24.

Ahn, Hyo Jung, Soo Jin Jung, Tae Hyun Kim, Min Kyung Oh, and Hye-Kyoung Yoon. 2015. "Differences in Clinical Outcomes between Luminal A and B Type Breast Cancers according to the St. Gallen Consensus 2013." *Journal of Breast Cancer* 18 (2): 149–59.

Alba, Emilio, Ana Lluch, Nuria Ribelles, Antonio Anton-Torres, Pedro Sanchez-Rovira, Joan Albanell, Lourdes Calvo, et al. 2016. "High Proliferation Predicts Pathological Complete Response to Neoadjuvant Chemotherapy in Early Breast Cancer." *The Oncologist* 21 (2): 150–55.

Allison, Kimberly H., M. Elizabeth H. Hammond, Mitchell Dowsett, Shannon E. McKernin, Lisa A. Carey, Patrick L. Fitzgibbons, Daniel F. Hayes, et al. 2020. "Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 38 (12): 1346–66.

Alsafadi, S., V. Scott, P. Pautier, A. Goubar, V. Lazar, P. Dessen, L. Lacroix, et al. 2011. "P5-01-07: Identification of SORBS2 as a Candidate Marker To Predict Metastatic Relapse in Breast Cancer." *Cancer Research*. https://doi.org/10.1158/0008-5472.sabcs11-p5-01-07.

Alvarez-Baron, Claudia P., Philip Jonsson, Christoforos Thomas, Stuart E. Dryer, and Cecilia Williams. 2011. "The Two-Pore Domain Potassium Channel KCNK5: Induction by Estrogen Receptor Alpha and Role in Proliferation of Breast Cancer Cells." *Molecular Endocrinology* 25 (8): 1326–36.

Aranda-Gutierrez, Alejandro, and Hector M. Diaz-Perez. 2020. "Histology, Mammary Glands." In *StatPearls*. Treasure Island (FL): StatPearls Publishing.

Asztalos, Szilard, Thao N. Pham, Peter H. Gann, Meghan K. Hayes, Ryan Deaton, Elizabeth L. Wiley, Rajyasree Emmadi, et al. 2015. "High Incidence of Triple Negative Breast Cancers Following Pregnancy and an Associated Gene Expression Signature." *SpringerPlus* 4 (November): 710.

Badillo, Solveig, Balazs Banfai, Fabian Birzele, Iakov I. Davydov, Lucy Hutchinson, Tony Kam-Thong, Juliane Siebourg-Polster, Bernhard Steiert, and Jitao David Zhang. 2020. "An Introduction to Machine Learning." *Clinical Pharmacology and Therapeutics* 107 (4): 871–85.

Bao, Y. I., Antao Wang, and Juanfen Mo. 2016. "S100A8/A9 Is Associated with Estrogen Receptor Loss in Breast Cancer." *Oncology Letters* 11 (3): 1936–42.

Bartlett, John M. S., Jane Bayani, Andrea Marshall, Janet A. Dunn, Amy Campbell, Carrie Cunningham, Monika S. Sobol, et al. 2016. "Comparing Breast Cancer Multiparameter Tests in the OPTIMA Prelim Trial: No Test Is More Equal Than the Others." *Journal of the National Cancer Institute* 108 (9). https://doi.org/10.1093/jnci/djw050.

Beltrán-Anaya, Fredy Omar, Sandra Romero-Córdoba, Rosa Rebollar-Vega, Oscar Arrieta, Verónica Bautista-Piña, Carlos Dominguez-Reyes, Felipe Villegas-Carlos, et al. 2019. "Expression of Long Non-Coding RNA ENSG00000226738 (LncKLHDC7B) Is Enriched in the Immunomodulatory Triple-Negative Breast Cancer Subtype and Its Alteration Promotes Cell Migration, Invasion, and Resistance to Cell Death." *Molecular Oncology* 13 (4): 909–27.

Bergenfelz, Caroline, Alexander Gaber, Roni Allaoui, Meliha Mehmeti, Karin Jirström, Tomas Leanderson, and Karin Leandersson. 2015. "S100A9 Expressed in ER(-)PgR(-) Breast Cancers Induces Inflammatory Cytokines and Is Associated with an Impaired Overall Survival." *British Journal of Cancer* 113 (8): 1234–43.

Bernstein, L., and M. F. Press. 1998. "Does Estrogen Receptor Expression in Normal Breast Tissue Predict Breast Cancer Risk?" *Journal of the National Cancer Institute*.

Bhattarai, Shristi, Geetanjali Saini, Keerthi Gogineni, and Ritu Aneja. 2020. "Quadruple-Negative Breast Cancer: Novel Implications for a New Disease." *Breast Cancer Research: BCR* 22 (1): 1–11.

Boughorbel, Sabri, Rashid Al-Ali, and Naser Elkum. 2016. "Model Comparison for Breast Cancer Prognosis Based on Clinical Data." *PloS One* 11 (1): e0146413.

Bouzubar, N., K. J. Walker, K. Griffiths, I. O. Ellis, C. W. Elston, J. F. R. Robertson, R. W. Blamey, and R. I. Nicholson. 1989. "Ki67 Immunostaining in Primary Breast Cancer: Pathological and Clinical Associations." *British Journal of Cancer* 59 (6): 943–47.

Brumec, Maša, Monika Sobočan, Iztok Takač, and Darja Arko. 2021. "Clinical Implications of

Androgen-Positive Triple-Negative Breast Cancer." *Cancers* 13 (7). https://doi.org/10.3390/cancers13071642.
Burstein, Matthew D., Anna Tsimelzon, Graham M. Poage, Kyle R. Covington, Alejandro Contreras, Suzanne A. W. Fuqua, Michelle I. Savage, et al. 2015. "Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-Negative Breast Cancer." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 21 (7): 1688–98.
Cao, Q., X. Chen, X. Wu, R. Liao, P. Huang, Y. Tan, L. Wang, G. Ren, J. Huang, and C. Dong. 2018. "Inhibition of UGT8 Suppresses Basal-like Breast Cancer Progression by Attenuating Sulfatide-αVβ5 Axis." *The Journal of Experimental Medicine* 215 (6). https://doi.org/10.1084/jem.20172048.
Carey, Lisa A., E. Claire Dees, Lynda Sawyer, Lisa Gatti, Dominic T. Moore, Frances Collichio, David W. Ollila, Carolyn I. Sartor, Mark L. Graham, and Charles M. Perou. 2007. "The Triple Negative Paradox: Primary Tumor Chemosensitivity of Breast Cancer Subtypes." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 13 (8): 2329–34.
Carey, Lisa A., Charles M. Perou, Chad A. Livasy, Lynn G. Dressler, David Cowan, Kathleen Conway, Gamze Karaca, et al. 2006. "Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study." *JAMA: The Journal of the American Medical Association* 295 (21): 2492–2502.
Cheang, Maggie C. U., Miguel Martin, Torsten O. Nielsen, Aleix Prat, David Voduc, Alvaro Rodriguez-Lescure, Amparo Ruiz, et al. 2015. "Defining Breast Cancer Intrinsic Subtypes by Quantitative Receptor Expression." *The Oncologist* 20 (5): 474–82.
Cheng, Shun-Wen, Po-Chih Chen, Tzong-Rong Ger, Hui-Wen Chiu, and Yuan-Feng Lin. 2021. "Serves as a Potential Marker to Predict a Favorable Response in Triple-Negative Breast Cancer Patients Receiving a Taxane-Based Chemotherapy." *Journal of Personalized Medicine* 11 (3). https://doi.org/10.3390/jpm11030197.
Chen, Jie, Jin Zhu, Shuai-Jun Xu, Jun Zhou, Xiao-Fei Ding, Yong Liang, Guang Chen, and Hong-Sheng Lu. 2022. "Transmembrane 4 L Six Family Member 1 Suppresses Hormone Receptor--Positive, HER2-Negative Breast Cancer Cell Proliferation." *Frontiers in Pharmacology* 13 (January): 770993.
Chen, Xi, Jiang Li, William H. Gray, Brian D. Lehmann, Joshua A. Bauer, Yu Shyr, and Jennifer A. Pietenpol. 2012. "TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer." *Cancer Informatics* 11 (July): 147–56.
Chiu, I-Jen, Yung-Ho Hsu, Jeng-Shou Chang, Jou-Chun Yang, Hui-Wen Chiu, and Yuan-Feng Lin. 2020. "Lactotransferrin Downregulation Drives the Metastatic Progression in Clear Cell Renal Cell Carcinoma." *Cancers* 12 (4). https://doi.org/10.3390/cancers12040847.
Chuan, Tian, Tian Li, and Cui Yi. 2020. "Identification of CXCR4 and CXCL10 as Potential Predictive Biomarkers in Triple Negative Breast Cancer (TNBC)." *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 26 (January): e918281.
Colomer, R., I. Aranda-López, J. Albanell, T. García-Caballero, E. Ciruelos, M. Á. López-García, J. Cortés, F. Rojo, M. Martín, and J. Palacios-Calvo. 2018. "Biomarkers in Breast Cancer: A Consensus Statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathology." *Clinical & Translational Oncology: Official Publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico* 20 (7): 815–26.
Condon, Jennifer C., Daniel B. Hardy, Kelly Kovaric, and Carole R. Mendelson. 2006. "Up-Regulation of the Progesterone Receptor (PR)-C Isoform in Laboring Myometrium by Activation of Nuclear Factor-kappaB May Contribute to the Onset of Labor through Inhibition of PR Function." *Molecular Endocrinology* 20 (4): 764–75.
Cserni, G., S. Bianchi, V. Vezzosi, H. Peterse, A. Sapino, R. Arisio, A. Reiner-Concin, et al. 2006. "The Value of Cytokeratin Immunohistochemistry in the Evaluation of Axillary Sentinel Lymph Nodes in Patients with Lobular Breast Carcinoma." *Journal of Clinical Pathology* 59 (5): 518–22.
Dai, Xiaofeng, Liangjian Xiang, Ting Li, and Zhonghu Bai. 2016. "Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes." *Journal of Cancer* 7 (10): 1281–94.
Dalton, Lori, Virginia Ballarin, and Marcel Brun. 2009. "Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics." *Current Genomics* 10 (6): 430–45.
Daniel, Andrea R., Ming Qiu, Emily J. Faivre, Julie Hanson Ostrander, Andrew Skildum, and Carol A. Lange. 2007. "Linkage of Progestin and Epidermal Growth Factor Signaling: Phosphorylation of Progesterone Receptors Mediates Transcriptional Hypersensitivity and Increased Ligand-Independent Breast Cancer Cell Growth." *Steroids* 72 (2): 188–201.
David Nathanson, S., Shravan Leonard-Murali, Charlotte Burmeister, Laura Susick, and Patricia Baker. 2020. "Clinicopathological Evaluation of the Potential Anatomic Pathways of Systemic Metastasis from Primary Breast Cancer Suggests an Orderly Spread Through the Regional Lymph Nodes." *Annals of*

Surgical Oncology 27 (12): 4810–18.
Del Bano, Joanie, Rémy Florès-Florès, Emmanuelle Josselin, Armelle Goubard, Laetitia Ganier, Rémy Castellano, Patrick Chames, Daniel Baty, and Brigitte Kerfelec. 2019. "A Bispecific Antibody-Based Approach for Targeting Mesothelin in Triple Negative Breast Cancer." *Frontiers in Immunology* 10 (July): 1593.
Denkert, C., S. Loibl, B. M. Müller, H. Eidtmann, W. D. Schmitt, W. Eiermann, B. Gerber, et al. 2013. "Ki67 Levels as Predictive and Prognostic Parameters in Pretherapeutic Breast Cancer Core Biopsies: A Translational Investigation in the Neoadjuvant GeparTrio Trial." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 24 (11): 2786–93.
Dent, Rebecca, Maureen Trudeau, Kathleen I. Pritchard, Wedad M. Hanna, Harriet K. Kahn, Carol A. Sawka, Lavina A. Lickley, Ellen Rawlinson, Ping Sun, and Steven A. Narod. 2007. "Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 13 (15 Pt 1): 4429–34.
Deocesano-Pereira, Carlos, Raquel Arminda Carvalho Machado, Henrique Cesar de Jesus-Ferreira, Thiago Marchini, Tulio Felipe Pereira, Ana Claudia Oliveira Carreira, and Mari Cleide Sogayar. 2019. "Functional Impact of the Long Non-coding RNA MEG3 Deletion by CRISPR/Cas9 in the Human Triple Negative Metastatic Hs578T Cancer Cell Line." *Oncology Letters* 18 (6): 5941–51.
Dey, Nandini, Benjamin G. Barwick, Carlos S. Moreno, Maja Ordanic-Kodani, Zhengjia Chen, Gabriella Oprea-Ilies, Weining Tang, et al. 2013. "Wnt Signaling in Triple Negative Breast Cancer Is Associated with Metastasis." *BMC Cancer* 13 (November): 537.
Dhahri, Habib, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi. 2019. "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms." *Journal of Healthcare Engineering* 2019 (November). https://doi.org/10.1155/2019/4253641.
Donzelli, Sara, Elisa Milano, Magdalena Pruszko, Andrea Sacconi, Silvia Masciarelli, Ilaria Iosue, Elisa Melucci, et al. 2018. "Expression of ID4 Protein in Breast Cancer Cells Induces Reprogramming of Tumour-Associated Macrophages." *Breast Cancer Research: BCR* 20 (1): 1–15.
Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. 2015. "Breast Cancer Statistics and Prediction Methodology: A Systematic Review and Analysis." *Asian Pacific Journal of Cancer Prevention: APJCP* 16 (10): 4237–45.
Echavarria, Isabel, Sara López-Tarruella, Antoni Picornell, Jose Ángel García-Saenz, Yolanda Jerez, Katherine Hoadley, Henry L. Gómez, et al. 2018. "Pathological Response in a Triple-Negative Breast Cancer Cohort Treated with Neoadjuvant Carboplatin and Docetaxel According to Lehmann's Refined Classification." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 24 (8): 1845–52.
Ehmsen, Sidse, Lea Tykgaard Hansen, Martin Bak, Charlotte Brasch-Andersen, Henrik J. Ditzel, and Rikke Leth-Larsen. 2015. "S100A14 Is a Novel Independent Prognostic Biomarker in the Triple-Negative Breast Cancer Subtype." *International Journal of Cancer. Journal International Du Cancer* 137 (9): 2093–2103.
Elston, C. W. 2005. "Classification and Grading of Invasive Breast Carcinoma." *Verhandlungen Der Deutschen Gesellschaft Fur Pathologie* 89: 35–44.
Ensenyat-Mendez, Miquel, Pere Llinàs-Arias, Javier I. J. Orozco, Sandra Íñiguez-Muñoz, Matthew P. Salomon, Borja Sesé, Maggie L. DiNome, and Diego M. Marzese. 2021. "Current Triple-Negative Breast Cancer Subtypes: Dissecting the Most Aggressive Form of Breast Cancer." *Frontiers in Oncology* 11 (June): 681476.
"Estrogen Receptor and Breast Cancer." 2001. *Seminars in Cancer Biology* 11 (5): 339–52.
Fakhri, Ghina B., Reem S. Akel, Maya K. Khalil, Deborah A. Mukherji, Fouad I. Boulos, and Arafat H. Tfayli. 2018. "Concordance between Immunohistochemistry and Microarray Gene Expression Profiling for Estrogen Receptor, Progesterone Receptor, and HER2 Receptor Statuses in Breast Cancer Patients in Lebanon." *International Journal of Breast Cancer* 2018 (May). https://doi.org/10.1155/2018/8530318.
Farabaugh, Susan M., David N. Boone, and Adrian V. Lee. 2015. "Role of IGF1R in Breast Cancer Subtypes, Stemness, and Lineage Differentiation." *Frontiers in Endocrinology* 6 (April): 59.
Ferreira, Paulo Michel Pinheiro, and Cláudia Pessoa. 2017. "Molecular Biology of Human Epidermal Receptors, Signaling Pathways and Targeted Therapy against Cancers: New Evidences and Old Challenges." *Brazilian Journal of Pharmaceutical Sciences* 53 (2). https://doi.org/10.1590/s2175-97902017000216076.
Fitzgibbons, Patrick L., Deborah A. Dillon, Randa Alsabeh, Michael A. Berman, Daniel F. Hayes, David G. Hicks, Kevin S. Hughes, and Sharon Nofech-Mozes. 2014. "Template for Reporting Results of Biomarker Testing of Specimens from Patients with Carcinoma of the Breast." *Archives of Pathology &*

Laboratory Medicine 138 (5): 595–601.
Flamant, Lionel, Edith Roegiers, Michael Pierre, Aurélie Hayez, Christiane Sterpin, Olivier De Backer, Thierry Arnould, Yves Poumay, and Carine Michiels. 2012. "TMEM45A Is Essential for Hypoxia-Induced Chemoresistance in Breast and Liver Cancer Cells." *BMC Cancer* 12 (1): 1–16.
Fujimoto, Nariaki, and Shigeyuki Kitamura. 2004. "Effects of Environmental Estrogenic Chemicals on AP1 Mediated Transcription with Estrogen Receptors Alpha and Beta." *The Journal of Steroid Biochemistry and Molecular Biology* 88 (1): 53–59.
Gajria, Devika, and Sarat Chandarlapaty. 2011. "HER2-Amplified Breast Cancer: Mechanisms of Trastuzumab Resistance and Novel Targeted Therapies." *Expert Review of Anticancer Therapy* 11 (2): 263–75.
Ganggayah, Mogana Darshini, Nur Aishah Taib, Yip Cheng Har, Pietro Lio, and Sarinder Kaur Dhillon. 2019. "Predicting Factors for Survival of Breast Cancer Patients Using Machine Learning Techniques." *BMC Medical Informatics and Decision Making* 19 (1): 1–17.
Gasca, J., M. L. Flores, R. Jiménez-Guerrero, M. E. Sáez, I. Barragán, M. Ruíz-Borrego, M. Tortolero, F. Romero, C. Sáez, and M. A. Japón. 2020. "EDIL3 Promotes Epithelial–mesenchymal Transition and Paclitaxel Resistance through Its Interaction with Integrin αVβ3 in Cancer Cells." *Cell Death Discovery* 6 (1): 1–14.
Gerdes, J., U. Schwab, H. Lemke, and H. Stein. 1983. "Production of a Mouse Monoclonal Antibody Reactive with a Human Nuclear Antigen Associated with Cell Proliferation." *International Journal of Cancer. Journal International Du Cancer* 31 (1): 13–20.
Geyer, Felipe C., Maria A. Lopez-Garcia, Maryou B. Lambros, and Jorge S. Reis-Filho. 2009. "Genetic Characterization of Breast Cancer and Implications for Clinical Management." *Journal of Cellular and Molecular Medicine* 13 (10): 4090–4103.
Gong, Chen, Jin Zou, Mingsheng Zhang, Jie Zhang, Shanshan Xu, Siqi Zhu, Mengqi Yang, et al. 2019. "Upregulation of MGP by HOXC8 Promotes the Proliferation, Migration, and EMT Processes of Triple-Negative Breast Cancer." *Molecular Carcinogenesis* 58 (10): 1863–75.
Graham, J. D., C. Yeates, R. L. Balleine, S. S. Harvey, J. S. Milliken, A. M. Bilous, and C. L. Clarke. 1995. "Characterization of Progesterone Receptor A and B Expression in Human Breast Cancer." *Cancer Research* 55 (21): 5063–68.
Gutierrez, Carolina, and Rachel Schiff. 2011. "HER2: Biology, Detection, and Clinical Implications." *Archives of Pathology & Laboratory Medicine* 135 (1): 55–62.
Hammond, M. Elizabeth H., Daniel F. Hayes, Mitch Dowsett, D. Craig Allred, Karen L. Hagerty, Sunil Badve, Patrick L. Fitzgibbons, et al. 2010. "American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, April. https://doi.org/10.1200/JCO.2009.25.6529.
Harbeck, Nadia, Christoph Thomssen, and Michael Gnant. 2013. "St. Gallen 2013: Brief Preliminary Summary of the Consensus Discussion." *Breast Care* 8 (2): 102–9.
Hassiotou, Foteini, and Donna Geddes. 2013. "Anatomy of the Human Mammary Gland: Current Status of Knowledge." *Clinical Anatomy* 26 (1): 29–48.
Henssen, Anton G., Richard Koche, Jiali Zhuang, Eileen Jiang, Casie Reed, Amy Eisenberg, Eric Still, et al. 2017. "PGBD5 Promotes Site-Specific Oncogenic Mutations in Human Tumors." *Nature Genetics* 49 (7): 1005–14.
"Hormone Receptors in Breast Cancer." 1978. *British Medical Journal* 2 (6130): 77–78.
Hou, Can, Xiaorong Zhong, Ping He, Bin Xu, Sha Diao, Fang Yi, Hong Zheng, and Jiayuan Li. 2020. "Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development." *JMIR Medical Informatics* 8 (6): e17364.
Hwang, Harry C., and Allen M. Gown. 2016. "Evaluation of Human Epidermal Growth Factor Receptor 2 (HER2) Gene Status in Human Breast Cancer Formalin-Fixed Paraffin-Embedded (FFPE) Tissue Specimens by Fluorescence In Situ Hybridization (FISH)." *Methods in Molecular Biology* 1406: 61–70.
"Immunotherapy for Triple-Negative Breast Cancer: A Molecular Insight into the Microenvironment, Treatment, and Resistance." 2021. *Journal of the National Cancer Center* 1 (3): 75–87.
Inic, Zorka, Milan Zegarac, Momcilo Inic, Ivan Markovic, Zoran Kozomara, Igor Djurisic, Ivana Inic, Gordana Pupic, and Snezana Jancic. 2014. "Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information." *Clinical Medicine Insights. Oncology* 8 (September): 107–11.
Iqbal, Nida, and Naveed Iqbal. 2014. "Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers:

Overexpression and Therapeutic Implications." *Molecular Biology International* 2014 (September): 852748.
Ishikawa, Takashi, Yasushi Ichikawa, Daisuke Shimizu, Takeshi Sasaki, Mikiko Tanabe, Takashi Chishima, Kazuaki Takabe, and Itaru Endo. 2014. "The Role of HER-2 in Breast Cancer." *Journal of Surgery and Science* 2 (1): 4–9.
Jacobsen, Britta M., Jennifer K. Richer, Stephanie A. Schittone, and Kathryn B. Horwitz. 2002. "New Human Breast Cancer Cells to Study Progesterone Receptor Isoform Ratio Effects and Ligand-Independent Gene Regulation." *The Journal of Biological Chemistry* 277 (31): 27793–800.
Jacobsen, Britta M., Stephanie A. Schittone, Jennifer K. Richer, and Kathryn B. Horwitz. 2005. "Progesterone-Independent Effects of Human Progesterone Receptors (PRs) in Estrogen Receptor-Positive Breast Cancer: PR Isoform-Specific Gene Regulation and Tumor Biology." *Molecular Endocrinology* 19 (3): 574–87.
Jézéquel, Pascal, Olivier Kerdraon, Hubert Hondermarck, Catherine Guérin-Charbonnel, Hamza Lasla, Wilfried Gouraud, Jean-Luc Canon, et al. 2019. "Identification of Three Subtypes of Triple-Negative Breast Cancer with Potential Therapeutic Implications." *Breast Cancer Research: BCR* 21 (1): 65.
Jung, Yoonsuh, and Jianhua Hu. 2015. "A -Fold Averaging Cross-Validation Procedure." *Journal of Nonparametric Statistics* 27 (2): 167–79.
Keam, Bhumsuk, Seock-Ah Im, Kyung-Hun Lee, Sae-Won Han, Do-Youn Oh, Jee Hyun Kim, Se-Hoon Lee, et al. 2011. "Ki-67 Can Be Used for Further Classification of Triple Negative Breast Cancer into Two Subtypes with Different Response and Prognosis." *Breast Cancer Research: BCR* 13 (2): 1–7.
Kim, Ga-Eon, Ji Shin Lee, Yoo-Duk Choi, Kyung-Hwa Lee, Jae Hyuk Lee, Jong Hee Nam, Chan Choi, et al. 2014. "Expression of Matrix Metalloproteinases and Their Inhibitors in Different Immunohistochemical-Based Molecular Subtypes of Breast Cancer." *BMC Cancer* 14 (1): 1–10.
Kloten, Vera, Martin Schlensog, Julian Eschenbruch, Janina Gasthaus, Janina Tiedemann, Jolein Mijnes, Timon Heide, Till Braunschweig, Ruth Knüchel, and Edgar Dahl. 2016. "Abundant NDRG2 Expression Is Associated with Aggressiveness and Unfavorable Patients' Outcome in Basal-Like Breast Cancer." *PloS One* 11 (7): e0159073.
Koopmans, Tim, and Yuval Rinkevich. 2018. "Mesothelial to Mesenchyme Transition as a Major Developmental and Pathological Player in Trunk Organs and Their Cavities." *Communications Biology* 1 (1): 1–14.
Kuroda, Hajime, Tsengelmaa Jamiyan, Rin Yamaguchi, Akinari Kakumoto, Akihito Abe, Oi Harada, Bayarmaa Enkhbat, and Atsuko Masunaga. 2021. "Prognostic Value of Tumor-Infiltrating B Lymphocytes and Plasma Cells in Triple-Negative Breast Cancer." *Breast Cancer* 28 (4): 904–14.
Lee, Se Kyung, Soo Youn Bae, Jun Ho Lee, Hyun-Chul Lee, Hawoo Yi, Won Ho Kil, Jeong Eon Lee, Seok Won Kim, and Seok Jin Nam. 2015. "Distinguishing Low-Risk Luminal A Breast Cancer Subtypes with Ki-67 and p53 Is More Predictive of Long-Term Survival." *PloS One* 10 (8): e0124658.
Lehmann, Brian D., Joshua A. Bauer, Xi Chen, Melinda E. Sanders, A. Bapsi Chakravarthy, Yu Shyr, and Jennifer A. Pietenpol. 2011. "Identification of Human Triple-Negative Breast Cancer Subtypes and Preclinical Models for Selection of Targeted Therapies." *The Journal of Clinical Investigation* 121 (7): 2750–67.
Lehmann, Brian D., Bojana Jovanović, Xi Chen, Monica V. Estrada, Kimberly N. Johnson, Yu Shyr, Harold L. Moses, Melinda E. Sanders, and Jennifer A. Pietenpol. 2016. "Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection." *PloS One* 11 (6): e0157368.
Liang, Yuan-Ke, Ze-Kun- Deng, Mu-Tong Chen, Si-Qi Qiu, Ying-Sheng Xiao, Yu-Zhu Qi, Qin Xie, et al. 2021. "CXCL9 Is a Potential Biomarker of Immune Infiltration Associated With Favorable Prognosis in ER-Negative Breast Cancer." *Frontiers in Oncology* 11 (August): 710286.
Li, Christopher I., Yuping Zhang, Marcin Cieślik, Yi-Mi Wu, Lanbo Xiao, Erin Cobain, Mei-Tzu C. Tang, et al. 2021. "Cancer Cell Intrinsic and Immunologic Phenotypes Determine Clinical Outcomes in Basal-like Breast Cancer." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 27 (11): 3079–93.
Lim, Gyeong Back, Young-Ae Kim, Jeong-Han Seo, Hee Jin Lee, Gyungyub Gong, and Sung Hee Park. 2020. "Prediction of Prognostic Signatures in Triple-Negative Breast Cancer Based on the Differential Expression Analysis via NanoString nCounter Immune Panel." *BMC Cancer* 20 (1): 1–13.
Li, Rui-Xin, Zi-Hua Chen, and Zhi-Kang Chen. 2014. "The Role of EPH Receptors in Cancer-Related Epithelial-Mesenchymal Transition." *Chinese Journal of Cancer* 33 (5): 231–40.
Liu, Rong-Zong, Elizabeth Garcia, Darryl D. Glubrecht, Ho Yin Poon, John R. Mackey, and Roseline Godbout. 2015. "CRABP1 Is Associated with a Poor Prognosis in Breast Cancer: Adding to the

Complexity of Breast Cancer Cell Response to Retinoic Acid." *Molecular Cancer* 14 (1): 1–16.
Liu, Yin, Puspa R. Pandey, Sambad Sharma, Fei Xing, Kerui Wu, Amar Chittiboyina, Shih-Ying Wu, Abhishek Tyagi, and Kounosuke Watabe. 2019. "ID2 and GJB2 Promote Early-Stage Breast Cancer Progression by Regulating Cancer Stemness." *Breast Cancer Research and Treatment* 175 (1): 77–90.
Liu, Yi-Rong, Yi-Zhou Jiang, Xiao-En Xu, Ke-Da Yu, Xi Jin, Xin Hu, Wen-Jia Zuo, et al. 2016. "Comprehensive Transcriptome Analysis Identifies Novel Molecular Subtypes and Subtype-Specific RNAs of Triple-Negative Breast Cancer." *Breast Cancer Research: BCR* 18 (1): 33.
Liu, Zhixian, Mengyuan Li, Zehang Jiang, and Xiaosheng Wang. 2018. "A Comprehensive Immunologic Portrait of Triple-Negative Breast Cancer." *Translational Oncology* 11 (2): 311–29.
Loibl, Sibylle, Joyce O'Shaughnessy, Michael Untch, William M. Sikov, Hope S. Rugo, Mark D. McKee, Jens Huober, et al. 2018. "Addition of the PARP Inhibitor Veliparib plus Carboplatin or Carboplatin Alone to Standard Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer (BrighTNess): A Randomised, Phase 3 Trial." *The Lancet Oncology* 19 (4): 497–509.
Lv, Xuemei, Miao He, Yanyun Zhao, Liwen Zhang, Wenjing Zhu, Longyang Jiang, Yuanyuan Yan, et al. 2019. "Identification of Potential Key Genes and Pathways Predicting Pathogenesis and Prognosis for Triple-Negative Breast Cancer." *Cancer Cell International* 19 (1): 1–12.
Marwah, Nisha, Ashima Batra, Sanjay Marwah, Veena Gupta, Samta Shakya, and Rajeev Sen. 2018. "Correlation of Proliferative Index with Various Clinicopathologic Prognostic Parameters in Primary Breast Carcinoma: A Study from North India." *Journal of Cancer Research and Therapeutics* 14 (3): 537.
Matheson, Timothy D., and Paul D. Kaufman. 2017. "The p150N Domain of Chromatin Assembly Factor-1 Regulates Ki-67 Accumulation on the Mitotic Perichromosomal Layer." *Molecular Biology of the Cell* 28 (1): 21–29.
Matros, Evan, Zhigang C. Wang, Gabriela Lodeiro, Alexander Miron, J. Dirk Iglehart, and Andrea L. Richardson. 2005. "BRCA1 Promoter Methylation in Sporadic Breast Tumors: Relationship to Gene Expression Profiles." *Breast Cancer Research and Treatment* 91 (2): 179–86.
McElwee, John L., Sunish Mohanan, Obi L. Griffith, Heike C. Breuer, Lynne J. Anguish, Brian D. Cherrington, Ashley M. Palmer, et al. 2012. "Identification of PADI2 as a Potential Breast Cancer Biomarker and Therapeutic Target." *BMC Cancer* 12 (October): 500.
McQuerry, Jasmine A., David F. Jenkins, Susan E. Yost, Yuqing Zhang, Daniel Schmolze, W. Evan Johnson, Yuan Yuan, and Andrea H. Bild. 2019. "Pathway Activity Profiling of Growth Factor Receptor Network and Stemness Pathways Differentiates Metaplastic Breast Cancer Histological Subtypes." *BMC Cancer* 19 (1): 1–14.
Mehta, Gaurav A., Pooja Khanna, and Michael L. Gatza. 2019. "Emerging Role of SOX Proteins in Breast Cancer Development and Maintenance." *Journal of Mammary Gland Biology and Neoplasia* 24 (3): 213–30.
Mina, Alain, Rachel Yoder, and Priyanka Sharma. 2017. "Targeting the Androgen Receptor in Triple-Negative Breast Cancer: Current Perspectives." *OncoTargets and Therapy* 10 (September): 4675–85.
Montagna, Emilia, Patrick Maisonneuve, Nicole Rotmensz, Giuseppe Cancello, Monica Iorfida, Alessandra Balduzzi, Viviana Galimberti, et al. 2013. "Heterogeneity of Triple-Negative Breast Cancer: Histologic Subtyping to Inform the Outcome." *Clinical Breast Cancer* 13 (1): 31–39.
Mulac-Jericevic, Biserka, John P. Lydon, Francesco J. DeMayo, and Orla M. Conneely. 2003. "Defective Mammary Gland Morphogenesis in Mice Lacking the Progesterone Receptor B Isoform." *Proceedings of the National Academy of Sciences of the United States of America* 100 (17): 9744–49.
Mulac-Jericevic, B., R. A. Mullinax, F. J. DeMayo, J. P. Lydon, and O. M. Conneely. 2000. "Subgroup of Reproductive Functions of Progesterone Mediated by Progesterone Receptor-B Isoform." *Science* 289 (5485): 1751–54.
Nandi, S., R. C. Guzman, and J. Yang. 1995. "Hormones and Mammary Carcinogenesis in Mice, Rats, and Humans: A Unifying Hypothesis." *Proceedings of the National Academy of Sciences of the United States of America* 92 (9): 3650–57.
Narita, D., E. Seclaman, A. Anghel, R. Ilina, N. Cireap, S. Negru, I. O. Sirbu, S. Ursoniu, and C. Marian. 2016. "Altered Levels of Plasma Chemokines in Breast Cancer and Their Association with Clinical and Pathological Characteristics." *Neoplasma* 63 (1): 141–49.
Nielsen, Torsten O., Samuel C. Y. Leung, David L. Rimm, Andrew Dodson, Balazs Acs, Sunil Badve, Carsten Denkert, et al. 2020. "Assessment of Ki67 in Breast Cancer: Updated Recommendations From the International Ki67 in Breast Cancer Working Group." *Journal of the National Cancer Institute*, December. https://doi.org/10.1093/jnci/djaa201.

Perou, C. M., T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, et al. 2000. "Molecular Portraits of Human Breast Tumours." *Nature* 406 (6797): 747–52.
Pike, M. C., D. V. Spicer, L. Dahmoush, and M. F. Press. 1993. "Estrogens, Progestogens, Normal Breast Cell Proliferation, and Breast Cancer Risk." *Epidemiologic Reviews* 15 (1): 17–35.
Podo, Franca, Lutgarde M. C. Buydens, Hadassa Degani, Riet Hilhorst, Edda Klipp, Ingrid S. Gribbestad, Sabine Van Huffel, et al. 2010. "Triple-Negative Breast Cancer: Present Challenges and New Perspectives." *Molecular Oncology* 4 (3): 209–29.
Powell, George, Heather Roche, and William R. Roche. 2011. "Expression of Calretinin by Breast Carcinoma and the Potential for Misdiagnosis of Mesothelioma." *Histopathology* 59 (5): 950–56.
Prossnitz, Eric R., Jeffrey B. Arterburn, and Larry A. Sklar. 2007. "GPR30: A G Protein-Coupled Receptor for Estrogen." *Molecular and Cellular Endocrinology* 265-266 (February): 138–42.
Qiu, Ming, and Carol A. Lange. 2003. "MAP Kinases Couple Multiple Functions of Human Progesterone Receptors: Degradation, Transcriptional Synergy, and Nuclear Association." *The Journal of Steroid Biochemistry and Molecular Biology* 85 (2-5): 147–57.
Querzoli, P., G. Albonico, S. Ferretti, R. Rinaldi, E. Magri, M. Indelli, and I. Nenci. 1996. "MIB-1 Proliferative Activity in Invasive Breast Cancer Measured by Image Analysis." *Journal of Clinical Pathology* 49 (11): 926–30.
Rajula, Hema Sekhar Reddy, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. 2020. "Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment." *Medicina* 56 (9): 455.
Rakha, Emad A., Maysa E. El-Sayed, Andrew H. S. Lee, Christopher W. Elston, Matthew J. Grainge, Zsolt Hodi, Roger W. Blamey, and Ian O. Ellis. 2008. "Prognostic Significance of Nottingham Histologic Grade in Invasive Breast Carcinoma." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 26 (19): 3153–58.
Ricciardelli, Carmela, Noor A. Lokman, Carmen E. Pyragius, Miranda P. Ween, Anne M. Macpherson, Andrew Ruszkiewicz, Peter Hoffmann, and Martin K. Oehler. 2017. "Keratin 5 Overexpression Is Associated with Serous Ovarian Cancer Recurrence and Chemotherapy Resistance." *Oncotarget* 8 (11): 17819–32.
Richer, Jennifer K., Britta M. Jacobsen, Nicole G. Manning, M. Greg Abel, Douglas M. Wolf, and Kathryn B. Horwitz. 2002. "Differential Gene Regulation by the Two Progesterone Receptor Isoforms in Human Breast Cancer Cells." *The Journal of Biological Chemistry* 277 (7): 5209–18.
Rodriguez, Mayra Z., Cesar H. Comin, Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio, Luciano da F. Costa, and Francisco A. Rodrigues. 2019. "Clustering Algorithms: A Comparative Approach." *PloS One* 14 (1): e0210236.
"Roles of S100 Family Members in Drug Resistance in Tumors: Status and Prospects." 2020. *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie* 127 (July): 110156.
Ropo Ebenezer, Ogunsakin, and Siaka Lougue. 2019. "Bayesian Generalized Linear Mixed Modeling of Breast Cancer." *Iranian Journal of Public Health* 48 (6): 1043–51.
Rötzer, Vera, Eva Hartlieb, Franziska Vielmuth, Martin Gliem, Volker Spindler, and Jens Waschke. 2015. "E-Cadherin and Src Associate with Extradesmosomal Dsg3 and Modulate Desmosome Assembly and Adhesion." *Cellular and Molecular Life Sciences: CMLS* 72 (24): 4885–97.
Roy, D., H. W. Strobel, and J. G. Liehr. 1991. "Cytochrome b5-Mediated Redox Cycling of Estrogen." *Archives of Biochemistry and Biophysics* 285 (2): 331–38.
Russnes, Hege G., Ole Christian Lingjærde, Anne-Lise Børresen-Dale, and Carlos Caldas. 2017. "Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters." *The American Journal of Pathology* 187 (10): 2152–62.
Shah, Deep, and Clodia Osipo. 2016. "Cancer Stem Cells and HER2 Positive Breast Cancer: The Story so Far." *Genes & Diseases* 3 (2): 114–23.
Shen, T., K. B. Horwitz, and C. A. Lange. 2001. "Transcriptional Hyperactivity of Human Progesterone Receptors Is Coupled to Their Ligand-Dependent down-Regulation by Mitogen-Activated Protein Kinase-Dependent Phosphorylation of Serine 294." *Molecular and Cellular Biology* 21 (18): 6122–31.
Sizemore, Gina M., Steven T. Sizemore, Darcie D. Seachrist, and Ruth A. Keri. 2014. "GABA(A) Receptor Pi (GABRP) Stimulates Basal-like Breast Cancer Cell Migration through Activation of Extracellular-Regulated Kinase 1/2 (ERK1/2)." *The Journal of Biological Chemistry* 289 (35): 24102–13.
Sizemore, Steven T., Gina M. Sizemore, Christine N. Booth, Cheryl L. Thompson, Paula Silverman, Gurkan Bebek, Fadi W. Abdul-Karim, Stefanie Avril, and Ruth A. Keri. 2014. "Hypomethylation of the MMP7 Promoter and Increased Expression of MMP7 Distinguishes the Basal-like Breast Cancer Subtype from

Other Triple-Negative Tumors." *Breast Cancer Research and Treatment* 146 (1): 25–40.
Sjöberg, Elin, Martin Augsten, Jonas Bergh, Karin Jirström, and Arne Östman. 2016. "Expression of the Chemokine CXCL14 in the Tumour Stroma Is an Independent Marker of Survival in Breast Cancer." *British Journal of Cancer* 114 (10): 1117–24.
Song, Yan-Yan, and Ying Lu. 2015. "Decision Tree Methods: Applications for Classification and Prediction." *Shanghai Archives of Psychiatry* 27 (2): 130–35.
Sørlie, Therese, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, et al. 2001. "Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications." *Proceedings of the National Academy of Sciences of the United States of America* 98 (19): 10869–74.
Sun, Li-Min, Cheng-Li Lin, Sean Sun, Chung Y. Hsu, Zonyin Shae, and Chia-Hung Kao. 2019. "Long-Term Use of Tamoxifen Is Associated With a Decreased Subsequent Meningioma Risk in Patients With Breast Cancer: A Nationwide Population-Based Cohort Study." *Frontiers in Pharmacology* 10. https://doi.org/10.3389/fphar.2019.00674.
Sun, Xiaoming, and Paul D. Kaufman. 2018. "Ki-67: More than a Proliferation Marker." *Chromosoma* 127 (2): 175–86.
Tan, Puay-Hoon, Boon-Huat Bay, George Yip, Sathiyamoorthy Selvarajan, Patrick Tan, Jeanie Wu, Chee-How Lee, and Kuo-Bin Li. 2004. "Immunohistochemical Detection of Ki67 in Breast Cancer Correlates with Transcriptional Regulation of Genes Related to Apoptosis and Cell Death." *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc* 18 (3): 374–81.
Temian, Daiana Cosmina, Laura Ancuta Pop, Alexandra Iulia Irimie, and Ioana Berindan-Neagoe. 2018. "The Epigenetics of Triple-Negative and Basal-Like Breast Cancer: Current Knowledge." *Journal of Breast Cancer* 21 (3): 233–43.
Tian, J-M, B. Ran, C-L Zhang, D-M Yan, and X-H Li. 2018. "Estrogen and Progesterone Promote Breast Cancer Cell Proliferation by Inducing Cyclin G1 Expression." *Brazilian Journal of Medical and Biological Research = Revista Brasileira de Pesquisas Medicas E Biologicas / Sociedade Brasileira de Biofisica ... [et Al.]* 51 (3): 1–7.
Tominaga, K., T. Shimamura, N. Kimura, T. Murayama, D. Matsubara, H. Kanauchi, A. Niida, et al. 2016. "Addiction to the IGF2-ID1-IGF2 Circuit for Maintenance of the Breast Cancer Stem-like Cells." *Oncogene* 36 (9): 1276–86.
Turashvili, Gulisa, and Edi Brogi. 2017. "Tumor Heterogeneity in Breast Cancer." *Frontiers of Medicine* 0. https://doi.org/10.3389/fmed.2017.00227.
Vickers, Andrew J., and Angel M. Cronin. 2010. "Traditional Statistical Methods for Evaluating Prediction Models Are Uninformative as to Clinical Value: Towards a Decision Analytic Framework." *Seminars in Oncology* 37 (1): 31–38.
Wang, Dong-Yu, Zhe Jiang, Yaacov Ben-David, James R. Woodgett, and Eldad Zacksenhaus. 2019. "Molecular Stratification within Triple-Negative Breast Cancer Subtypes." *Scientific Reports* 9 (1): 19107.
Wang, Ning, Kristin A. Eckert, Ali R. Zomorrodi, Ping Xin, Weihua Pan, Debra A. Shearer, Judith Weisz, Costas D. Maranus, and Gary A. Clawson. 2012. "Down-Regulation of HtrA1 Activates the Epithelial-Mesenchymal Transition and ATM DNA Damage Response Pathways." *PloS One* 7 (6): e39446.
Wang, Yipeng, Yibin Xie, Yanan Niu, Peng Song, Ye Liu, Joseph Burnett, Zhihua Yang, et al. 2021. "Carboxypeptidase A4 Negatively Correlates with p53 Expression and Regulates the Stemness of Breast Cancer Cells." *International Journal of Medical Sciences* 18 (8): 1753–59.
"Website." n.d. Accessed March 24, 2021. https://doi.org/10.1007/978-3-030-36664-3_28.
Weigel, Marion T., and Mitch Dowsett. 2010. "Current and Emerging Biomarkers in Breast Cancer: Prognosis and Prediction." *Endocrine-Related Cancer* 17 (4): R245–62.
Wei, L. L., B. M. Norris, and C. J. Baker. 1997. "An N-Terminally Truncated Third Progesterone Receptor Protein, PR(C), Forms Heterodimers with PR(B) but Interferes in PR(B)-DNA Binding." *The Journal of Steroid Biochemistry and Molecular Biology* 62 (4): 287–97.
Wu, Hsing-Jung, Ji Won Oh, Dan F. Spandau, Sunil Tholpady, Jesus Diaz 3rd, Laura J. Schroeder, Carlos D. Offutt, et al. 2017. "Estrogen Modulates Mesenchyme-Epidermis Interactions in the Adult Nipple." *Development* 144 (8): 1498–1509.
Xu, Shen, Shan Yu, Daming Dong, and Leo Tsz On Lee. 2019. "G Protein-Coupled Estrogen Receptor: A Potential Therapeutic Target in Cancer." *Frontiers in Endocrinology* 10 (October): 725.
Yager, James D., and Nancy E. Davidson. 2006. "Estrogen Carcinogenesis in Breast Cancer." *The New*

England Journal of Medicine 354 (3): 270–82.
Yalaza, Metin, Aydın İnan, and Mikdat Bozer. 2016. "Male Breast Cancer." *The Journal of Breast Health* 12 (1): 1–8.
Yanagawa, Masumi, Kenzo Ikemot, Shigeto Kawauchi, Tomoko Furuya, Shigeru Yamamoto, Masaaki Oka, Atunori Oga, Yukiko Nagashima, and Kohsuke Sasaki. 2012. "Luminal A and Luminal B (HER2 Negative) Subtypes of Breast Cancer Consist of a Mixture of Tumors with Different Genotype." *BMC Research Notes* 5 (1): 1–8.
Yao, Hui, Guangchun He, Shichao Yan, Chao Chen, Liujiang Song, Thomas J. Rosol, and Xiyun Deng. 2017. "Triple-Negative Breast Cancer: Is There a Treatment on the Horizon?" *Oncotarget* 8 (1): 1913–24.
Yassin, Nisreen I. R., Shaimaa Omran, Enas M. F. El Houby, and Hemat Allam. 2018. "Machine Learning Techniques for Breast Cancer Computer Aided Diagnosis Using Different Image Modalities: A Systematic Review." *Computer Methods and Programs in Biomedicine* 156 (March): 25–45.
Yin, Li, Jiang-Jie Duan, Xiu-Wu Bian, and Shi-Cang Yu. 2020. "Triple-Negative Breast Cancer Molecular Subtyping and Treatment Progress." *Breast Cancer Research: BCR* 22 (1): 1–13.
Zaha, Dana Carmen. 2014. "Significance of Immunohistochemistry in Breast Cancer." *World Journal of Clinical Oncology* 5 (3): 382–92.
Zheng, Dandan, Chengwei Jiang, Ning Yan, Yayun Miao, Keren Wang, Ge Gao, Yan Jiao, Xiangkai Zhang, Miao He, and Zhaoying Yang. 2020. "Wntless (Wls): A Prognostic Index for Progression and Patient Survival of Breast Cancer." *OncoTargets and Therapy* 13 (December): 12649–59.
Zheng, Hongmei, Sumit Siddharth, Sheetal Parida, Xinhong Wu, and Dipali Sharma. 2021. "Tumor Microenvironment: Key Players in Triple Negative Breast Cancer Immunomodulation." *Cancers* 13 (13). https://doi.org/10.3390/cancers13133357.
Zhu, Xiuzhi, Li Chen, Binhao Huang, Yue Wang, Lei Ji, Jiong Wu, Genhong Di, et al. 2020. "The Prognostic and Predictive Potential of Ki-67 in Triple-Negative Breast Cancer." *Scientific Reports* 10 (1): 1–10.