Artificial intelligence applied to digital pathology to discover new predictors of breast cancer patient outcome

Ingrid Garberis

université PARIS-SACLAY

Artificial intelligence applied to digital pathology to discover new predictors of breast cancer patient outcome

L'intelligence artificielle appliquée à la pathologie numérique pour découvrir de nouveaux prédicteurs de l'évolution des patientes atteintes d'un cancer du sein

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°582 Cancérologie : biologie - médecine – santé (CBMS)
Spécialité de doctorat: Sciences de la vie et santé.
Graduate school : Life Sciences and Health. Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche
Prédicteurs moléculaires et nouvelles cibles en oncologie
(Université Paris-Saclay, INSERM) sous la direction
de **Fabrice ANDRE**, PU-PH à Gustave Roussy et le co-encadrement
de **Magali LACROIX-TRIKI**, PH à Gustave Roussy

Thèse soutenue à Paris-Saclay, le 19 décembre 2022, par

Ingrid Judith GARBERIS

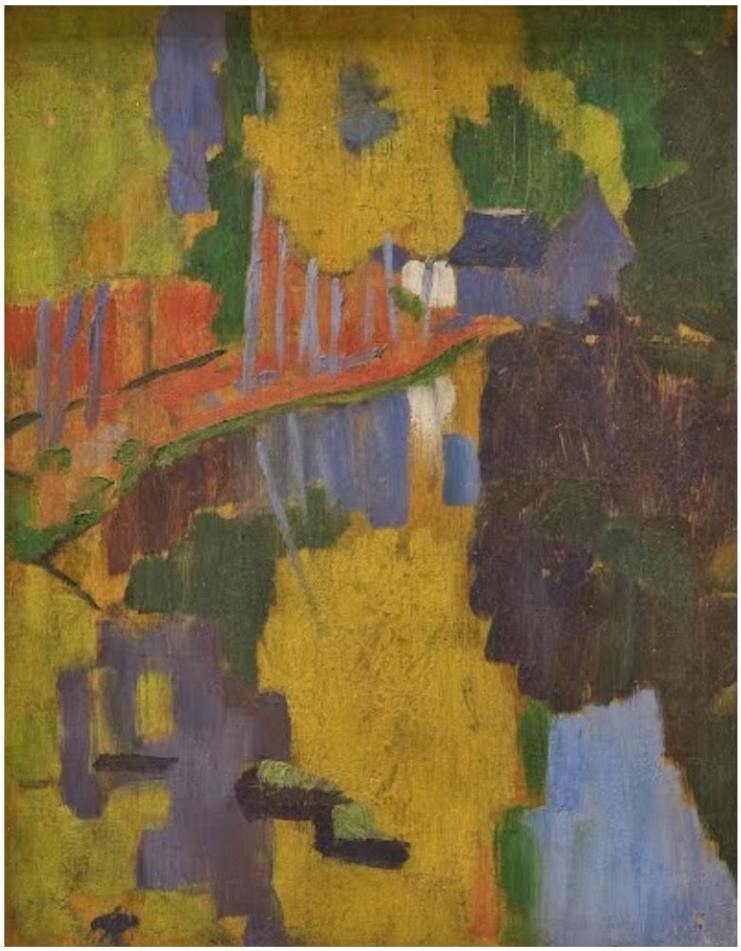
Composition du Jury

Jean-Yves SCOAZEC PU-PH, Gustave Roussy, Villejuif	Président
Emmanuelle CHARAFE-JAUFFRET MCU-PH, Institut Paoli-Calmettes, Marseille	Rapporteur & Examinatrice
Nathalie RIOUX-LECLERCQ PU-PH, CHU Rennes	Rapporteur & Examinatrice
Frédérique PENAULT-LLORCA PU-PH, Centre Jean Perrin, Clermont-Ferrand	Examinatrice

THÈSE DE DOCTORAT

NNT : 2022UPASL103

A picture is worth a thousand words…



The Talisman (Paul Sérusier, oil on wood, 1888)
© Musée d'Orsay, dist. RMN / Patrice Schmidt

Acknowledgements

Je voudrais tout d'abord remercier aux membres du jury : Pr Nathalie Rioux-Leclercq et Pr Emmanuelle Charafe-Jauffret, qui ont accepté d'être rapporteurs et d'évaluer mon manuscrit de thèse, ainsi que Pr Frédérique Penault-Llorca pour l'honneur qu'elle me fait d'être examinatrice.

Je tiens à remercier grandement Pr Jean-Yves Scoazec, qui a bien voulu présider ce jury ainsi que pour son aide tout au long de mon parcours scientifique à Gustave Roussy.

Je remercie également mon directeur de thèse, Pr Fabrice André, de m'avoir ouvert les portes de son labo pour faire ma thèse, ainsi que pour la confiance et pour l'opportunité de participer à de nombreux projets aussi complexes qu'enrichissants.

Un grand merci à Dr Magali Lacroix-Triki pour son encadrement. Merci beaucoup d'avoir accepté cette tâche (même si je ne t'ai pas trop laissé de chance de dire non !) et d'être là dans les moments les plus amusants mais aussi dans les plus durs. Je tiens à t'exprimer ma profonde gratitude pour ton soutien scientifique ainsi que personnel, pour tes conseils, pour ta gentillesse, ta patience et ta sympathie.

J'adresse tous mes remerciements aux Owkinautes et spécialement à Victor, Charlie, Valentin, Kevin. Cette thèse est le fruit d'une longue collaboration avec Owkin, commencée avec le Data Challenge en 2019 et poursuivie par de réunions aussi productives qu'amusantes, de congrès ici et là, et de sorties ! toujours dans la joie et la bonne humeur.

Je remercie toutes les personnes avec qui j'ai partagé ces années de thèse sans lesquelles cette période aurait été plus sombre, plus difficile :

Aux collègues du B2M, notamment aux membres de mon équipe, « l'équipe sein ». Bojana, Ibrahim, Benjamin, mes copains et copines de tous les jours, toujours à l'écoute, je suis très contente d'avoir partagé le bureau et cette aventure avec vous. Merci à Sofia et à Juan, à Michele, Diep, Anna, Olivia, Tony, Semih et Alicia. Vous avez tous fait le poids beaucoup moins lourd tellement de fois… !

Un grand merci à Véro pour toute l'aide et les conseils, ce fut un vrai plaisir d'avoir eu des projets en commun et de partager les matinées au labo ainsi que les

cours de sport et les rentrées en voiture. A Aimie, Amélie P. et Cécile T. pour votre précieuse aide et gentillesse.

Aux équipes « prostate », « ginéco », « ex CTC » et à l'ensemble de l'unité 981 pour les beaux moments partagés dedans et dehors le labo. Grâce à vous toutes et tous, venir au labo a été moins « venir au taff » et plus « rencontrer des ami.e.s ». Jojo : je suis désolée de n'avoir finalement pas pris aucun des sujets de thèse qu'on a envisagés ensemble à la machine à café. A Loïc pour le soutien, à Ludo qui passe nous chercher pour (petit ?) déjeuner à 11h, à Mathieu pour les blagues, à Luce et à Cath pour nous supporter tous ! A Tala, Agathe et Marianne pour les pauses et les invitations à prendre une petite sucrerie pour « se ressourcer ». A Chloé pour l'organisation de magnifiques événements comme la chasse aux œufs, le petit dej' de Noël et les A.P.E.R.O. que tant améliorent le quotidien au labo, à Hadia et Eliza pour le soutien et les conseils. A Nico D. pour tous les dépannages. A Widad, très spécialement, pour ta disponibilité, ta sympathie, ton assistance pour les démarches compliquées !

Aux centraliens qui ont passé par notre équipe et m'ont appris les bases d'une nouvelle discipline : Clément, Younes, Loïc L.B., j'espère avoir bien assimilé les concepts qui m'ont tant aidé avec mon sujet de thèse ainsi qu'avec les nouveaux chemins qui s'ouvrent pour l'anapath !

A la plateforme PETRA, que j'ai connue en tant que module HCP lors de mon premier contrat à GR et notamment à Virginie, Doris, Adeline, merci pour tout ! A Nicolas S. pour tout l'apprentissage en pathologie numérique et pour la patience infinie avec le scanner.

A Laetitia G., je n'oublierai pas ces moments de galère à côté de la Discovery ! A Julien A., pour les connaissances partagées en pathologie et analyse d'images.

Un grand merci également aux assistantes de l'ED582 et particulièrement à Léa Poisot, qui a répondu avec calme et patience aux questions diverses dont je l'accablais.

Merci aux membres de la team argentine à GR : Laura I., Ivana, Virginia, Gonzalo, José, et presque argentine : Elaine, Laura M., pour le partage, le soutien, les conseils, le ramen ! A mes ami.e.s dans les quatre coins du monde : la « famille » argentine à Paris Fer, Mati O. & Anne & Maga, Joaqui & Gioi, Juan G., Isa & Seba, et ailleurs Fede V., J.I. Della Batta, Mati B.,

Michele & Titou, Jotge, Jakob, Thibault, Marco & Mica, Benji & Camille et la « famille » française dans les Cévennes, Martina à N.Y., Gemela & Diego et les biscottis à Buenos Aires, Chicha & Martin à San Luis, ma pseudo-tata Marga à Caseros. Vous êtes les meilleur.e.s !

Mention spéciale pour leur lecture critique d'une première version (très préliminaire) de ce manuscrit à Elaine Limkin, Nicolas Signolle, Loïc Le Bescond.

Merci du fond du cœur à Pablo & famille : Noemí, Rubén & Susana, Daniela, Clari, Alex, les tíes Carlos & Cristina et famille. Pablo : merci pour ta patience infinie, pour ton écoute et pour l'encouragement tout au long de ces années. Tu as démontré à chaque épreuve être quelqu'un d'exceptionnel. Rien n'aurait été possible sans ton soutien : c'est tellement plus facile quand le poids est divisé par deux…

Enfin, mes plus profonds remerciements à mes parents Liliana & Roberto et à ma sœur Nadia : vous savez combien vous comptez pour moi. Je n'ai pas les mots pour exprimer toute ma gratitude envers vous pour votre encouragement constant et votre amour.

A la mémoire de mon père, à son honnêteté, sa tempérance et son dévouement à sa famille, qui a toujours célébré même mes plus petits accomplissements, et qui m'a appris à ne jamais me laisser abattre.

Parti trop tôt cette année, il n'a pas vu la fin de cette thèse.

De tout mon cœur, je lui dédie ce travail.

Table of contents

List of figures

List of tables

Abbreviations

AFA	Alcohol-Formalin-Acetic acid (fixative)
AI	Artificial Intelligence
ANN	Artificial Neural Network
AR	Androgen Receptor
CNN	Convolutional Neural Network
CT	Chemotherapy
DFS	Disease Free Survival
DIA	Digital Image Analysis
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DP	Digital Pathology
ER	Estrogen Receptor
FDA	United States Food and Drug Administration
FFPE	Formalin-Fixed, Paraffin-Embedded
HER2	Human Epidermal growth factor Receptor 2
HES	Hematoxylin-Eosin-Saffron (staining)
HR	Hormone Receptors
IF	Immunofluorescence
IHC	Immunohistochemistry
KRT	Keratin
mIHC	Multiplexed Immunohistochemistry
MFI	Metastasis Free Interval
ML	Machine Learning
MIL	Multiple Instance Learning
OS	Overall Survival
pCR	Pathological Complete Response
PCR	Polymerase Chain Reaction
PR	Progesterone Receptor
RF	Random Forest
RFS	Relapse Free Survival
ROI	Region Of Interest
SVM	Support Vector Machines
TCGA	The Cancer Genome Atlas
TILs	Tumor-Infiltrating Lymphocytes
TMA	Tissue MicroArray
TNBC	Triple Negative Breast Cancer
TSA	Tyramide Signal Amplification
WSI	Whole Slide Image

Preface

Cancer: a villain with a long career

The word *καρκινοσ* (pronounced *karkinos* and meaning, in Greek, *crab*) was employed for the first time in the 4th century BC by the Hippocratic physicians to describe non-healing swellings or ulcerous formations whose projections seemed to reach out as the claws of a crab. These crab-like tumors were, in the words of Galen (131-203 AD), filled with – and caused by – a «black bile formed in the liver» (Papavramidou et al., 2010). At that time, the actual content of these lumps was not exactly known but, in the case of breast cancer, the fact that they were easily detectable by visible signs or palpation due to their superficial location has allowed the existence of depictions dating back many years: first breast cancer descriptions from 3500 years ago were found in medical papyri from Ancient Egypt (Brawanski, 2012; Helgason, 1987; Lukong, 2017). Its rarity in prehistory could be due to the low life expectancy, although there are also no descriptions of the multiple neoplasms that can affect young people (Salaverry, 2013).

Many years later, Galen used the term *oncos* (the Greek word for *swelling*) to describe tumor masses or edema, which would subsequently constitute one of the capital signs of inflammation. However, all these groundbreaking scientists did not think of cancer as a curable disease (Gill et al., 2015). Towards the end of the middle age, where advancement of medical science was still halted by religious reasons, the Galenic humoral theory encountered several opponents. Novel hypotheses, covering a broad range of unusual explanations for the breast cancer origin were formulated, such as chemical imbalance, coagulation of defective lymph, curdled milk, depressive mental disorders and even celibacy, to cite just a few examples.

Nevertheless, fresh ideas began to flourish. In 1757, Henri Le Dran proposed the surgical removal of tumor masses before they spread to the regional lymph nodes, introducing the concept of a disease that progresses in stages. But it wouldn't be until 1882 when radical mastectomy for breast cancer was introduced by William Halsted and, even if this procedure did not improve overall survival, a significant decrease in local recurrence was evidenced (from 56-81% reported at the time, to only 6%). Moreover, Thomas Beatson, in his publication of 1896, presented a new major milestone in the breast cancer treatment: the anti-hormonal

therapy (Rayter & Mansi, 2003).

Concurrently, the 19th century saw a landmark event being born: the cell theory, credited to the scientists Theodor Schwann, Matthias Jakob Schleiden and Rudolf Virchow. T. Schwann, a German physiologist, together with M. J. Schleiden, a German botanist, stated in his writings that all living organisms are composed of one or more cells and the products of cells in their structures. R. Virchow, a German biologist and doctor, contributed with his tenet in Latin: "*Omnis cellula e cellula*", which translates to "*All cells arise from pre-existing cells*" through the process of division. This precept contested the theory of spontaneous generation, still current at that time (https://www.biologyonline.com/dictionary/cell-theory).

Virchow associated for the first time the source of cancers with otherwise normal cells (Wagner, 1999). He hypothesized that carcinomatous cells could derive from the activation of quiescent cells in normal tissues, perhaps triggered by an important irritation in the tissues (https://www.cancer.org/content/dam/CRC/PDF/Public/6055.00.pdf). It would take until the end of the 20th century for this theory to gain relevance, when a strong link was highlighted between certain cancers and long-term inflammation (Balkwill & Mantovani, 2001; Coussens & Werb, 2002; Mantovani et al., 2008).

The rise of surgical resections in earlier times made way to the emergence of innovative diagnostic and treatment methods, with the mammography and the radiation therapy arising in the first half of the 20th century and major discoveries in the field of pharmacology in the second half, like the initial U.S. approval of the anti-estrogen drug Tamoxifen for hormone receptor-positive cancers in 1977, or the monoclonal antibody Trastuzumab, approved by the U.S. Food and Drug Administration (FDA) in 1998 and indicated in patients with HER2-positive tumors, as well as the emergence of chemotherapeutic compounds.

In these last two decades, substantial developments in many fronts as genetics, molecular biology and bioinformatics led to significant progress in the understanding of the cancer biology, which allows, in turn, developing more efficient and effective tools in cancer prevention, early detection and treatment.

Although the advances in this area occur exponentially, the global cancer burden remains high, with more than 19 million new cases worldwide in 2020 (Global Cancer Observatory; available from http://gco.iarc.fr/). Whereas the average age of the population

increases, there is a shift from infections towards common chronic conditions of the elderly, like vascular diseases, neurological disorders and cancer, and frequently linked to environmental factors and lifestyle. In Robbins words, « *there is no escape: it seems that everything people do to earn a livelihood, to subsist, or to enjoy life turns out to be (…) possibly carcinogenic.* » (Kumar et al., 2017). Population aging is likewise the cause of the augmentation in the absolute burden of disease and numbers of deaths by cancers. However, if treatment and care expansion is continued, the cancer mortality and morbidity will significantly fall in developed countries and, in a sort of trickle-down effect, less advantaged communities will benefit from this trend too, attaining a global improvement in public health around the world (Gill et al., 2015).

We are facing a new era where precision medicine will aim to offer tailored therapy for cancer. The emphasis of medical care is shifting from the management centered on a particular disease, affecting people in the same way, to focusing on the different manifestations of the same disease in each individual patient. In addition, a new model of practice is required, where the health professionals can work together with the patients to choose the best treatment and support options in every case, as it is starting to thrive in the personalized medicine approach.

Perhaps, in future decades, novel and promising advances to find safer compounds and to overcome drug resistance, such as identify and block the alternative pathways that lead to tumor escape, will be found. Coupled with enhanced interventions at the radiological and surgical level, they will be able to bring the cure to a greater number of afflicted individuals while enabling cancer to turn into a chronic disease concomitant with a good quality of life, easing the burden of what this diagnosis represents for patients today (Lukong, 2017).

Introduction

Breast cancer

A. Epidemiology

1. A major public health issue worldwide

Breast cancer is the first cause of cancer in women worldwide, with a global incidence of 2 261 419 cases in 2020 and responsible for 6.9% of all cancer deaths (both sexes, all ages) (Global Cancer Observatory; available from: http://gco.iarc.fr/) (Sung et al., 2021).

Despite the fact that the incidence rates had stabilized or decreased in several countries, breast cancer is still the most common neoplasm in women (Allemand et al., 2008; Pollán et al., 2010). In France, the number of new breast cancer cases in females at all ages was, in 2020, 58 083, being the most frequent malignancy in women and whose age-standardized incidence rate for both sexes exceeded the prostate cancer rate (99.1% versus 99%, respectively) (Global Cancer Observatory; available from: http://gco.iarc.fr/).

Philippe Autier et al., in their study of the mortality trends in breast cancer between 30 European countries, have uncovered that there was a median reduction in breast cancer mortality of ≥20% in 15 countries. In the U.K., it has been demonstrated that intervening in several factors ranging from smoking cessation support to improving detection through screening and developing more effective treatments – such as improved surgery, radiotherapy and drugs like tamoxifen and, more recently, anastrozole and letrozole –, the death rate showed a diminution of 38% since the start of 1980s (*Death Rates in Top Four Cancer Killers Fall by a Third over 20 Years*, 2014). In France, from 1989 to 2006, mortality decreased by 11%. The greatest reductions in mortality were observed in the group of women aged <50 years old (median 37%) (Autier et al., 2010). This improvement in survival has been most likely due to progress in the different treatment alternatives (medical, surgical, radiological) together with screening programs, and more accurate diagnostic and staging. According to all published data from European studies, as demonstrated by the EUROSCREEN Working Group, the reduction in breast cancer mortality associated with mammographic population-based service screening programs is in the range of 38–48% for women screened with sufficient follow-up time (Broeders et al., 2012).

In a report of the Munich Cancer Registry, overall survival (OS) improved among women

without metastases at diagnosis. Survival after development of metastases was not improved, but the anatomic sites involved varied over the years, shifting from bones to liver and central nervous system, probably due to the increasing use of systemic treatments (Hurk et al., 2011).

B. Morphological classification

Breast cancer is a clinically and molecularly heterogeneous disease, which covers a broad spectrum of entities, some of them with a particular approach and treatment. Classifications become of crucial importance to address questions such as which tumors will have an indolent course and which will grow rapidly, or which patients will require more aggressive treatments (Weigelt & Reis-Filho, 2009).

Invasive breast cancers are classified by pathologists into different subgroups regarding several clinicopathological variables as the histological grade and type and other features such as the presence of lymph-vascular invasion, lymph node involvement, and the expression of biomarkers by classical immunohistochemistry, mainly hormone receptors (HR) and HER2, with different prognostic and predictive implications (**Table 1**).

	MOLECULAR SUBTYPE				
	Luminal A	Luminal B		HER2 overexpression	Triple negative
		HER2 negative	HER2 positive		
% of all BC	50	15		15-30	12-17
Histological grade	Low	High		Intermediate to high	High
5-year survival (%)(*)	94,3	90,5		84	76,9
Metastasis	Bone	Bone, lungs		Bone, brain, liver, lungs	Lungs, brain
Particularities	Good prognosis	Aggressive clinical behavior			Younger women, afro-american, BRCA1 mutations
Estrogen receptor	Positive	Positive	Positive	Negative	Negative
Progesterone receptor	Positive	Low	Any	Negative	Negative
HER2	Negative	Negative	Positive	Positive	Negative
Ki67	Low	High	High	High	High

ER: estrogen receptor; HER2: human epidermal growth factor receptor 2; Ki67: proliferation index; PR: progesterone receptor

Table 1. Breast cancer molecular subtypes according to immunochemistry markers (modified from (Merino Bonilla et al., 2017)). (*) Cancer StatFacts: https://seer.cancer.gov/statfacts/html/breast-subtypes.html.

In order to inform about the tumor's aggressiveness, the Nottingham combined histologic grade (Elston-Ellis modification of Scarff-Bloom-Richardson grading system), also known as the Nottingham grading system, is the most widely used today, and it hoists independent prognostic significance (Rakha et al., 2008). It takes into account three

components: 1) the architecture, more precisely the tubule formation; 2) the nuclear pleomorphism; and 3) the mitotic rate (the proliferative activity) (Sinn & Kreipe, 2013; Weigelt & Reis-Filho, 2009). Each of these categories are scored 1-3 and the sum of these individual values, which varies between 3 and 9, will correlate with a grade from I for well-differentiated to III for poorly differentiated (Elston & Ellis, 1991).

The latest edition of the World Health Organization (WHO) classification of breast cancers (5th edition, published in 2019) recognizes as many as 44 distinct histological subtypes, which provide important information for clinical management (Cserni, 2020). The greater part is constituted by invasive carcinomas of no special type (NST) (formerly named invasive ductal carcinoma, not otherwise specified [IDC-NOS] before the 2012 edition of the WHO classification). This entity represents up to 75% of all breast cancers and constitutes a diagnosis of exclusion (Reis-Filho & Lakhani, 2008).

Conversely, special type carcinomas, which account for up to 25% of all breast cancers, harbor a particular histological pattern in more than 90% of their mass. With the exception of invasive lobular carcinomas (the most frequent in the special type group) and apocrine cancers, and in contrast with the IDC-NSTs, these special types are very homogeneous, pertaining to only one molecular subtype (Weigelt et al., 2008, 2009). It is extremely important to recognize and report these histological subtypes since each one carries a particular prognosis and they are associated, for the majority of them, with an improved survival compared to IDC-NST, even when it comes to mixed tumors merging more than one morphological variant in the same lesion (Ellis et al., 1992). Mixed morphologies were defined as the occurrence of special types in 10-90% of the tumor, admixed with IDC-NST.

ER was the most important biomarker associated to breast cancer for years, because of its significance in treatment, together with PR, also measured in tissue samples and which provides clues for prognosis. Additionally, with the PR gene dependent on ER, its expression might indicate that the ER pathway is intact. Both proteins are routinely assessed by IHC on formalin-fixed, paraffin-embedded (FFPE) tissue slides, and the cut point to distinguish "positive" from "negative" cases for ER and PR is 1%, this means that in patients whose tumors show at least 1% ER-positive cells, hormone therapy should be considered. In France, the established cut-off is 10% (Cornish, 2020). Equally, HER2, an oncogene that encodes a transmembrane glycoprotein with tyrosine kinase activity and the third main actor among the

breast cancer biomarkers, is evaluated in order to refine prognosis and individualize the patients who could benefit from therapies targeting this protein (Hammond et al., 2010; Slamon et al., 1987). The screening test is performed by IHC and equivocal cases (score 2+) are confirmed by in situ hybridization (Wolff et al., 2018).

Putting aside the HR-positive tumors and the HER2-positive tumors, what is left is a hodgepodge with a "triple-negative breast cancer" (TNBC) tag encompassing all tumors that lack the expression of ER, PR and HER2. These tumors share the absence of recognized molecular targets for therapy, and therefore a poor prognosis (Chacón & Costanzo, 2010). They also generally have a larger size, with no correlation between tumor size and node status, they have a special pattern of metastasis with a propensity for visceral involvement and harbor a more aggressive biological behavior (Dent et al., 2007).

In the conventional practice, this wide diagnosis of exclusion, which affects 12 to 17% of women afflicted with breast cancer, emerges from the accurate assessment of the status of estrogen receptor (ER), progesterone receptor (PR) and HER2 status. The same phenotype characterizes the basal-like cancers, along with the positivity for keratin-5 (Foulkes et al., 2010).

Of note, is it worth mentioning here that some rare histological subtypes harboring a TNBC phenotype are associated to a very good prognosis, such as adenoid cystic, apocrine and secretory carcinomas.

C. Molecular classification

1. Perou and the molecular portraits revolution

Tumor classification is a dynamic process, combining multiple sources of information. Traditionally, TNM staging system, published by the American Joint Committee on Cancer (AJCC) and based on tumor size, regional lymph node involvement and metastasis, has been reported as a measure of the extent of cancer, to establish a clinical stage (before treatment) and a pathological stage (after treatment, i.e. surgery), and to give advice on the prognosis (**Figure 1**).

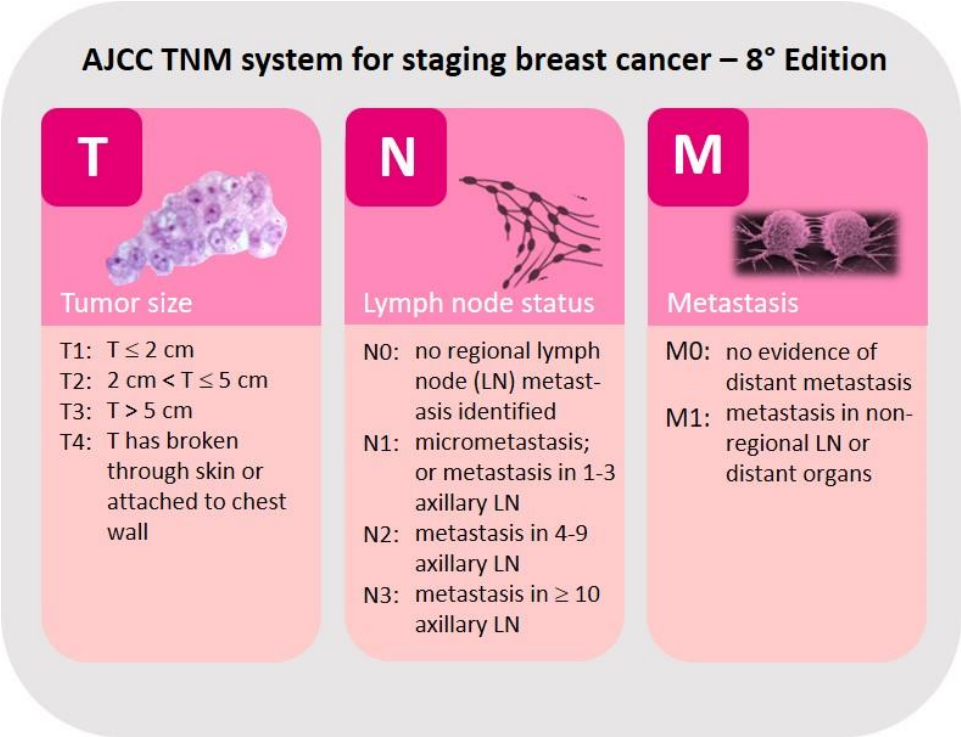


Figure 1. TNM staging system for breast cancer (adapted from AJCC Cancer Staging Form Supplement (last updated 7 January 2021)).

The advent of genomic techniques enabled to relate the different biological behaviors of morphological groups with tumor gene expression profiles. This had led to the development of a new and more accurate classification to complete and improve the anatomical criteria informed in the TNM. The 8th and latest version of this staging system, effective since January 1st, 2018, integrated biomarkers status and multigene panel status, switching from morphological to molecular features, and entailed a major change in the study of these entities. The interest of these recent changes is not only ontological but also the contribution in terms of practicality and applicability to the stratification of patients who could benefit from a tailored therapy (Koh & Kim, 2019). What was historically thought as a single disease with several variants, emerged in fact as distinct disorders likely to be discriminated by its molecular predominant features, as ER expression. As Perou et al. remarked on their revered gene-expression profiling study, ER-positive and ER-negative breast cancers are indeed distinct diseases, and same conclusions could be stated as the characterization efforts evolve. The identification of novel groups might be related to different molecular features of mammary epithelial biology, namely ER+/luminal-like, basal-like, HER2-enriched and normal breast-like, in accordance with prognosis and responses to treatment (Dent et al., 2007; Perou et al., 2000) (**Table 2**).

	HIGH HR EXPRESSION		LOW HR EXPRESSION		
	ER+		BASAL-LIKE	ERBB2+	NORMAL BREAST
	LUMINAL A	LUMINAL B			
HIGHLY EXPRESSED GENES	BREAST LUMINAL EPITHELIAL CELLS		BREAST BASAL EPITHELIAL CELLS	ERBB2 ONCOGENE	NORMAL BREAST GENE EXPRESSION PATTERN (BASAL EPITHELIAL CELLS AND ADIPOSE CELLS)
IMMUNOHISTO-CHEMISTRY	KERATINS 8/18		KERATINS 5/6, 14, 17, EGFR	HER2	
Biomarker status	ER+ HER2- PR+ KI67-	ER+ HER2+/- PR+ KI67+	ER- HER2- PR-	ER- HER2+ PR-	ER+ HER2- PR+ KI67-
GRADE	1/2	2/3	3	2/3	1/2/3
OUTCOME	Good	Intermediate/ Poor	Poor	Poor	Intermediate

ER: estrogen receptor; HR: hormone receptor; PR: progesterone receptor

Table 2. Tumor gene expression profiles (Dai et al., 2015; Perou et al., 2000).

According to the research conducted by Sørlie et al., based on the study of expression patterns of 540 stably expressed genes (« intrinsic genes »), tumors were classified into five intrinsic subtypes: luminal A, luminal B, HER2 over expression, basal and normal-like, setting the standard for new classifications. The first main divergence criterion was the expression of ESR1 gene, coding for ER and other genes representative of luminal epithelial cells. When present, variation in this expression separated luminal tumors into two groups: the largest pool, termed luminal subtype A, held the highest ER cluster expression; the smaller group, or luminal subtype B, showed low to moderate expression of the genes mentioned above and high expression of a different set of genes of unknown function. Tumors negative for genes from the luminal/ER cluster were characterized by the presence of epithelial basal cell keratins such as KRT5, KRT14 and KRT17, and thus named basal-like subtype. A HER2-enriched subtype was identified from high expression of the genes in the ERBB2 amplicon, including HER2. Finally, normal breast tissue-like group exhibited an expression profile including genes of non-epithelial cells like adipocytes and others (Sørlie, 2004; Sørlie et al., 2003). This class may correspond actually to an artefact related to the transcriptomic analysis technique, due to an overproportion of benign cells over carcinomatous cells in the tumor samples analyzed. Furthermore, it has been demonstrated that the risk and timing of disease recurrence also depends on tumor characteristics, such as the growth rate through the analysis of Ki67 expression and the molecular subtypes (Ribelles et al., 2013). In a retrospective study performed by Shim et al., relapse rates were ranked by subtype, with luminal A tumors having

the lowest (5.02%) and triple negative tumors having the highest risk of recurrence (16.76%) (Shim et al., 2014). The same trend was found by Sørlie regarding the overall survival in different subgroups, where luminal B tumors turned out to be a particular clinical group with a worse outcome due to relapse, and basal-like and HER2-positive subtypes correlated with the shortest survival time (Sørlie, 2004).

To illustrate the complexity achieved by classifications, in the TNBC group, 6 different subtypes have been described by Lehmann et al. on the basis of gene expression profiles: basal-like 1 (BL1) enriched in cell cycle components; basal-like 2 (BL2) of probable basal/myoepithelial origin; immunomodulatory (IM) enriched for genes related to immune cell processes; mesenchymal (M); mesenchymal stem–like (MSL); both linked to cell motility and ECM interaction; and luminal androgen receptor (LAR) (Lehmann et al., 2011).

Classifications become even more complex. In breast cancer, cell type and steroid receptor signaling are deeply related. Through the study of gene expression microarrays, Farmer et al. identified a third class of tumors, together with basal and luminal subtypes, characterized by apocrine features and positivity for the androgen receptor (AR). Based on this data, a classification regarding the steroid receptor activity could be argued: luminal (ER and AR positive), basal (ER and AR negative) and molecular apocrine (ER negative and AR positive). Continuous improvement in these categories could have a positive impact on treatment through the identification of novel drug targets and adjustments in actual endocrine therapy (Farmer et al., 2005).

A different insight of classifications was presented by Curtis et al. on the basis of an integrated analysis taking account of genomic and transcriptomic data. Ten novel subgroups with particular clinical outcomes were defined by cluster analysis stratifying tumors in 10 integrative clusters (IntClust). Once more, breast cancer heterogeneity, especially of TNBC, was brought to light in view of the basal-like tumors distributed among different groups (Curtis et al., 2012).

Notwithstanding the efforts of the different study groups trying to find the ideal stratification for breast cancer entities, the current classification could be simplified to three basic subtypes in order to come to a deeper understanding of these categories, i.e., luminal, HER2 over-expression and triple negative tumors subtypes (Prat & Perou, 2011; Reis-Filho & Lakhani, 2008).

Luminal tumors harbor a profile including HR expression along with luminal keratins 8 and 18, accounting for 70% of all the breast tumors. Taking HER2 expression as a criterion, an approximate differentiation in three subgroups can be made: 1) luminal A tumors (negative for HER2) with a higher expression of ER-related genes, 2) luminal B tumors positive for HER2, which have a tendency to be of higher grade and, therefore, of poorer prognosis, and 3) luminal B tumors negative for HER2, that display a great proliferation capacity, determined by the Ki67 score. This biomarker is usually assessed by manual counting of positive nuclei on IHC stained slides, but is subject to certain limitations such as the lack of consensus regarding the cut-off for proliferation rate and the intra and interobserver variability.

HER2 over-expressed tumors frequently exhibit TP53 mutations and a histological high grade. Although these tumors are sensitive to neo-adjuvant chemotherapy and a targeted anti-HER2 treatment is available, they harbor a higher risk of relapse in the absence of complete pathological response. This, added to the fact that not all HER2 over-expression tumors respond to trastuzumab, confers this subtype a poor prognosis.

Basal tumors are (generally) negative for HR and HER2, and display high expression of keratins 5, 6, 14, 17, EGFR (epidermal growth factor receptor) and proliferation genes, also a high frequency of TP53 mutations. They are likely to be of high grade and larger size, affect younger patients and harbor an aggressive clinical course with a high risk of relapse (Dai et al., 2015; Rakha et al., 2008).

Molecular subtypes were introduced at the St Gallen 2011 conference (Goldhirsch et al., 2011), and slightly modified at the St Gallen 2013 meeting (Goldhirsch et al., 2013), in order to find the way to transpose them into a hybrid morphological / immunohistochemical classification applicable in routine clinical practice and used as orientation in decision-making regarding treatment choices.

The proposition that came out of this conference was a tumor classification according to immunohistochemical criteria with an adapted treatment for each immunohistochemical class (**Table 3**).

Luminal A		ER and PR-positive
		HER2-negative
		Ki67 low (<20%)
		Low risk of recurrence according to genomic tests
Luminal B	HER2-negative	ER-positive
		HER2-negative
		At least one of the following:
		Ki67 high (at least 20%)
		PR-negative
		High risk of recurrence according to genomic tests
	HER2-positive	ER-positive
		HER2-positive or amplified
		Any Ki67 status
		Any PR status
HER2-positive		ER and PR-negative
		HER2-positive or amplified
TNBC		ER and PR-negative
		HER2-negative

ER: estrogen receptor; HER2: human epidermal growth factor receptor 2; Ki67: proliferation index; PR: progesterone receptor; TNBC: triple-negative breast cancer.

Table 3. Molecular classification according to St Gallen 2013 (adapted from (Goldhirsch et al., 2013)).

It is important to recall that, even if it is known that molecular analysis will provide major insights into the classification improvement and refinement, the value of morphological assessment, relatively simple, inexpensive and not very time-consuming, cannot be underestimated. Often the identification of certain features on the histological samples can guide the reasoning, as it happens, for example, with the presence of metaplastic elements which are tightly linked to a basal-like phenotype. As Weigelt & al. rightly pointed out, there is a genotypic-phenotypic correlation between morphological patterns and molecular traits despite the molecular subtype (Weigelt et al., 2009). This is the case of basal-like ductal carcinomas and metaplastic cancer, which, although sharing a molecular subtype, are genetically different. This discrepancy could explain the different sensitivities of these entities to treatments as chemotherapy. To cite another example, the medullary and adenoid cystic tumors have an excellent prognosis even though they belong, according to their gene expression profiles, to the basal-like subtype, more frequently associated with a poor outcome

(Weigelt et al., 2008; Weigelt & Reis-Filho, 2009).

Nevertheless, sometimes a special histological type constitutes a distinct pathological entity. Such is the case of micropapillary carcinomas, with peculiar histological features, a recurrent luminal-B phenotype and significant association with genetic aberrations shared with high-grade tumors. These findings suggest that the identification of micropapillary carcinomas is not just a matter of terminology, since it denotes a whole separate disease with a poorer prognosis than IDC-NSTs (Marchiò et al., 2008).

2. **Molecular signatures**

Advances in the field of tumor genotyping have been of practical use in treatment decision-making. Gene-expression profiling is a relevant means in personalized medicine: it has the power to enhance the accuracy of the clinicopathological risk assessment in predicting cancer prognosis, aiming to better stratify breast cancer patients for tailored treatment, including therapeutic de-escalation (Buyse et al., 2006). The analysis of a set of genes that compose a « molecular signature », when performed on cancer cells, can help predict which patients will most likely benefit from adjuvant chemotherapy to reduce the risk of relapse and, equally important, which patients might not need it, avoiding unnecessary heavy treatments and the resultant adverse side effects. It is for such patients that it seems essential to identify predictive factors of treatment efficacy. But this risk distinction is not always clear, principally when it comes to early stage cancers or low-risk tumors, where the maximum usefulness of these signatures is revealed. Additionally, genes issued from poor prognosis tumor profiling could function as a rationale for the development of new tailored therapeutics (van 't Veer et al., 2002).

Examples of these commercially available multigene panels, the so-called pure signatures and the most widely used multigene assays worldwide, are *Oncotype DX®- Recurrence Score®* (RS) (Genomic Health) (Paik et al., 2004) and *Mammaprint® Netherland Kanker Institute 70-gene signature* (Agendia BV) (Cardoso et al., 2016; S. Tian et al., 2010; van de Vijver et al., 2002). These two assays are performed in centralized laboratories (in USA for Oncotype DX® and in Netherlands for MammaPrint®) with results available in 8-10 days. Combined signatures (panel of genes associated to clinico-pathological characteristics) include *EndoPredict®* (EPclin) (Myriad Genetics Salt Lake City, UT, US) (Fitzal et al., 2015) and *Prosigna® risk of recurrence (ROR)* (formerly denominated PAM50 test) (NanoString Technologies, Seattle,

WA, US) (Gnant et al., 2014; T. Nielsen et al., 2014; T. O. Nielsen et al., 2010; Parker et al., 2009) (van de Vijver et al., 2002). Other signatures also used are *Breast Cancer Index (BCI)* (BioTheranostics, San Diego, CA, USA) (Ma et al., 2008, p.) and HalioDX (Ignatiadis et al., 2016). All signatures are suitable for the individualization of low-risk tumors that might not need chemotherapy. The two combined tests (EndoPredict® and Prosigna®) are decentralized and can be executed in any pathology laboratory, and can also identify which patients would profit of more than 5 years of endocrine therapy.

All assays except MammaPrint® were designed for ER-positive breast cancer patients. The ROR and EPclin signatures integrate clinical parameters, such as tumor size and node involvement, the latter being the most significant prognostic clinical indicator in early-stage breast cancer. It has been demonstrated that it is useful to predict response to pre-operative neoadjuvant therapy (Dubsky, 2020). Low-score breast tumors were unlikely to respond to neoadjuvant CT, tumor response attained 27% when treated with neo-adjuvant endocrine therapy (NET), while breast tumors with high scores were resistant to neoadjuvant endocrine therapy and, in general, strongly linked to a poor tumor response.

Some of these signatures have been validated in the prospective setting. MammaPrint® has been included in the MINDACT clinical trial, which assessed the CT benefit in groups of different clinical and genomic prognosis, the latter estimated with this molecular signature (Cardoso et al., 2016). Oncotype DX® utility was proven by two clinical trials: TAILORx, which confirmed the validity of the RS to identify the patients that could benefit of endocrine therapy only (Sparano et al., 2015), and RxPONDER, a phase 3, randomized clinical trial to assess the significance of CT for patients with N1 and a low/intermediate RS (Kalinsky et al., 2021).

More details of the main molecular signatures can be found in **Table 4**.

	70-gene signature Mammaprint® (Agendia, Netherlands)	21-gene signature Oncotype DX® (Genomic Health, US)	PAM 50 Prosigna (Nanostring, US)	Genomic Grade MapQuant DXtm (Ipsogen/HalioDX, France)	HOXB13: IL17BR BCI (Biotheranostics, US)	11-gene assay Endopredict® (Myriad Genetics, US)
Method	*Microarray*	*qRT-PCR*	*n-counter*	*Microarray qRT-PCR*	*qRT-PCR*	*qRT-PCR*
Material	Cryo/FFPE	FFPE	FFPE	Cryo/FFPE	FFPE	FFPE
Analyzed data	70 genes	Genes: ER, PR, BCL2, SCUBE2, Ki67, STK15, BIRC5, CCNB1, MYBL2, HER2, GRB7, MMP11, CTSL2, GSTM1, CD68, BAG1	50 genes and pathological criteria (tumor size and node status)	97 genes	Genes: HOXB13, IL17BR, BUB1, CENPA, NEK2, RACGAP1, RRM2	Genes: DHCR7, AZGP1, MGP, STC2, BIRC5, UBE2C, RBBP8, IL6ST and pathological criteria (tumor size and node status)
Prognostic value	**Recurrence** (5 years)	**Recurrence** (10 years)	**Recurrence** (10 years)	**Recurrence**	**Recurrence** (5 and 10 years)	**Recurrence** (10 years)
Indications	ER+/N– or N+ (1-3) ER–/N– or N+ (1-3)	ER+/HER2– /N–/ET ER+/HER2–/N+ (1-3)	HR+/HER2– N– or N+	ER+/N– (grade 2) under tamoxifen	ER+/N– under tamoxifen	ER+/HER2– N– or N+ (1-3) under ET
Results	 High Low	RS = 0 to 100 High > 30 Intermediate Low < 18	Molecular type ROR = 0 to 100 High Intermediate Low	 High Equivocal Low	0 to 10 High Intermediate Low	0 to 15 High Low
Prospective assay	MINDACT LESS	TAILORx RxPONDER	OPTIMA UK	ASTER 70s		UNIRAD

qRT-PCR: real-time quantitative reverse transcription-polymerase chain reaction; **Cryo:** cryopreservation; **FFPE:** formalin fixed-paraffin embedded; **ER (+ or –):** estrogen receptor status; **N (+ or –):** lymph node status; **HER2:** human epidenrmal growth factor receptor 2; **ET:** endocrine therapy; **HR (+ or –):** hormone receptor status; **RS:** recurrence score ; **ROR:** risk of recurrence score; **CT:** chemotherapy; **MINDACT:** Microarray In Node-negative and 1-3 positive lymph-node Disease may Avoid ChemoTherapy; **TAILORx:** Trial Assigning Individualized Options for Treatment Rx; **RxPONDER:** Rx for Positive Node, Endocrine Responsive Breast Cancer; **OPTIMA:** Optimal Personalised Treatment of early breast cancer usIng Multi-parameter Analysis; **ASTER 70s:** Adjuvant Systemic Treatment for (ER)-Positive HER2-negative Breast Carcinoma in Women over 70 According to Genomic Grade.

Table 4. Summary of the main prognostic molecular signatures used in early invasive breast cancer (adapted from (Joyon et al., 2017; Naito & Urasaki, 2018)).

Gene expression among breast cancer subtypes, as shown by the examination of molecular signatures, is not random. Interestingly, and despite the few genes in common among these panels, the performances are comparable, that is to say, they pinpoint the same cluster of poor prognosis patients (Fan et al., 2006; Sestak et al., 2018). Proliferation-related genes, which appeared to lead the signatures' performance, turned out to be one of the strongest prognostic factors in patients with ER-positive disease. Their expression in luminal tumors was quite heterogeneous, with luminal A tumors showing a low expression of proliferation-related genes, which explains the fact that signatures are helpful in predicting the risk of relapse in ER-positive tumors. ER-negative tumors, comprising both triple-negative and HER2-positive tumors, displayed a high expression of proliferation-related genes and are thus classified as high-risk of recurrence by all tests (Reis-Filho & Pusztai, 2011; Sotiriou & Pusztai, 2009). Relevant information has been extracted from a meta-analysis performed by Desmedt et al. on microarray studies, respecting the clinical outcome in HER2-positive and basal-like subgroups. Proliferation was identified as a major predictor in prognosis, along with histological grade, in ER-positive tumors. However, immune response and tumor invasion appeared to be the leading molecular processes linked to survival in triple-negative and HER2-

positive subgroups, respectively (Desmedt et al., 2008).

As it follows from the above considerations, one of the limitations of molecular signatures is that they are not useful for TNBC or HER2-enriched because these high-grade tumors always fall in the poor prognosis category. Another important restrains are their very expensive cost and their unavailability in less developed laboratories around the world.

More accessible prediction tools are "*Adjuvant!*" software, a web-based prognostication and treatment benefit method for breast cancer, which uses an algorithm that combines prognostic factors for treatment decision making, by calculating the reduction in the risk of relapse (Olivotto et al., 2005), *Predict Breast* (https://breast.predict.nhs.uk/) (Wishart et al., 2010, 2011, 2012) and *CTS* (Clinical Treatment Score) (Sestak et al., 2018). While the first was abandoned in 2011, the other two are still in use in clinical practice. Predict Breast is a prognostication model that uses registry data to predict overall and breast cancer specific survival in treated early breast cancer patients. CTS is a prognostic algorithm developed on the Translational Study of Anastrozole or Tamoxifen Alone or Combined (TransATAC) cohort. It includes clinicopathologic information about nodal status, tumor size, grade, age, and treatment, and the 4-marker immunohistochemical score (IHC4) (which combines prognostic information of 4 widely used IHC markers) to estimate distant recurrence for 0 to 10 years and 5 to 10 years after diagnosis (Buus et al., 2016; Cuzick et al., 2011; Dowsett et al., 2013; Sgroi et al., 2013).

3. **Decrypting the outcome of cancer (predictive and prognostic factors)**

The stratification of patients into different outcome classes, according to patient and tumor characteristics, allows the estimation of individual outcome prediction, constructed on the basis of validated prognostic and predictive markers. These markers can be clinical, morphological, biological or molecular, and some factors are both prognostic and predictive. As clearly defined by Rakha, a *prognostic factor* is any characteristic that is predictive of the patient's outcome unrelated to systemic therapy, while a *predictive factor* is a feature that correlates with response or lack of response to a specific treatment. Thus, prognostic markers, related to intrinsic tumor characteristics (such as tumor growth, invasion capacity and metastatic potential) aid to decide whether a patient should be treated with adjuvant chemotherapy and with which treatment regimen, while the choice among the different treatment options is assisted by the predictive markers. Up to date, the three most important

prognostic factors remain lymph node invasion, tumor size and histological grade (Fitzgibbons et al., 2000; Galea et al., 1992; Rakha, 2013; Rakha et al., 2008).

Strategies to better define breast cancer, including a complete characterization of the tumor microenvironment and the actors intervening in cancer development and progression, like tumor epithelial, myoepithelial, and stromal cells, are ongoing challenges which endeavor to find a better approach for patients. Currently treatment options for early breast cancer can be seen in **Figure 2**.

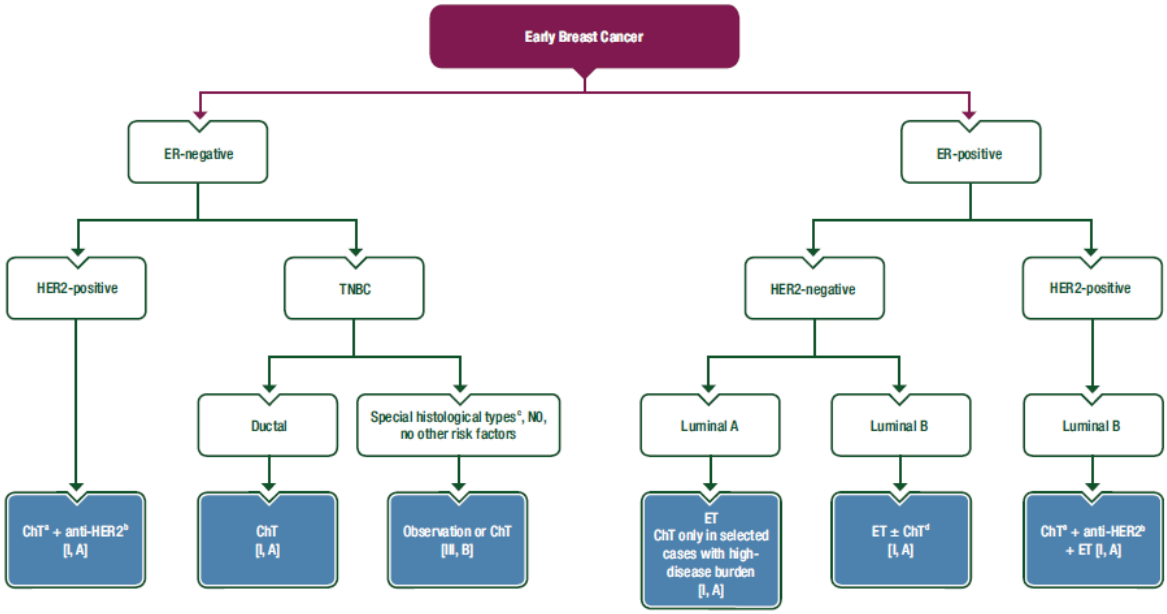


Figure 2. (Neo)-adjuvant systemic treatment choice by marker expression and intrinsic phenotype. **(A)** With possible exception of selected cases with very low risk T1abN0. **(B)** Anti-HER2: trastuzumab **+/-** pertuzumab. **(C)** Adenoid cystic or apocrine, secretory carcinoma, low-grade metaplastic carcinoma. **(D)** Depending on level of ER and PgR expression, proliferation, genomically assessed risk, tumour burden and/or patient preference. **(E)** Except for very low-risk patients T1abN0 for whom ET/anti-HER2 therapy alone can be considered. ChT, chemotherapy; ER, estrogen receptor; ET, endocrine therapy; HER2, human epidermal growth factor receptor 2; N0, node-negative; PgR, progesterone receptor; TNBC, triple-negative breast cancer (from (Cardoso et al., 2019)).

One of the most important future challenges will probably be to facilitate the transition between traditional healthcare and personalized medicine, and to guarantee an extended access to the system with an emphasis on early breast cancer cases to find the best treatment strategy, avoiding toxic effect of unnecessary therapies improving patients' outcomes (Andre et al., 2010; Collins & Varmus, 2015; Torti & Trusolino, 2011).

Pathology: the great digital transformation

A. At the crossroads of morphology, molecular diagnosis, personalized treatment and computer sciences

It is known that diseases have an anatomical substrate. Since its origins as a medical discipline, intermingled with other branches of medicine still emerging, pathology put its efforts into unraveling the causes behind the injuries that afflicted living beings. This achievement started to bear fruit through the knowledge both from anatomy, with the practice of dissections and autopsies, and from biology, looking for alterations at the cellular level that could generate each clinical picture. The concept of organ-based disease had arisen.

The relationship that unites pathology with technological development is very close and long-standing, and maybe the most outstanding milestone to illustrate it is the way the microscope changed the practice of this specialty from the mid-nineteenth century onwards, when this instrument became more available, efficient and at a decreasing cost.

Through the emergence of microscopy and the committed work of figures like Virchow, Morgagni and Bichat in the correlation of histological and cytological findings with specific physio-pathological processes, histopathology became more and more important, until becoming an individual domain among specialties (Chan & Salto-Tellez, 2012). Modern practice changed progressively with the emergence of technical advances as the fixation, embedding and staining procedures for tissues study. It changed at an even more accelerated pace since early days of last century with immunohistochemical staining, molecular methods, genomics and, more recently, image processing and analysis by computational approaches (**Figure 3**).

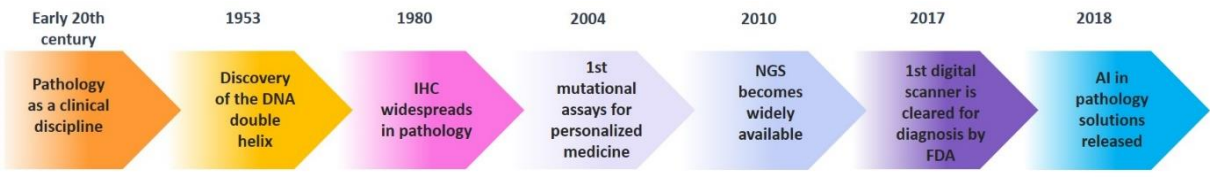


Figure 3. Chronology of transforming milestones in pathology (adapted from (Salto-Tellez et al., 2019, p.)).

The long road of pathology moves towards a greater degree of detail, dominated these last years by the switch from cell-based to gene-based causes of disease, to understand them and explain them at a molecular level (van den Tweel & Taylor, 2010). This conversion from the general to the particular is transforming medicine at the individual level, with the rising trend of personalized treatment. Medical care has fluctuated from a generic approach of the diseases to the subclassification of patients into smaller and more homogeneous groups based on distinctive biomolecular features of both the patient and his/her disease. The role of the pathologist, with increasing demands on diagnostics, is key in providing the scientific evidence necessary to characterize tumors and enhance treatment.

As always, this transformation is boosted by technological advances. First, current molecular diagnoses are mostly based on traditional or upgraded versions of immunohistochemistry. Second, computational science and information technology brought a fresh approach that is changing the specialty once again through big data management and digital imaging systems, which are slowly replacing the traditional use of microscopes in histopathology. A workflow in a modern pathology laboratory could be schematized as in **Figure 4**.

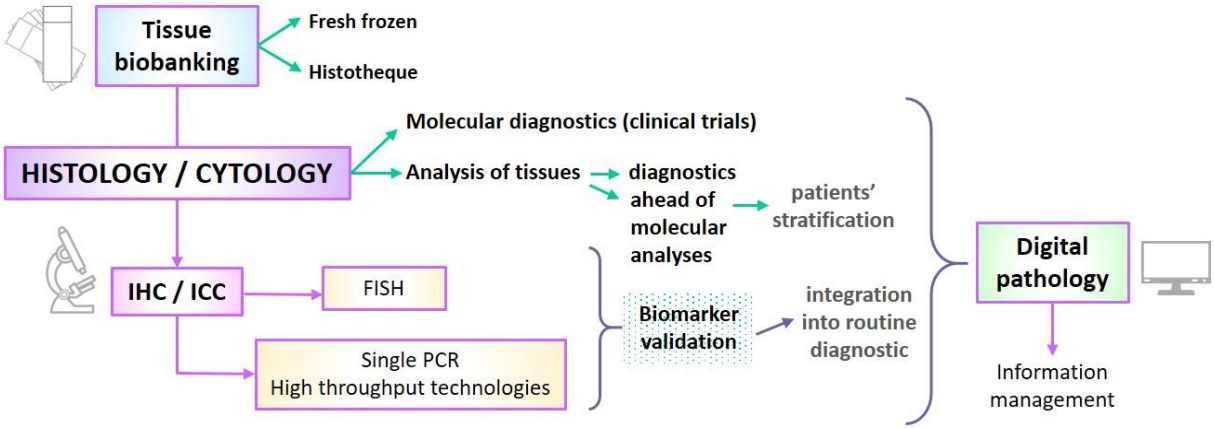


Figure 4. Tissue pipeline. Representation of the pipeline that a tissue sample would take, passing through several traditional and molecular pathological tasks until deliverance of optimal results. Digital pathology can be applied at the different steps of the path from management of databases to tools that can be integrated in routine diagnostic (adapted from (Salto-Tellez et al., 2014)).

The first virtual microscope, developed in Baltimore in 1997, emerged from a collaboration between the Department of Computer Science from the University of Maryland and the Department of Pathology from the Johns Hopkins Medical Institution. This instrument, described as "a realistic emulation of a high-power light microscope", has surpassed today the

possibilities of light microscopy. At that time, while the hardware was beginning to be available in a relatively easy way, there was a lack of software that would allow the new technology to work in the same way that a traditional light microscope did. The main problem was the extremely large quantities of data and the need of both a powerful compression and a minimal loss of information. Just to make a cursory comparison, at that time, an image occupied 35-210 GB of raw data per slide versus 2-3 GB occupied by an image today (Ferreira et al., 1997).

With the expansion of the volume and data complexity of cancer specimens, pathology laboratories face the challenge of an increasing workload, not only in the morphological setting but also in the "special techniques" field. The bundle of immunohistochemical and molecular tests required for diagnosis, prognosis or therapeutic decision-making is continuously escalating, and it demands new working modalities to absorb this extra burden. Digital pathology (DP), which is becoming broadly available, may enable the implementation of artificial intelligence in the form of computer learning tools to boost certain tasks, such as tumors classification, that will ultimately accelerate the administration of appropriate therapies and accurate prognostication. Eventually DP, through telepathology facilities, could also allow the flexible use of the pathologist expertise from different locations, with centralized slide production and dispersal of diagnostic pathologists across or between regions (P. H. Tan et al., 2020; Williams & Treanor, 2020).

B. Digital pathology and image analysis

1. From the conventional microscope to the computer screen

A brief introduction to digital pathology

Personalized medicine started to open spaces through the traditional health care workflow and changed the way in which diagnoses are made. Several constraints are still encountered, sometimes related to the increasing demand, such as the lack of pathologists in remote locations or the lack of training in special determinations or techniques required, or related to insufficiently equipped laboratories. Other times, the concerns are linked to the employed methods for assessment, such as intra and interobserver variability.

Modern pathology has found the answer to some of these issues in the implementation of a digital workflow. The term *digital pathology* encompasses all associated technologies that exploit digital images to enable improvements and innovations in current practice, while *digital*

microscopy, another commonly used term, refers only to high resolution scanning of histological slides and their subsequent storage.

Digital pathology involves the scanning of glass slides containing pathological specimens into digital image files with purposes of interpretation, automated analysis and archiving (Jara-Lazaro et al., 2010). The combination of digitized pathology data with image processing techniques is a real breakthrough both in diagnostics and research, in particular after the introduction of artificial intelligence and its diverse subfields specifically adapted to image classification through pattern recognition (see next chapter). Digital methods provide efficient tools that allow the pathologists to choose when it is mandatory to deliver results in a highly quantitative manner (i.e. biomarkers scoring) to increase its reproducibility and to assist clinical decision making (Robertson et al., 2018).

It could be said that the functional unit of digital pathology is the so-called *whole slide image* (WSI) or virtual or digital slide. This term refers to the full image of a histological section captured with a scanning device and converted into a high-resolution digital slide that can be viewed, managed, shared and analyzed on a computer screen/system. The slide scanner is composed of four central elements: a light source, a motorized slide stage, objective lenses, and a high-resolution camera for image capture. Two characteristics are essential for slide scanners: the speed of scanning (to obtain images in times compatible with clinical diagnosis) and the resolution and quality of the images that are obtained, as they are the basis of the analysis by pathologists. The balance between acquisition speed and resolution will be dictated by the type of activity for which the scanner will be used (research projects, intra-operative consultations, etc.) (Frenois, 2019; Zarella et al., 2018).

The acquisition time varies depending on the tissue surface, the used magnification and the scanner type but, in general, it requires between less than a minute (with a high-speed scanner) and 20 minutes for a complete slide at a high magnification. The product is a single high-resolution digital file of between 200 MB (Megabytes) and 5 GB (Gigabytes) always depending on the tissue size and the chosen magnification. Special attention must be paid to the quality of WSI to prevent subsequent problems when performing image analysis. Image artifacts such as tissue folds, blurred zones and pen marks need to be eliminated or corrected, as well as batch effects, introduced at the moment of the staining, derived of the presence of sub-cohorts prepared with different colorations and digitizing devices, or when several batches

are required to analyze large clinical series (Heeke et al., 2019; Kothari et al., 2013).

Routinely, once scanned, quality of WSI is first checked to assess if they are exploitable or not, and then the insufficiently good WSI have to be rescanned. These additional tasks entail a double loss of time in the workflow. Novel methods have been developed to overcome this matter, allowing the automatic assessment of WSI while the scanning process is in progress, or later as a quality control tool (Ameisen et al., 2013, 2014).

Most of the digital microscopy systems use similar functioning principles. They are basically composed of an optical microscope system (motorized microscope or scanner), an acquisition system with a software to control the scan procedure, and display devices coupled to a digital slide viewer. These devices work with conventional slides (25 x 75 x 1 mm) which are digitized in different focused planes through the z-axis to emulate the fine focus of a conventional microscope (see below) (Rojo et al., 2006). Besides the listed above, extra elements may be required to implement a digital pathology structure, such as image storage systems, data sharing systems for the transmission of images and more complex image analysis software.

But… how does it work?

The process starts when a proper slide is placed in the scanner and an overview of the whole slide is acquired that will serve to localize the sample. Subsequently, the region of interest is manually or automatically delimited, and the image settings (acquisition mode, magnification) and focus points are adjusted. Focus points, manually or automatically introduced, are selected areas in the image, where the optical system will adjust itself to obtain a maximal sharpness. Then, during the scan stage, the virtual slide is created through the sequential acquisition of square or rectangular microscopic fields – *patches* – that are later stitched into a unique, seamless image. The software then generates the intermediate magnifications from the most detailed image using subsampling algorithms, which allow the visualization of the full WSI at all possibly magnifications. The product is an image termed *pyramidal* (**Figure 5**), where every point is designated by coordinates for both the location in the image (width, length and height; X, Y and Z – see below – respectively) and magnification (G).

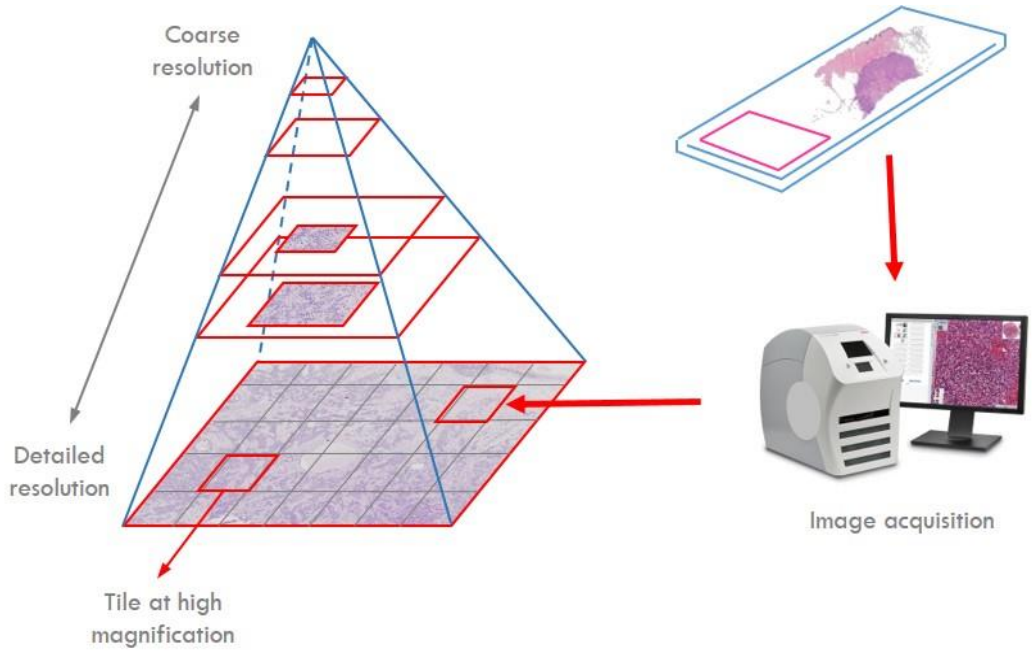


Figure 5. Visualization of a multi-resolution representation of an image (pyramidal image). The image acquisition system takes a series of shots (tiles) moving from one corner of the slide to the opposite corner. These sets of files are assembled into an image of maximum dimension. The resizing algorithms then generate the intermediate magnifications from the most detailed image, allowing the image to be visualized at different resolutions, from the initial detailed image to a low magnified image. When the user browses the slide from the entire preview, he can access tiles of different magnifications, from low to high, up to maximum resolution (base image) (adapted from (Ameisen et al., 2012)).

The result of the whole acquisition procedure is one or multiple image files and their associated metadata files, compressed in a specific file format depending on the employed device (Ameisen et al., 2012). Generally, to overcome image size problems, instead of being saved as a whole, images are split in small tiles, lighter to manipulate, saved in tiled TIFF format. As the viewer software can generate an image of the region to be displayed at a desired magnification, it is possible to scroll continuously between the lower and maximum magnification. By this method, the slide viewer will only access and display the tiles of the tissue area being requested at a given time, overcoming the issue of large file size of digital images.

An additional technique is of great usefulness when tridimensional microstructures or cell clusters are present such as in cytology preparations, or just in case of thick tissue sections, allowing to explore their width by examining the slide through the entire thickness of the cut. Known as *z-stacking* or *focal plane merging*, it consists of capturing successive images obtained at varying focus levels in the thickness of the section, meaning that in each image different areas of the sample will be in focus. Then these patches are combined into a single final image, to increase the depth of field (**Figure 6**). The result emulates the fine focusing of a classical

microscope. However, the use of this technique increases the acquisition time and data in proportion to the number of scanned and saved planes.

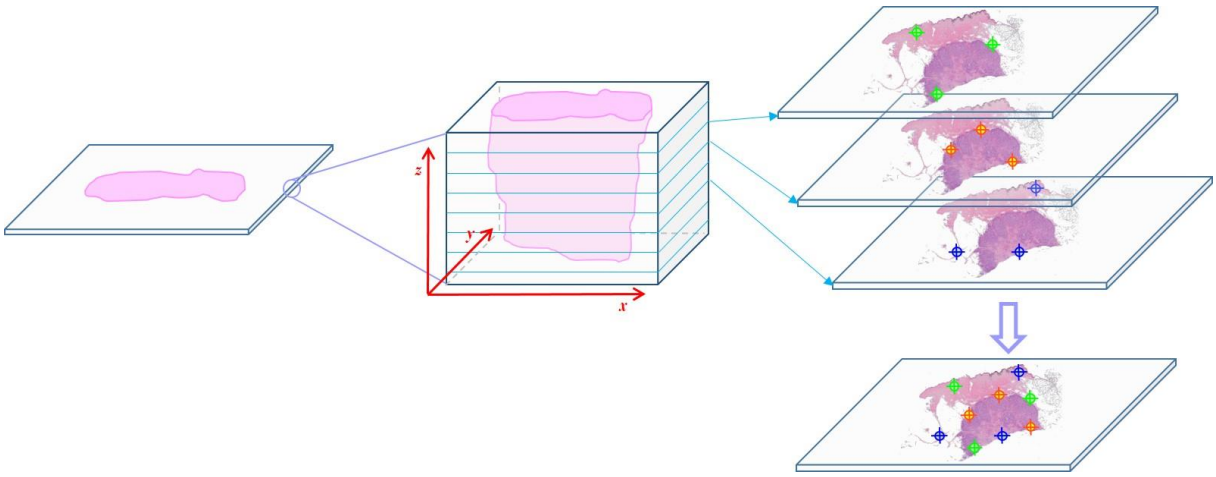


Figure 6. Z stacking. The stack acquisition allows to scan the thickness of the tissue section to capture the signals at different levels. Several images are acquired at every plane and then compiled in a single picture while keeping, for each position (x,y), the most intense pixel on the z axis.

The so-produced WSI are a great source of data at multiple scales depending on the resolution (e.g. x20, x40) and z-stack level, along with color, shape and texture information. Such characteristics, added to the fact that pathological slides are not macroanatomically oriented (contrary to the case of radiological images), make WSI ideal for a wide range of computer tools, plus the point that this sort of visual information cannot be easily assessed by the human eye, mostly when it comes to very subtle shades of color or texture (Niazi et al., 2019). Extra assets of digital slides over glass slides are that they are safe from breaking, loss, and fading staining, keeping their quality constant over time. With the aid of viewer software, and differing from what happens with a conventional microscope, a whole slide can be seen at once, zooming-in on areas of interest, without changing the objectives or having to refocus, and so avoiding fragmented views that could cause the pathologist to miss relevant things (**Figure 7**). Moreover, several slides can be displayed side by side. This option may be particularly advantageous when looking at multiple pictures of tumors over time, to compare structural details between slides or to evaluate different stainings on the same tissue area. It is also useful to annotate and to share the slides for teaching purposes, second opinions or simply to replace a multi-head microscope. Furthermore, storage issues such as the need of conditioned large rooms, and the laborious task of storing and retrieving glass slides (during which, inevitably, a number of them will be misplaced and lost) can be surmounted. Clearly, there are still drawbacks: the slides must be scanned, with the cost in equipment and human

resources that this carries; then, there has to be checked that scans are not blurry and, in a more general vision, an IT infrastructure (servers, network, screens, etc.) is needed (with redundant storage units to prevent loss of information in case of breakdown). But certainly, the opportunity given by digitized WSI to be treated by AI-based algorithms to improve the performances is a real game changer in the routine diagnostic workflow.

Figure 7. Visualization software. Example of the dashboard (a) in OLYMPUS OlyVIA 2.9 software. It allows the display of images with .vsi format. Here, zooming tool has been used to enlarge the tissue (b) and image properties are visible (c).

2. **Digital image analysis**

Digital image analysis (DIA) refers to the intention of obtaining significant information from images in an objective and reproducible manner via specialized software, reducing human bias (Riber-Hansen et al., 2012). In other words, DIA is an attempt to make the software "understand" or "read" the histological slides.

The main objective of DIA is to increase the amount and quality of data that can be acquired from a tissue section, by providing quantitative measurements of histological features

that could make possible to carry out statistical analyses. These features present in the WSI can be catalogued in three increasing levels according to their biological interpretability. The first layer or *pixel level* includes raw data corresponding to image properties such as color and texture, which completely lack of biological correlate but is precisely useful because of its objectivity. The next, called *object level* is based on the notion of segmentation (see below): the features, such as shape, texture and spatial distribution, are now related to tissue objects, for instance, cell structures. The upper or *semantic level* gathers and integrates the features from lower levels to formulate a meaningful biological concept in the form of a classification, e.g., the percent of tumor cells in a WSI (Kothari et al., 2013).

Given the magnitude of the information contained in the unique features of the WSI, the first requirement is to choose which areas are most important or relevant for a particular study. The selection of these *regions of interest* (ROI) is necessary to limit the bias on the results, avoiding the inclusion of zones with artifacts, or normal tissue zones when the study concerns the tumor cells. Besides, WSI are cropped into smaller non-overlapping or overlapping squares or *tiles* of, e.g. 256 x 256 pixels, to optimize processing time and computer efficiency.

Pre-processing aside, the analysis usually starts with biological object demarcation (cells, compartments, etc.) in order to recognize and quantitatively assess certain signals (for example, cell staining), morphology and tissue architecture, to be ultimately applied to assist diagnosis, prognosis and prediction (Aeffner et al., 2019). The process of partitioning a digital image into multiple sections or sets of pixels, assigning a particular label to every pixel, is known as *segmentation*. The aim of segmentation is to "translate" the information contained in an image to something more suitable and easier to be analyzed by the computer. This duty can be performed through visual assessment by an experienced pathologist, which proves to be a difficult task especially when it comes to mIHC/IF, or by semi-automated or automated analysis, in which the software pre-selects ROIs to be verified by the pathologist, or the ROIs are fully recognized by a trained software, respectively (Stack et al., 2014). The automated DIA uses a series of mathematical algorithms that process images, enabling the classification of their elements based on their color, texture, and/or context. Color is defined by the amount of red, green, and blue present in a pixel usually on an 8-bit scale of 0 to 255. For brightfield microscopy, which is absorptive, 255 represents the brightest of maximal color (white) intensity and 0 represents the absence of color (black); however, for fluorescence-based digital

pathology, which is absorbance coupled with emission, single color images representing individual fluorophores are converted to a gray scale, and the intensity is then measured on a similar scale (Webster & Dunstan, 2014).

To summarize, WSIs that overwent the scanning stage can be subsequently assessed with an image analysis software, with the benefit of automatizing the detection of parameters that cannot be accurately perceived by the human eye. These kinds of software usually include user-trainable algorithms, which can be adjusted until optimal results are obtained. After an initial step where several parameters are corrected, the different phases of the analysis include tissue segmentation to examine each compartment separately (e.g. tissue vs background or tumor vs stroma vs background if an antibody has been used to stain tumor cells), cell segmentation with the use of nuclear staining as a counterstain, and biomarkers' signals identification and quantification with the phenotyping tool, that uses machine learning algorithms. This is an example of what a DIA workflow consists of: an iterative process where algorithm parameters are adjusted according to the desired objective; next, the user runs the algorithm on a subset of images, and then the performance is evaluated and the parameters are possibly readjusted until sufficient algorithm performance is achieved. The contribution of the pathologists reviewing the results within the process is essential in at least two senses, to verify cell detection and correct phenotyping on one hand, and to contribute with their knowledge about the pre-analytical variables and the correlation with biological and pathophysiological expertise of the specialty on the other (Aeffner et al., 2019).

3. **Functions and utility of digital slides**

The so-obtained virtual slides can be suitable for different purposes. According to Al-Janabi et al., digital pathology may have four basic main applications: diagnostics, research, education and archiving (Al-Janabi et al., 2012). Several elements are common to the different spheres of application, such as the slide annotation tools, that are employed for training purposes, diagnostic practice and research, through identification of ROIs, tumor size measurements, etc. These benefits are appreciated at different stages, and from the point of view of the patient, faster turnaround times for diagnosis, improved access to expert opinion and more robust reports are just some of the assets of digital pathology (Williams & Treanor, 2020). Functions of digital pathology are exposed in **Figure 8**.

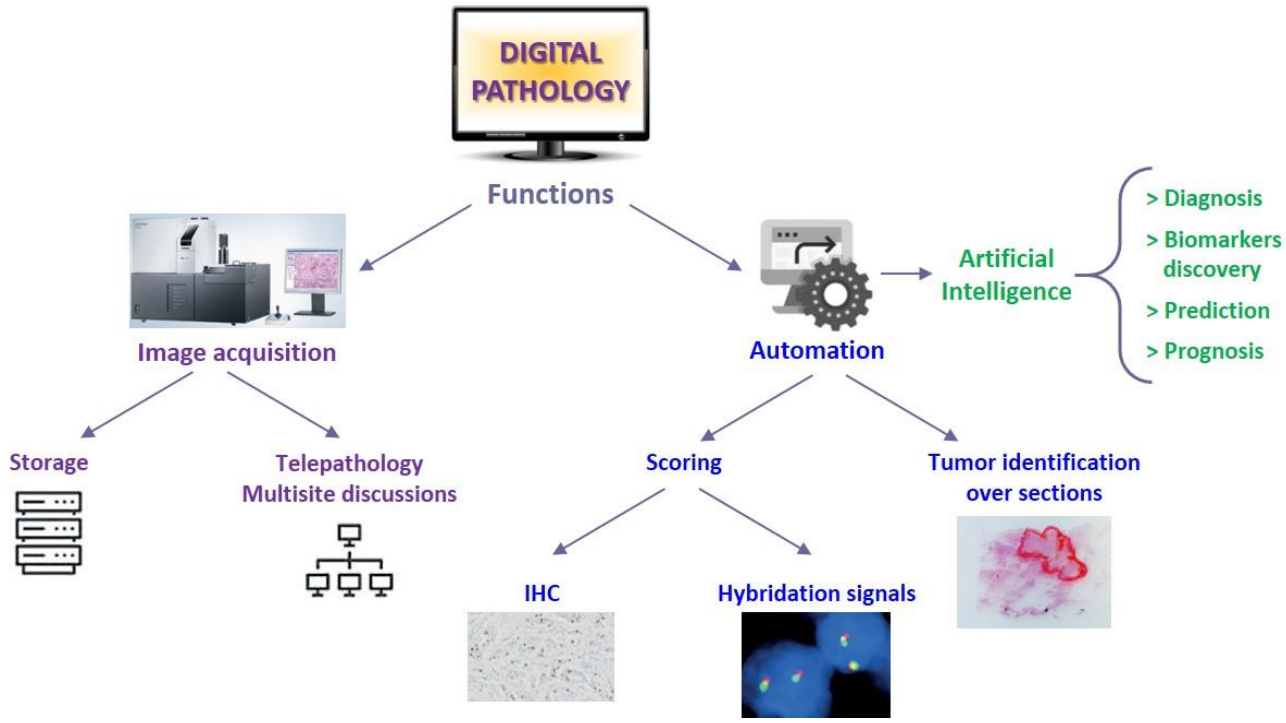


Figure 8. Functions of digital pathology. A summary of the different utilities that can be achieved with digital pathology are depicted: the digitalization of pathology slides for storage and multi-site discussion, the automated scoring of IHC and automated counting of hybridization signals, the automated identification of tumors over tissue slides for subsequent microdissection, and the more recent functionalities involving artificial intelligence (adapted from (Salto-Tellez et al., 2014)).

Diagnostics and immunohistochemistry assessment

This functionality refers to the replacement of conventional light microscopic examination of stained glass slides (H&E, IHC) with examination of WSI on screen by a pathologist to make a diagnosis (Williams & Treanor, 2020). The digital slides contain information that may be extracted in a computerized manner to standardize scorings and to automatize or semi-automatize diagnostic process. The basis of many of these assignments relies in the translation of digital arrangements and intensities into quantitative scores and in pattern recognition, assignable to characteristic features of a given diagnosis and quantitation of various biomarkers. The possibility to compare unknown histological entities that may appear in daily practice with diagnoses portrayed in textbooks or already present in archival cases, represents a great advance and a non-negligible gain of time for the patient and also for the pathologist. More interesting, digital pathology may permit the implementation of innovative AI and computer learning tools to refine tumor classifications that will ultimately simplify and accelerate treatment choices according to an accurate patients' stratification, and also a more precise prognostication, taking the quality and speed of patient care to a superior

level.

The patients, ultimately, would derive the most benefit from this technology, which may help them to receive the correct diagnosis sooner. An expert professional in very specific subjects can be reached anywhere as long as that person has a screen and an internet connection, and WSI are easier to handle in the case of presentation for discussion at multidisciplinary meetings or tumor boards. In addition, since information sharing is facilitated, the transmission of medical information in case of decentralization of medical care (patient transferred to another health care center for follow-up) is ensured and accelerated.

Image analysis has been used to quantify routinely evaluated markers in breast cancer. Excellent performances have been demonstrated for ER and PR, while more sophisticated systems are required for HER2 due to the distinctive localization of these proteins (*membrane staining* for HER2 versus *nuclear staining* for HR) (Garberis et al., 2021). In fact, and contrary to what happens with nuclei, the lack of membranous counterstain makes segmentation more complex (Shamai et al., 2019).

Diagnosis over frozen sections will be discussed in the section "D. Telepathology".

Research

A digital pathology images archive, well organized and associated to clinical information, signifies an invaluable scientific resource for the research community. Digitized slides can be rapidly retrieved for academic purposes and clinical trial review, and together with the corresponding databases, represent the ideal substrate for the development of computerized algorithms (Williams & Treanor, 2020). Publicly available digital slide repositories are an excellent contribution for knowledge development, not only making raw data available for confirmation and validation in scientific publication context, but also sharing valuable resources to further development and test of algorithms in diverse research teams, and to compare histopathological information from different clinical trials (Hipp et al., 2011).

Education and presentations

Digital pathology allows the constitution of excellent resources for undergraduate and postgraduate education. The access to instructive and unusual cases, even between different centers, is facilitated, and small tissue samples or singular findings are no longer a constraint. Since all students evaluate the same tissue section, it also eliminates the slide-to-slide

variability. In addition, the use of WSI as an educational tool permits the trainer and any number of trainees to share cases in real time, and to have an instantaneous feedback of the experience.

Other education contexts are also beneficiated by the use of digital slides. Case presentations at tumor boards, and pathology presentations in general, can dispose of high-quality images with less preparation time, and flexibility concerning image regions on a slide and different resolutions, that can be chosen during the presentation.

Archiving

The switch from physical depositories of slides to high-resolution images storage has several advantages. On one hand, it eliminates the risks linked to the conservation of glass supports (discoloration, damage, loss over time) while it also makes it easy for the pathologists to access to archived cases and to compare them with current analyses. To this object it is important to maintain a minimum searchable database with specimen-specific reports, so that slides may be rapidly retrieved, and to guarantee a periodic backup of the digital archive. On the other hand, it reduces the healthcare costs associated with the production of additional slides and duplicate tests when a patient is transferred to another hospital (Chong et al., 2020). However, the costs of digital slides storage must be taken in consideration, even if, due to the rapid advances in computational field, these prices are constantly decreasing with the increase in server capacity development.

When working with fluorescent techniques, it should be considered that the signal emitted by fluorochromes gradually fades out over a few weeks, months or years, even under optimal storage conditions. Taking into consideration the ability of certain scanners to acquire fluorescent slides, another important benefit of digital pathology is the possibility of conserving these images for long time.

4. Multiplexed images analysis

Thus far, to explain the basis of digital pathology, examples have been taken from slides stained with classic chromogenic methods. Nonetheless, there are image analysis tools capable of identifying and quantitating all targets of interest generated by a plethora of techniques, and that includes brightfield mIHC and mIHC/IF. This purpose is feasible using a multispectral camera able to capture intervals (≥ 10 nm) across the entire visible spectrum (420 nm to 720

nm) and a spectral library, which contains the information of each fluorophore emission spectra, as well as the autofluorescence spectrum from an unstained tissue, in order to provide the correct parameters to unmix and quantitate the cases of study. Several image analysis software packages can support this type of assessment.

Fluorescent mIHC has a plus that makes it suitable for DIA, due to the additive nature of the fluorescent signals and the directly proportional relationship that links the signals' intensity to the concentration of antigens to be detected. The feasibility of the analysis is the result of the different spectral characteristics of the fluorophores and their fluorescent intensity quantitation. Taking the example of the Vectra system, the original observations are spectrally unmixed into as many images as the number of "colored" markers. Then the software enables the visualization of these different pictures by constructing a composite image in which several layers that can be shut on and off independently to aid visual assessment (Stack et al., 2014; van der Loos, 2008).

To alleviate the task of automated tissue segmentation when several biomarkers are used, specific cellular attributes can be pointed out by the employment of biological landmarks. Thus, highlighting some tissue reference points, the software (and the pathologist) can relatively easily separate, for instance, tumor from non-tumor zones, or to pinpoint each cell individually.

5. Current obstacles and potential solutions

As a new technology in course of adoption, digital pathology still encounters several technical barriers to overcome, such as data storage, transmission, processing and interoperability, apart from its still high cost. Extra equipment (servers, networks, computers, scanners) and staff, as well as additional steps such as slide preparation and scanning should be integrated into the workflow, in such a way that times to deliver an attended result are not too extended, and ensuring high-quality images necessary to arrive at a consistent diagnosis. Detailed information about different scanning devices can be consulted in (Patel et al., 2021). As examples of the parameters that could negatively influence image quality, the erroneous selection of focus points during the scanning or the absence of color normalization can be mentioned. Fortunately, methodologies for color management such as spectral analysis and proper calibration can help to deal with differences in image acquisition, staining conditions, and other coloring issues. Focusing and compression ratio are other problems that can be

introduced by slide scanners, and that are not simple to handle by the user, basically because algorithms behind these operations are non-modifiable. Among the adaptable factors influencing image quality, glass histopathology production is one that could be easily improved, avoiding section wrinkles, thick tissue sections, dull microtome blades, and bubbles in the mounting media (Weinstein et al., 2009).

Visualization is another important point of improvement to soften the transition from analog to digital slides assessment, mainly because of the low screen refreshment and especially because of the poor representation of colors by certain screens (either because some colors cannot be displayed -*gamut problem* or color difference from one device to another- or because some colors are replaced by others). With the aid of technology, the "desk equipment" should be optimized in such a way that the pathologist proceeds with a similar method as using a conventional microscope (for example, creating a special mouse that allows to z-navigate on images in a more intuitive way) to make the most of a new work environment. In sum, a major evolution of the pathologist's workstation should be considered to adapt it to technical advances.

Data sharing limitations slow down the impact of the digital slides embracing. On the one hand, the open data and open code are not largely established; scarcity of publicly available datasets including both patient data and annotated images, as TCGA, restricts the progress in the field. On the other hand, technical challenges due to the enormous amounts of imaging data generated by slide scanners complicate the information sharing. This difficulty is intensified by the fact that techniques such as TMAs and novel molecular approaches generate even larger datasets, and by its dependence on the specific software used. Since no universal data format is in widespread use, each vendor implements its own proprietary data formats, analysis tools, viewers and software libraries. Pathologists become a kind of "hostages" of the digital solution used in the laboratory, and they do not have the possibility of testing the services of others vendors, seriously affecting not only the pathologists, but also the interoperability (Goode et al., 2013). The lack of standardized procedures makes more difficult to validate tools on new datasets and, vice versa, datasets generated with different pieces of software may not be able to work with; even if efforts are taking place to achieve to a normalization of methods, as is the case of DICOM regarding WSI (https://dicom.nema.org/Dicom/DICOMWSI/). Digital Imaging and Communications in

Medicine (DICOM) is the international standard for the use of medical images and their related information. With its origins in radiology, it expanded to cover other specialties that use images in their daily practice as well. The *DICOM Standards Committee Working group 26*, in an effort to make digital pathology more universal, developed standards for storage, display, sharing and management of WSI. A pathology images system based on DICOM stimulates a rapid growth of digital slides technology applications, providing the basis for image annotation and DIA (Cornish, 2020; Herrmann et al., 2018; Rojo et al., 2009; Singh et al., 2011). The FDA approved the first WSI system for primary diagnosis in surgical pathology in 2017 (Evans et al., 2018).

Nonetheless, open-source solutions are beginning to gain ground to ease the handling of WSI (W. C. C. Tan et al., 2020). Libraries to simplify the reading and manipulation of digital slides of multiple vendor formats are freely available, such as OpenSlide (https://openslide.org/) (Goode et al., 2013) or Bio-Formats (https://docs.openmicroscopy.org/bio-formats). A detail of supported image formats can be found in https://qupath.readthedocs.io/en/latest/docs/intro/formats.html. Open-source software is also integrating image analysis routines, being two of them QuPath (https://qupath.github.io/) (Bankhead et al., 2017) and ASAP (Automated Slide Analysis Platform) (https://computationalpathologygroup.github.io/ASAP/#home). In a study using a colorimetry-based evaluation method, these WSI viewers generated similar images (Cheng et al., 2020). There is still a long way to go; however, to convert such appliances into widely used products, starting by providing training, support, maintenance and adaptation to specific required tasks to obtain its maximal profits.

C. Telepathology

1. Pathologists here, there and everywhere…

The advent of digital pathology not only can help to reduce the physical transport and archival storage of glass slides. It also facilitates the sharing of digital pathology images and accompanying clinical information by the means of network connections to remote sites to be assessed by a pathologist, giving rise to what is called telepathology. Falling under the category of telehealth, the delivery of pathology services over a distance have a range of applications, such as clinical remote interpretations (telediagnosis), teleconsultation, colleagues' advice for

difficult or rare cases, slide conferences and panels, educational purposes or medical research (Al-Janabi et al., 2012).

Telepathology refers also to the sending of image files to make a diagnosis, but the greatest advantages come from the use of image exchange platforms and the possibility of sharing information in real-time. The scope of this technology is vast, it covers not only the problem of remote places where there is a shortage of pathologists but also the need of pathology expertise in small centers where the full-time presence of a pathologist is not justified. This is particularly interesting when it comes to respond requests for intra-operative consultations regarding the analysis of frozen sections. Moreover, it improves patient care quality through easier and faster consultation between professionals, resulting in more accurate diagnoses and therefore better treatment management.

Some of the evoked arguments against an extensive application of telepathology are its elevated cost, and the wrong beliefs on telepathology as a method neither fast enough nor precise (statement based on the quality of the obtained images) for its use over frozen sections, even if a good average accuracy has been demonstrated (Dietz et al., 2020; Gephardt & Zarbo, 1996).

Canada has been at the forefront of the implementation of telepathology. A successful telepathology program for frozen section diagnoses, in the absence of an on-site pathologist, has been running since 2006 at the University Health Network (UHN) in Ontario, Canada. The results with WSI in terms of accuracy are equivalent to that accomplished with light microscopy, with at least two pathologists reviewing the case and communicating the diagnosis to the surgeon within 14 to 16 minutes from receiving the tissue sample. A final report is signed out when the slides assessed by telepathology are reexamined by light microscopy in conjunction with the FFPE sections of the case (Evans et al., 2009, 2010). The Eastern Quebec Telepathology Network is another fruitful experience that started in 2011 and that allowed to provide high quality pathology services since then, alleviating the difficulty of covering such a large territory (Têtu et al., 2014).

As reflected in the UHN experience, it is crucial to establish that the pathologist will arrive at the same diagnosis regardless of the methodology employed. Before the implementation of a digital pathology system, all centers should carry their own validation procedure, needed to ensure that the WSI system can reliably work in a clinical daily practice

(Hanna et al., 2015). For instance, a variety of tasks, such as diagnosis, frozen sections assessment, teaching and supervision of residents, have been executed both in a conventional fashion (using light microscopy) and by means of remote digital pathology services to compare their performances in a study at the Førde Central Hospital in Norway. The results have shown a very high concordance between the diagnoses achieved by both methods, with a shorter turnaround time for remote digital reporting of cases (Vodovnik & Aghdam, 2018). Further information on validation is listed in https://digitalpathologyassociation.org/_data/cms_files/files/DPA-Healthcare-White-Paper--FINAL_v1.0.pdf (access October 28th, 2022). In France, the Pathology Laboratory at the Rennes University Hospital has experienced the transition towards a modern workflow in less than two years, being the first in deploying a comprehensive digital pathology system (https://www.atlanpolebiotherapies.eu/news/the-first-in-france-the-university-hospital-of-rennes-has-deployed-the-1st-comprehensive-digital-pathology-system-thanks-to-the-nominoe-fund/; https://healthcare-in-europe.com/en/news/bringing-digital-pathology-to-the-hospital-environment.html).

2. **...with a "little" help of technology**

Several years have gone by since the first photographs of blood smears were sent *via* video, in 1968 in the USA. Throughout this time, telepathology applied the available imaging technology in different ways. Four modes can be mentioned: 1) *static image type*, prior to 2000, involving histological photographs (*snapshots*) transmitted from camera-equipped microscopes via e-mail or stored on a shared server to be assessed by the pathologist/s anytime (i.e., asynchronous telepathology); 2) *dynamic image type*, more common from 2000 to 2007, where either the images are transmitted in real time (streaming) from a video camera (*video microscopy*), or the pathologist can use the robotic functions of a microscope to operate it remotely (*robotic microscopy*); 3) *WSI system,* also called *"virtual" microscopy*, the most popular since 2007, where the development of slide scanners that can rapidly create a high-quality virtual image from an histological section to be addressed via Internet allows a faster view of the digital image at various magnifications; and 4) *hybrid* methods, also known as *multi-modality telepathology*, which combine WSI with video or robotic microscopy (Dietz et al., 2020; Pantanowitz et al., 2014).

WSI system comprises both the slide scanner and the required software to view the

virtual slides on a computer monitor or cell phone screen. The representative images corresponding to the pathological diagnoses are saved on the server computer. In this way, they are available for long-distance consultations, allowing the pathologist to work from home or from any location around the world, and to easily access to revisions or opinions from colleagues. Moreover, the investment and labor costs of WSI are lower than those of robotic microscopy (Jara-Lazaro et al., 2010; Ribback et al., 2014).

Nevertheless, several factors including costs, technology restrictions, resistance from pathologists (e.g., reluctance, skepticism, technophobia) and lack of standards, limit the widespread use of telepathology. Other issues to consider, derived of the use of telecommunication systems, slow down the implementation of these technologies, such as security and privacy concerns, that are crucial for their acceptance for clinical use (Farahani & Pantanowitz, 2015).

D. Future directions: towards the slideless era?

The convergence of large tissue collections preserved in biobanks and novel methods as mIHC, along with automated computer-aided imaging technologies, allows glimpsing a future where the management of complex information will have a direct impact in routine health care, improving prognostic and predictive patient stratification. A transition phase for pathologists towards automatic scoring methods may result in more accurate characterization of diseases. Even more interesting, the possible implementation of WSI for routine pathologic diagnoses could give rise to "slideless" laboratories.

However, to fully accept WSI as a diagnostic modality, the integration of additional medical information is required. A correct diagnosis cannot be considered without the rest of the electronic medical record (e.g. gross pathology description, prior pathology reports, clinical history, etc.) (Pantanowitz et al., 2011), and a growing number of results from molecular determinations must be added to this list. Thus, the large scale of data denotes a hindrance for the pathology lab, which needs to optimize the resources in order to support data mining (i.e., exploration and extraction of information from a large amount of data) as well as a model for storage and retrieval, so that the information can be easily capitalized (Becich, 2000). Other technical factors such as the need of increasing scan speed and the WSI quality should be also contemplated.

From a more organizational perspective, the quotidian use of WSI demands guidelines, especially for diagnostic procedures, standardization of technical specifications regarding format and storage, and information sharing (e.g. consultations provided across international borders).

Nowadays, the idea of a slideless practice reflects, for most of the people, an environment where pathologists work with WSIs, but histological sections on glass slides are still necessary in diagnostic surgical pathology, essentially, because they must be produced as the first step of any digital workflow. Nevertheless, the raise of new technologies may definitely change the functioning of pathology laboratories. In an extreme representation, the **surface imaging microscopy (SIM)** technology allows to obtain a digital image directly from the surface of the paraffine block where the tissue is embedded. If a quality equivalent to that of conventional HE sections is achieved, the era of glass slides together with its scanning may be over (Jara-Lazaro et al., 2010). Other astounding slideless techniques include the open-top light sheet microscopy, which generates 3D images of tissues without the need for sectioning or slide preparation (Glaser et al., 2017), or MUSE microscopy, which uses ultraviolet illumination to obtain images of tissue surfaces almost instantly (Fereidouni et al., 2017).

Other outstanding advances for tissue assessment include three-dimensional (3D) techniques for microscopy (W. Li et al., 2017, 2019). Among them, tissue clearing is based on the modification of the optical properties of usually opaque samples, to render them transparent while keeping their structure and fluorescent labels intact, which allows high-resolution microscopic imaging (T. Tian et al., 2021). These technologies will not only enhance our comprehension of the connections between cells and their microenvironment, but it could be directly applicable to diverse areas as diagnostic medicine (W. C. C. Tan et al., 2020).

While the pathology practice is not entering headlong into this almost science-fiction era for tissue analysis, especially because of the unequal access to the panoply of novel tools and techniques, some changes are starting to operate in daily life. The implementation of WSI in medical student education will accelerate the adoption of digital pathology and fresh technologies in general in the years to come, along with the integration of equipment that allows the immediate implementation of the knowledge acquired. In certain locations, scanners are already part of the laboratory workflow: the slides pass through the staining stage, and then through the digitization step on a continuum, and software solutions are bringing

accessibility to WSI even in mobile devices.

Computer-aided diagnosis (CAD) takes the relay for automating the processing of WSI (for example, detecting lesions) in order to provide aid and assistance to the pathologist by means of a digital analysis (Chong et al., 2020). A necessary step is to simulate and capitalize the analysis product of pathologist's professional expertise, by using mathematical algorithms capable of aiding in histopathological image analysis, such as cell detection, segmentation and automated detection of diverse parameters (mitoses, metastases in lymph nodes, and ROIs) (Jiang et al., 2020). Many image analysis tools are already capable of quantifying tissue immunostaining, including the immunohistochemical expression of ER, PR, HER2 and Ki67 (Couture et al., 2018; Garberis et al., 2021; Saha et al., 2017; Vandenberghe et al., 2017) with the same or even better accuracy than the manual method (Stålhammar et al., 2016, 2018).

All these great changes in the way of thinking and practicing pathology bring new thoughts and interrogations that will have to be answered. Regarding the old and dragged glass slide, the question remains if it should be still kept for long time after scanning, in other words, which is the limit of the WSI and what could be done if the need surges to return to the "true" tissue slide (for example in the need of rescanning because of a failure in stockage that would require returning to the tissue).

As expressed by Bertheau et al., "the pathologist will always keep close contact with the tissue or cell sample of the patient because, while the image is important, it does not summarize all the complexity of a tissue or a cell which remains above all a molecular edifice complex and lively" (Bertheau et al., 2012). Maybe, in a near future, a battery of special studies and techniques will be used to perform a full characterization of the specimens before reaching the pathologist's hands, so that he or she has a maximum of information from the beginning of the diagnostic process. In any case, it is clear that we are far from being able to understand all the scope of the digital revolution.

Artificial Intelligence and Medicine

A. Basics of artificial intelligence approaches in pathology

1. Introducing some concepts

Artificial intelligence (AI) is emerging in Medicine and particularly in pathology as a novel tool for improving the precision of diagnosis strategies. AI describes automated systems that can perform tasks considered to require "intelligence", imitating what a human being could do when confronted to the same situation. These systems are based on the creation and application of algorithms (Liu et al., 2019). Adopting computational neurobiology, mathematical logic and computer science, AI holds promise for enhancing automation, elimination of tedious tasks, improved accuracy, and efficiency. By identifying patterns of recognition, AI allows the processing of large datasets that contain information from clinical and pathological records, as well as data generated in genomics research which, due to its scale and diversity, could be more difficult to handle if manually analyzed (Topol, 2019).

Depending on tasks that the system is able to handle, one could speak of two different kinds of AI. In popular culture AI is akin to the concept of *strong AI* (artificial general intelligence, AGI), which defines a type of multitasking AI with human-level intelligence, but in practice it is just *weak AI*, which can perform a unique specific task with high accuracy (Tizhoosh & Pantanowitz, 2018).

Three concepts are often confused in the AI field: AI, machine learning, and deep learning. Robertson defined *machine learning* (ML) as the science of making computers analyze and learn from data without human instruction, carried out by artificial neural networks (ANN) (Robertson et al., 2018). ANN are computational models able to recognize features through the extraction of abstract attributes from datasets, subsequently utilized to achieve certain tasks such as classification. Thus, ML could be described as a subcategory of AI (narrow AI) in which some aspect(s) of human intelligence are exhibited, through the development of algorithms capable of "learning" to solve a problem directly from the data, that is to say, without being reprogrammed (**Figure 9**).

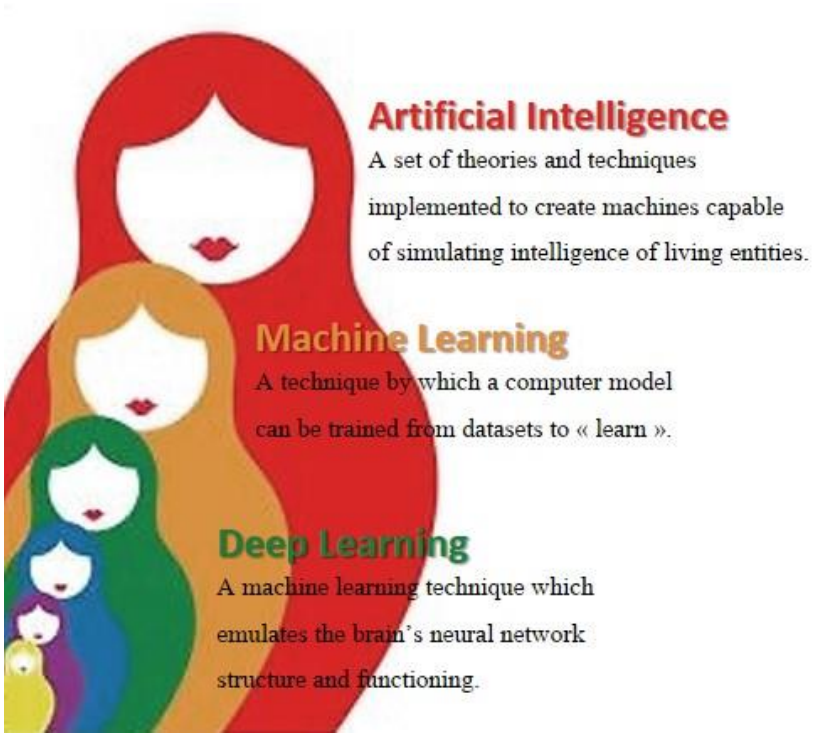


Figure 9. Perspective of the different concepts in the artificial intelligence sphere.

In the general case, a mathematical model is first created with a scientific objective (for example, to recognize images) and then trained to "learn" the weights that could carry out the task (in the example, to predict the configuration in future cases). The first step or *training phase* involves providing huge amounts of data to the system and permitting the algorithm to adjust itself and improve (McClelland, 2017; Q. Xie et al., 2019). In the second step or *prediction phase*, as its name implies, the ANN processes the input to produce predictions. The size and variety of the datasets, as well as their prospective character, are important factors to consider in order to avoid erroneous results (Mutasa et al., 2020). When the inputs of the system are images, they are best treated on *Graphics Processing Units* (GPU), highly specialized electronic circuits that allow the calculations in parallel for fast processing of pixel-based data, such as histological pictures.

There are different ML learning paradigms: supervised, weakly supervised, unsupervised and reinforcement learning. *Supervised learning* refers to train a model to predict outcome statuses that are provided as labeled data for training. In other words, this labeled data commonly corresponds to examples of the correct output, contained in the training data. The model learns by trying to establish the relationship between the input data and the assigned label. In *unsupervised learning*, the model is trained to recognize patterns within the input data

without the use of labels. Thus, a supervised algorithm, such as support vector machines (SVM), random forest (RF), or classification trees, needs human experts to manually delineate important information in the data, while an unsupervised algorithm, such as clustering, is able to extract features without labeled data, making of it an interesting tool in research due to its utility to describe hidden structures from images. Neural networks (NN) seem to be flexible as they can be either supervised, unsupervised or weakly supervised, depending on the type of architecture considered. *Reinforcement learning* could be summarized as a subset of unsupervised learning, where the NN employs self-supervised learning and obtains supervisory signals from the structure of the data itself, using any observed or unhidden part of the input to predict a hidden part (or property) of it (Esteva et al., 2019; Zhu et al., 2020). A small mention to define a method used in our work, *momentum contrastive learning* (MoCo) (X. Chen & Fan, 2020) which stands for a self-supervised learning algorithm that allows the training with data of lower quality. The "*contrastive*" refers to the training data for a binary classification task, which can be divided into positive examples (those that match the target) and negative examples (those that do not match the target). Contrastive self-supervised learning uses both positive and negative examples (Camp, 1996; Y. Tian et al., 2021).

Supervised learning has been traditionally the most common approach in digital pathology but, due to the costs of precise annotations, *weakly supervised learning* is becoming more popular. In the latter, the detail in annotations is reduced at the expense of model accuracy. For example, in multiple instance learning (MIL), a type of weakly supervised learning algorithm used in our study, the labels are not referred to each patch (patch-level labels) but to the entire WSI, and therefore easy to obtain. Each WSI of the dataset is considered a "bag" containing a set of instances (patches or *tiles*), and there is one single label per bag. If all its instances are negative, the bag is considered negative, and if at least one instance in the bag is positive, then the whole bag is considered positive (known as the standard MIL assumption). The objective of MIL is to predict the label (of a bag or an instance) based on training data that includes only bag labels (**Figure 10**). An example of high diagnostic accuracy obtained with weakly supervised learning is detailed in (Campanella et al., 2019). The most outstanding characteristic of MIL regarding the medical field is that it can be integrated into a deep learning model, which allows the creation of systems where WSIs are fed in as inputs and diagnoses are returned as outputs. DeepMIL models can automatically reveal novel, abstract features from

WSIs that perform better than traditional features at determining survival, treatment response, and genetic defects (Ilse et al., 2018).

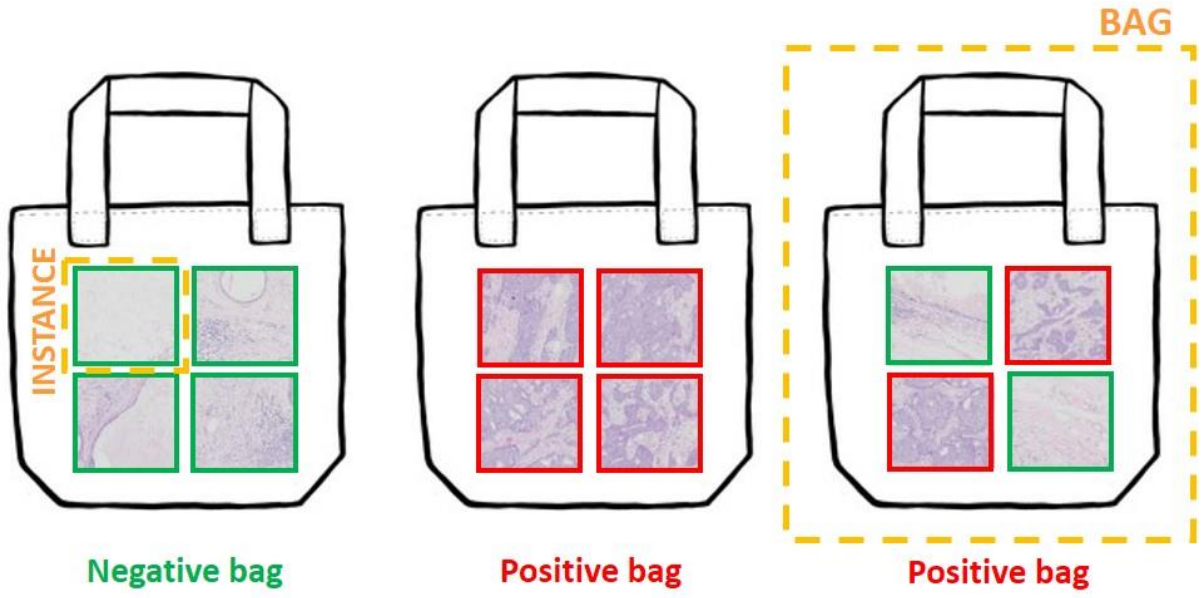


Figure 10. Multiple instance learning. Each WSI is represented by a bag containing a certain number of tiles (image segments) or instances, extracted from the original digitized slide. When all the instances in a bag are negative (non-tumor tiles) the bag is labeled as negative; when the bag contains at least a positive instance (tumor tile), it is labeled as positive.

Deep learning (DL) is a recent biologically-inspired ML approach, where depth is generated by a sequence of multiple cascading layers used by the algorithms to enhance the system. Each layer contains a variable number of neuron-like computing units, emulating neurons in the visual cortex in the brain, and is focused on a specific feature to learn, such as curves/edges in image recognition. In contrast to non-DL machine learning methods, DL does not need feature extraction or representation steps: data travels from the starting line to the output last layer without human intervention, which is why it is called an *end-to-end model* (Komura & Ishikawa, 2018). Several types of neural network exist, depending on how the units or *nodes* of each layer are connected, how the data is propagated inside the network, how the network learns the patterns, and to what extent the network can remember what it has learned. Some examples of DL approaches are feed forward (FF) NN, multi-layer perceptron (MLP), recurrent NN (RNN), and convolutional NN (CNN).

In *FFNN*, the first and simplest NN, the information moves in only one direction—forward—from the input nodes, through the hidden nodes (if any) and to the output nodes (Schmidhuber, 2015). There are no cycles or loops in the network. *MLP* is the classical type of

NN. It is a supervised learning algorithm suitable for classification and regression prediction problems where inputs are assigned a class or label. Data is often provided in a tabular format. Since it is flexible, it can be applied to different types of data, such as the pixels of an image that can be reduced down to one long row of data and fed into a MLP (Dayhoff & DeLeo, 2001). *CNNs* (see below) are also suitable for prediction, and they are useful for data that has a spatial relationship: the benefit over MLP is their ability to develop an internal representation of a two-dimensional image. This allows the model to learn position and scale in the data, which is important when working with images. *RNN,* derived from FFNN, differ in the ability of their connections to create a cycle. They were designed to work with sequence prediction problems, suited for both sequences of text and sequences of spoken language, and thus used for text data and speech data. They are not appropriate for tabular datasets nor image data input.

In DL, the data passes through the system in a process called *forward propagation*, where the output from a previous layer is taken up as input for the next layer. As the information gets across the different levels, the prior representation gains in complexity and abstraction, and when the network has been propagated through, a loss function is used to calculate the *error* (the difference between the output and the ground truth – see below) and the internal parameters or *weights* are readjusted (*back propagation*) to reduce this error; then the forward propagation is started over again. By this method, iterated over thousands of training cycles, computers can handle millions of labelled images and "learn" from them until they are ready to recognize these structures autonomously, that is to say, to classify unknown images without human aid (Aeffner et al., 2018) (**Figure 11**).

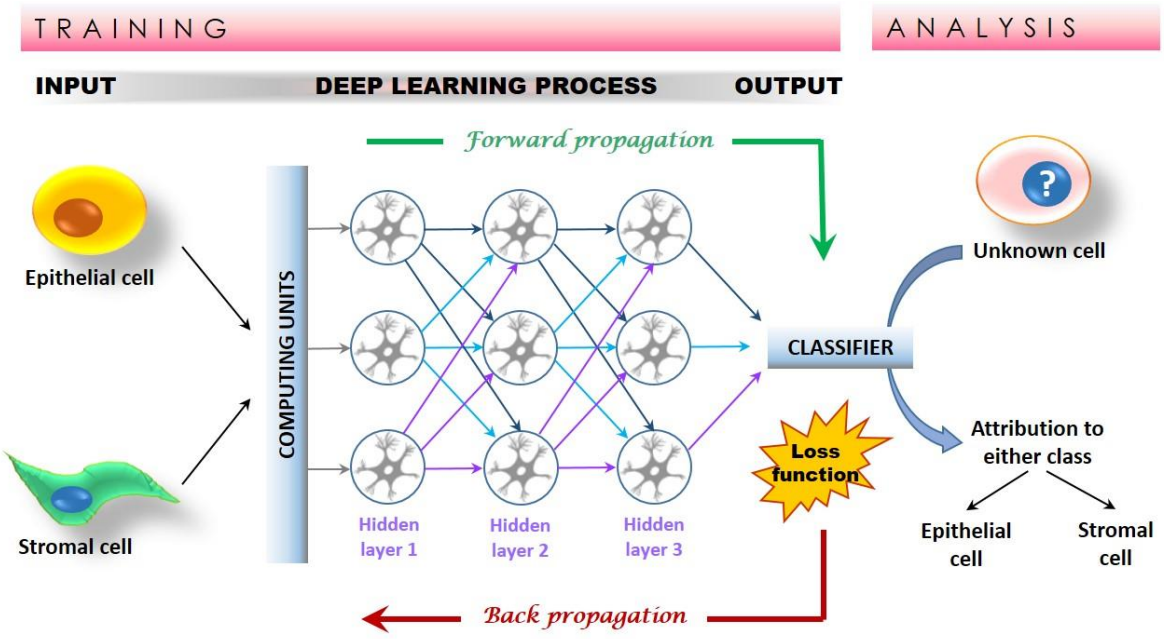


Figure 11. Simplified schematic of deep learning approach. Example images of epithelial cells and other cells are fed into a convolution layer that passes on features to a neural network to learn to identify epithelial cells autonomously in unknown images. As the information is passing through the neural network (comprised of connected computing units called "neurons" arranged in layers), the algorithm freely selects the most important image features of epithelial cells. Entering these layers of neurons, this feature information is transformed and passed on through weighted connections to the next layer. The algorithm completes thousands of training cycles to learn to recognize these structures autonomously in novel, unseen images (adapted from (Aeffner et al., 2018)).

As a case of representation learning, DL can automatically discover the representations needed for detection or classification in the raw data with which it is fed. This process, that involves a global recognition of patterns across spatial scales – from organ systems to cells and even molecules, requires arduous training by the machine, that can learn by itself, and an ability to associate the recognized lines and curves with figures and whole forms and to complete the empty spaces or missing information (Chan & Salto-Tellez, 2012). In the example of an image, the assemblage of pixel values conforms the raw data, and the first layer focuses on the learning of edges. The second layer will spot the motifs designed by particular arrangements of these edges and, in turn, these motifs may be part of larger combinations that correspond to parts of familiar objects. Then, subsequent layers could assess combinations of these parts to detect objects, in a hierarchical manner similar to that existing in speech and text. As explained by LeCun et al., "the key aspect of DL is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure" (LeCun et al., 2015).

2. The visual recognition: basis of the histological analysis both by the pathologist and the computer

It could be said that a large part of the cognitive abilities responsible for image processing reside in the visual cortex of the human brain, and that they are based on recognition. Basically, we identify an image that we have already seen, through comparisons and inferences regarding our prior experience. The same principle of image recognition is employed by the AI-based approach in pathology, while adding a wide range of options to advance and optimize workflow (Heeke et al., 2019).

In computer vision, the extraction of visual characteristics from a digital image consists of mathematical transformations calculated on its pixels. The *pixel* (px), the smaller element of a display surface, often presented as a colored square, is the unit of measurement for the definition of a digital image. Each pixel is associated with a color, usually reconstituted by a triad of primary components rendering red, green and blue tones by electrical excitation.

Convolutional neural networks (CNN or *ConvNets*), a specialized type of neural networks, are among the most famous DL models for pattern recognition and detection tasks, being efficient solutions to extract relevant features from digital images (Ertosun & Rubin, 2015). CNN are particularly well adapted to images because of their usefulness for multi-scale analysis and their capacity to capture well the invariances present in the images (Mallat, 2016).

This DL algorithm is named after the mathematical operation *convolution*, employed in at least one of their layers (*convolutional layer*). In the simplest version, as described by LeCun in (LeCun et al., 1998), a convolutional layer and a pooling layer are stacked up and followed by a few fully connected layers and a task-specific output layer to form a deep model. The pooling layer is used precisely to save space in memory by reducing the size of feature maps by spatially averaging the information. The limit in the number of layers is determined by the amount of memory available on the GPUs and by the acceptable amount of training time (Chang et al., 2019; Krizhevsky et al., 2017) as well as the number of parameters (i.e., a single layer but with 1 million neurons will be limiting). In this case, a neuron corresponds to a filter (a matrix of a certain dimension, which will traverse the image through the convolution operation).

The general architecture of ConvNets is directly inspired by the LGN-V1-V2-V4-IT hierarchy in the visual cortex ventral pathway (**Figure 12**).

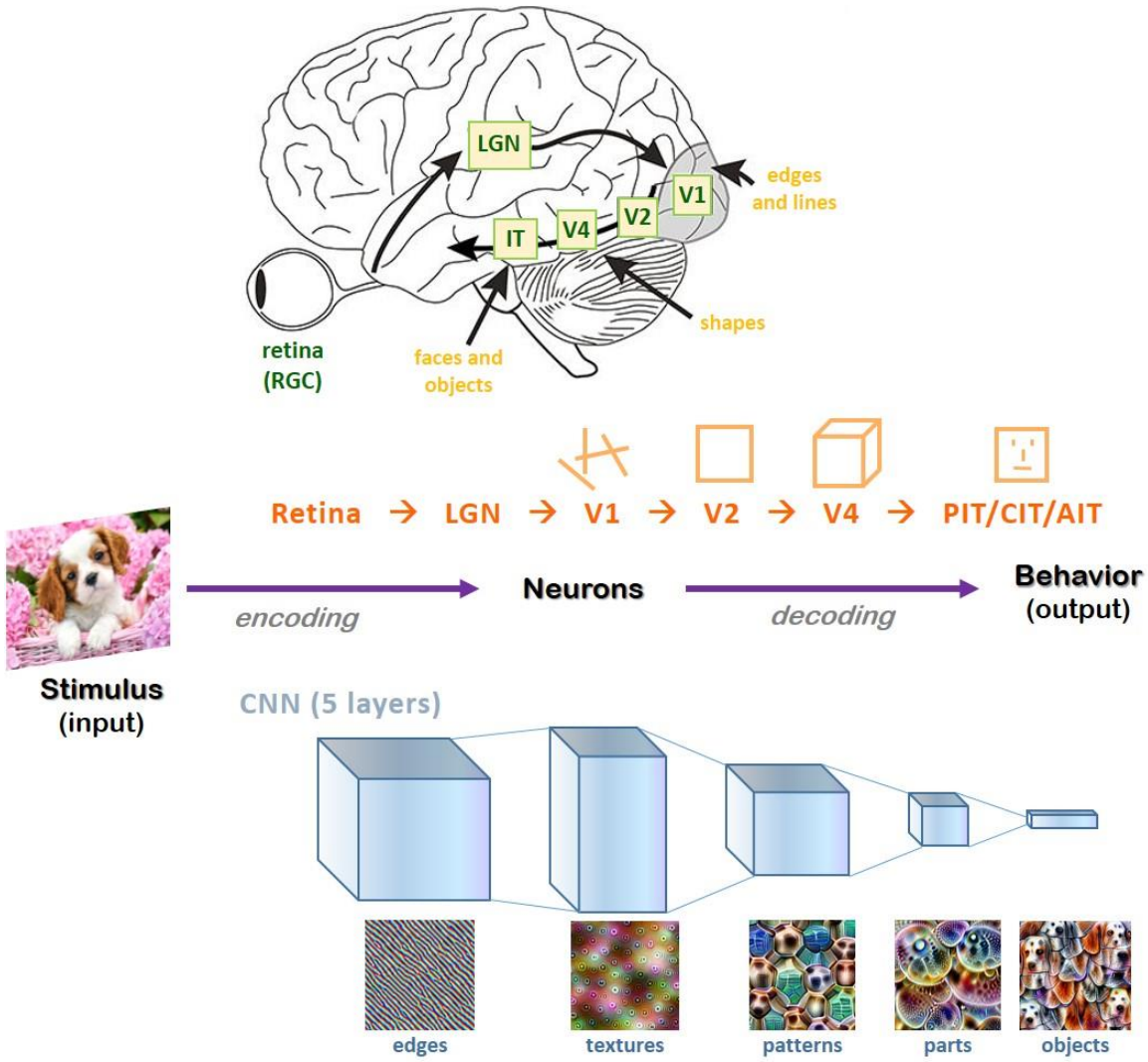


Figure 12. Comparison between the human visual pathway and a CNN. Most CNNs have at least 5 layers, the final of which feeds into a fully-connected layer. Frequently 2-3 fully connected layers are used in a row and the final layer of the network performs a classification. CNNs can replicate the creation of neural representations in a similar way as performed by the visual system through the same "untangling" process. That is, both systems take representations of different object categories that are inseparable at the image/retinal level and create hierarchical representations that allow for linear separability. Feature visualization shows how a CNN "sees" and how it could build up its understanding of images over many layers.
CNN: convolutional neural network; **LGN:** lateral geniculate nucleus; **RGC:** retinal ganglion cell; **V1,2,4:** visual cortex 1,2,4 (adapted from https://gracewlindsay.com/2018/05/17/deep-convolutional-neural-networks-as-models-of-the-visual-system-qa/, feature visualization extracted from https://distill.pub/2017/feature-visualization/).

Unlike traditional image analysis, where the features of interest are manually selected, CNN can perform automatic identification, guided or not by an element called *ground truth*. The ground truth can be a category, a quantity or a label, manually delineated by a pathologist in the form of an annotated dataset, provided to the algorithm along with the raw images set, that will orientate it in the prediction or characterization of an image. Depending on their use

of ground truth for construction or validation, the models can belong to the supervised learning or the unsupervised learning categories. In the first, the algorithm training is based on a previously defined label, i.e. the recognition of benign vs malignant regions according to manual annotations, while in the latter, natural divisions in the dataset are recognized without the need for a ground truth. This is the case of tumor clustering into subtypes according to similar image attributes (Abels et al., 2019).

3. Of 'black boxes' and 'glass boxes'

The term black box refers to a system where the inputs and outputs are perceivable but its process of decision remain 'hidden'. The operation of such system (in this case, an algorithm, or an ANN) is therefore only understood from the angle of its interactions, and there is no verifiable path to understand the rationale behind its decisions. This is not a lesser point to consider, mostly when it comes to features directly correlated to clinical endpoints. DL approaches, working as black boxes, do not allow explaining or proving the results obtained, and even the way in which the algorithm reached its deductions is difficult to comprehend. The main issues derived from the limited interpretability of DL models are, on the one hand, the lack of transparency that can make an algorithm look as non-reliable, and therefore hinder its acceptance among physicians and other experts involved in the diagnostic field and, on the other hand, the difficulty of correcting an algorithm for which the reasoning is unknown.

The opposite of a black box, called a glass box, is a system whose mechanisms are visible and allow the understanding of how they work. This is the objective of the *explainable AI* (XAI), through strategies such as the visualization of pixels that impacted the decision-making (Samek et al., 2015) or the interactive *ML+human-in-the-loop* approach that uses the strengths of human cognitive abilities where automatic approaches fail, mostly due to data complexity or insufficient training samples (Holzinger, 2018; Holzinger et al., 2017). The intervention of the pathologist decoding the steps followed by the network and the way in which features are used to elaborate a conclusion, may be solutions for the still reluctant acceptance of this method (Mobadersany et al., 2018). Moreover, the human-computer interaction could increase the performance of either AI or pathologist alone (Abels et al., 2019; Liu et al., 2019).

However, explainable AI approaches do not resolve the problem of interpretability in a completely satisfactorily way. In a trade-off between performance and interpretability, and as

pointed out by Rudin, the current efforts are centered on creating second models to explain the first black box models, rather than designing models that are inherently explainable, and this could produce, in the eagerness of elucidating the original algorithm, methods that are finally not faithful to it. Conceiving models that are interpretable *per se* instead, would aid earning clinicians and patients' trust, cutting with the algorithms' interpretability problem (Rudin, 2019).

B. Data and requirements for AI implementation in pathology

1. Data sets terminology

A data set, as the name implies, is a collection of data coming, for example, from a single database, that will be transformed into a numerical representation that the ML algorithm can process. The required size of the data set will depend on the complexity of the ML project. The amount of information contained in it must be good enough for the algorithm to work properly, allowing the split in different non-overlapping data subsets, with the necessary amount of available data, for the algorithm to acquire the ability to generalize and to converge towards an optimal solution. Data sets are usually divided into a training set, a validation set and a test set, and they take up about 60%, 10%, and 30% of the data, respectively, though these percentages vary depending the considered literature (Angermueller et al., 2016).

The *training set* is fed to the ML algorithm to teach a model, by searching for data patterns among input variables. Training is performed by updating the parameters iteratively until the model optimally fits the data.

The validation and testing samples are used for analyzing the performance of a ML model. By using samples that were not employed for building or adjusting it, the result of the model's effectiveness will be unbiased.

The *validation set* is used right after training, to tweak and adjust the hyperparameters by evaluating the performance for a given set of hyperparameters at each training. The hyperparameters are the variables which determine the network structure, e.g. the number of hidden units, and the variables which determine how the network is trained, e.g. the learning rate. Hyperparameters are set before the learning process begins; they are tunable and can directly affect how well a model trains itself. To improve the reproducibility, especially when working with limited data sets, the adjustments may be repeated using multiple random

partitions of training and validation sets, in a resampling procedure named *k-fold cross-validation*, where k-fold refers to a single parameter *k* corresponding to the number of groups that a given data sample is to be split into. This technique is useful to estimate the skills of a ML model on unseen data. The procedure involves randomly dividing the data set into k groups, or folds, then the first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds. The configuration of the groups does not change for the duration of the procedure: the samples stay always in the same group they were assigned, meaning that each sample is used in the validation set 1 time and used to train the model k-1 times. The results of a k-fold cross-validation run are often summarized with the mean of the model skill scores.

The *test set* is independent from the training or validation sets. It is used to understand how the ML model will work, to evaluate its accuracy, and to determine if it can be applied clinically. After this phase, the ML model is usually not adjusted anymore, in order to avoid overfitting (see below).

2. Considerations for pathological image analysis

There are two basic preconditions for the use of AI in pathology: 1) the digitization of histological slides, and 2) the availability of structured data (especially concerning the information present in pathology reports) of good quality and in enough quantity to be used as input. The presence of structured data, which is, translated into a language the computer can understand, such as large spreadsheets holding unique numbers or values, is unusual in health care, where most of the information is in an "unstructured" form consisting mostly of clinical chart notes (Bini, 2018).

Furthermore, apart from the hard work that it takes to convert all that poorly preserved data into something usable, it is of paramount importance to corroborate that this data is clean, artefact free and comprehensive to train correctly an algorithm. In fact, the unique characteristics of the pathological datasets can pose serious problems that end up affecting the image analysis. Each one of the steps until obtaining a WSI, and the devices implied in the process, can introduce undesirable effects damaging the quality of the digital images. These steps refer to tissue characteristics (fixation and sectioning, staining, orientation, heterogeneity of biological samples, morphological and architectural structure of histological regions such as overlaps, touching objects, weak boundaries, etc.) and acquisition conditions (presence of dust, uneven thickness of the scanned section, marker traces, scanning magnification, focus,

different equipment and file formats, etc.). To alleviate the impact of these variations before launching the automated analysis, different pre-processing techniques, such as normalization, can be applied (Chang et al., 2019).

The very large size of WSIs (in the order of tens of billions of pixels) can be a real problem, first to fit in memory, and second, because classic ML architectures do not work well on such large scales (tissue to cell). It is to consider that if they are resized to a smaller dimension, there is the possibility of losing information at the cellular level. Thus, WSIs need to be divided into squares, so-called "patches", of about 256 x 256 pixels, that will be analyzed independently. Once generated, these tiles will be used to construct the training and validation sets.

Then, most AI algorithms need a large set of training images. Two AI approaches can be distinguished: those based on image annotation, and those more autonomous based on DL. They use two broad classes of features respectively: handcrafted features and unsupervised features (Madabhushi & Lee, 2016).

Handcrafted features hold a certain degree of interpretability by their connection to particular attributes in the image, such as architecture, shape, color and texture. These characteristics are identified by expert pathologists for the "labeling" or annotation of images, which includes also the manual delineation of ROI. This task may turn out not only boring but also challenging when working with not enough sharpness or resolution in the learning set images. Hence, efforts have been made to find the way of automatizing their detection, using computer vision operations such as color space conversion, image blurring, sharpening, edge detection, morphological transformation, pixel value quantization, clustering, and thresholding (Chang et al., 2019).

More related to DL strategies that learn from unknown pattern, unsupervised features undergo interpretability issues regarding the feature selection performed by the deep network in an autonomous fashion, but count with the advantage of being rapidly operational for any problem.

Frequently, due to lack of expert annotations, patient privacy issues or very specific scientific questions, ANN are confronted to medical image datasets that are not large enough to achieve successful performances. Different strategies can be put into practice to bypass this

obstacle. One of them is *transfer learning*, in which the designed system is pre-trained on a different (and larger) annotated dataset and with a different task, and then the system's parameters are calibrated with data on the target dataset. Through the re-use of the learned features by the CNN from a different source task and data, apart from solving the problem of limited data, the effort of manual labeling is reduced, as well as the training time (Bayramoglu & Heikkilä, 2016; Tewary & Mukhopadhyay, 2021). Another tactic is *data augmentation*, which consists of artificially increasing the size of the dataset by performing operations that modify the appearance of the image (for example, by reducing the brightness, by rotating the image, etc.), without modifying its corresponding label. A third technique entails the processing of the input image at different magnifications, that is to say, different algorithms operating at various scales and obtaining information at different levels (cellular, structural). Finally, different ML methods can be combined to accomplish a single task to improve prediction. For example, a first step depicted by the trained DL network, employed merely as a representation method, may be continued by a second technique, such as RF or SVM, among others (Robertson et al., 2018).

Finally, and not less important, a frequent topic of dealing with medical data is the privacy of the patients and their medical records. When launching a project, it must be guaranteed that datasets comply with General Data Protection Regulation (GDPR). A usual solution is to give restricted access to the individuals who are involved in the project, but there is still the problem of the data used to train a deep network when the model must be released. Novel techniques propose a solution on this concern: the use of encrypted input data (Dowlin et al., s. d.; P. Xie et al., 2014).

3. Overfitting and spectrum bias

One of the most important and recurrent limitations in CNN algorithms applied to medical research is that the datasets are relatively small and frequently coming from a single institution. This goes against the principle of model improvement that comes from the training with larger datasets, but also with the need of prospectively collected data from multiple care centers, that will be required for further validation of the algorithm. These major issues are known as *overfitting* and *spectrum bias*.

Overfitting is a recurrently encountered challenge in AI algorithm development. It refers to a situation in which a CNN algorithm becomes exceedingly reliant on the provided training

data, to the degree that it has a negative effect for the model to generalize to new data, while concurrently inflating the model's performance with the training dataset (Park, 2018).

Very optimistic results in the validation set might raise the suspicion of overfitting, as well as a large performance gap between the validation and test sets and data-related factors (differences in patient populations, or in the source of the observations -e.g. diverse scanners employed-). The use of algorithms on new, unseen datasets, analyzed in a prospective manner, could be a possible approach to limit this phenomenon, and an important step in transitioning this fresh technology out of the research lab and into the clinic (see point *D. Validation of AI methods*).

The other mentioned pitfall in CNN models is given by the data collection. When cases are included in an unnatural ratio between the presence and absence of the event of interest (e.g., relapse) there is the risk of *spectrum bias*. It means that the dataset used for model development does not satisfactorily reproduce the target population, i.e., the patients to whom the algorithm will be applied in real-world practice. This is an important consideration when working with rare diseases.

C. The process

1. A few more definitions...

When conceiving a project that implies the use of DL, it can be chosen: 1) to work with an existing architecture or 2) to design a new network with a certain number of layers, convolutional attributes and type of input. Then, ANN have two phases: a learning phase and an operational phase. Next are cited some notions for a better understanding of the process.

Two concepts are related to the way in which an algorithm browses the information present in the dataset: *epoch* and *iteration*. The first means the number of times a learning algorithm process the entire dataset, while the second refers to the number of batches or steps through partitioned packets of the training data, needed to complete one epoch, being a *batch* the number of training samples or examples in one iteration. Allowing this iteration over parts or batches, are necessary considering the huge data size to be processed. To sum up, an epoch refers to one cycle through the full training dataset, and several iterations exist for each complete epoch. In guise of example, if it is desired to conduct a project with a big training set of 1 million images in total, and the batch size is set to 50K, this means the network needs 20

(1M/50K) iterations to complete one epoch. Then a certain number of epochs is required to train a neural network.

2. The learning phase

In concordance with the different partitions of the datasets described above, model development includes three sequential steps: 1) training, 2) validation, and 3) testing. These denominations should not be confused with the names given to the datasets decomposition. The basic mechanism for learning follows the principle of "trial and error", through the forward and backward propagation already mentioned.

In the *training step*, different categories to be recognized, or *instances,* are defined, each one with a list of relevant characteristics, or *features.* By learning from these features, the model can categorize new instances not seen before. Taking the example of how to teach a computer to identify which organ each sample comes from in a set of tissues, the set of all the labeled images is called the training data set or ground truth. This database of histopathological images allows creating some kind of "textbook cases" for the network to learn. After "absorbing" them, the ANN will have learned which combination of features is associated with any type of tissue. The information goes back and forth in the network as much until the desired behavior is obtained. As is deductible from the example of the set of tissues mentioned above, a key aspect for an effectively use of a DL approach is to provide a sufficiently rich set of exemplars, i.e., images from all the organs, to ensure a good representation of diversity in the training set (Janowczyk & Madabhushi, 2016).

In the *validation step*, the pathologist evaluates the degree of "maturity" of the network to estimate the performance of the model thus obtained. Continuing with the example, at this point the software is confronted with an image it has never seen before, and it recognizes with a certain accuracy which tissue it belongs to. It is to highlight that the accuracy goes up with the number and size of the training set, and that is why it is so important to precisely define the minimum number of cases required for training. Furthermore, the software can adjust itself and learn through feedback loops (right decision/wrong decision) (Ryad Zemouri et al., 2019). DL models, some of them with up to 100 layers deep, can extract their own relevant features without human input if the training set is large enough. It means that they can find by themselves, in our example, the features that define each tissue without being told what those features are, and with no need of structured data sets to learn from.

In the *testing step*, new image patches (cases that until now had never been "seen" by the model) are submitted to the network, for which a class prediction from the learned model is obtained.

3. The operational phase

Once the model has been validated, the operating phase then follows, where the network is ready to use in aiding the pathologists to elaborate their diagnoses. Kayser et al. detailed the steps followed by a virtual slide included in an AI-based diagnostic system (Kayser et al., 2009). A concise scheme of the workflow that contemplates the inclusion of AI methods in diagnoses assistance is illustrated in **Figure 13**.

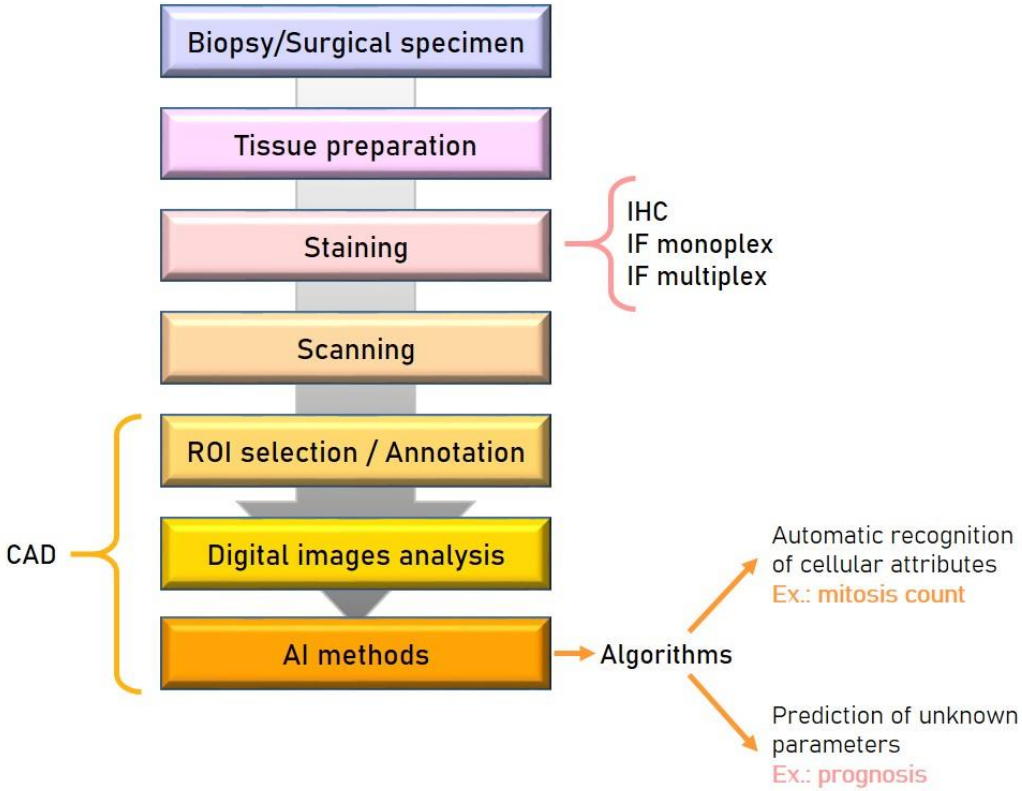


Figure 13. Workflow for a pathology lab that incorporates computer aided diagnosis (CAD). AI: artificial intelligence **IF:** immunofluorescence; **IHC:** immunohistochemistry; **ROI:** regions of interest (adapted from (Garberis et al., 2021).

D. Validation of AI methods

1. Some considerations about evaluation

Before getting into details of how an AI method can legitimately meet the operational needs of the medical community, a consideration must be made that validation has different

meanings regarding the field of medicine and the field of AI. While, in the first, it means the process of verifying the performance of a diagnostic or predictive model, in the latter, it refers to a specific step in development in which the model is fine-tuned, or the most optimized model from among minor variants is chosen after training. Another term, *test*, is used instead to denote the process of evaluating the model's performance (Park, 2018). This aspect should be taken into account when referring to the denomination of datasets. In the guide published by Liu et al., in an attempt to adapt the terminology of computer science to the clinical field, the validation set is referred to as "tuning set" while the test set becomes the "validation set" (Liu et al., 2019).

Another remark is the need of comparison of the DL model with an accepted reference standard or *gold standard*, which encompasses the ground truth before descripted. In many cases, manual pathology assessment of lesions and biomarker scores by an experienced pathologist serves as the reference standard, and alternative methodologies could be used as benchmark to counterbalance its subjective nature (Aeffner et al., 2017).

As with any other medical technologies, AI algorithms must embody clinical validation before their widespread application in the daily practice: they have to be *tested*. In fact, this is not a minor issue. First, to estimate the degree of confidence accorded to the result of an algorithm, and to comply with the rule of *internal validation*, it must be ensured that data has been accurately collected and processed and that the biases attributable to these steps have been limited or eliminated if possible (for example, data cleaning when it comes to databases, picture quality in projects involving image acquisition and, most important in the medicine field, scarcity of events in non-prevalent diseases or collected data that is not representative of the whole population of study). Internal validation refers to the assessment of the algorithm with the data that were used to develop the model. Typically used internal validation methods are *cross-validation* and *bootstrapping* (Hastie et al., 2009).

Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations (Kohavi, 2001). It is frequently used for prediction issues, when it is required to estimate how accurately a predictive model will perform in "real life". The objective of this method is to give an insight on how the model will generalize to an independent dataset by testing it on data that was not used to train it. Thus, data are partitioned into complementary subsets, then the analysis is performed on the designated

subset, the *training set*, and validated on another subset, the *testing set*. Multiple rounds of cross-validation are performed in order to reduce variability, and the validation results are combined (e.g. averaged) over the rounds to derive a more accurate estimate of the model's predictive performance. *Bootstrapping* states for a statistical procedure where a single dataset is resampled to create many simulated samples (Hesterberg, 2011). In this method, n samples are extracted at random from the original dataset a number k of times, and a model is trained on each of these new sub datasets. The predictions will then be averaged across the k different models to get the final predictions, leveraging the different training conditions of the models to reduce the uncertainty during inference. It must be noted that increasing the number of resamples will not increase the amount of information in the data.

Second, the generalization of the approach, or *external validation*, must be carried out. Ideally, this can be accomplished through the study of a sufficient number of samples that were not used for model development, and exhibiting a variety of characteristics that more or less represents the entire spectrum of the investigated problem. However, AI algorithms performance is usually calculated on test datasets that are nothing more than random subsamples or *splits* of the original dataset, hence external validation is difficult to achieve. Two strategies are employed to overcome this matter: the use of a completely external dataset, exploiting data from newly recruited patients (*temporal validation*) or from a different site (*geographic validation*). The less preferred use of a small section randomly chosen from the original dataset and kept untouched for use as a test dataset (*split-sample validation*) may address the internal validity of a model but would not correctly evaluate its generalizability (Park, 2018). External validation requires several features such as a diagnostic cohort design, the testing in multi-institutional data (interoperability), and a prospective data collection and here is where a lot of studies fail (Kim et al., 2019).

Last but not least, there is the interpretability problem: the lack of understanding of AI systems by a large sector of the medical community could cause the reject of the algorithm decision just because physicians do not understand how the DL model draws its conclusions.

2. **Algorithms' performance analysis**

Measurement tools in AI

Accuracy is the most widely used measure to evaluate the performance of an algorithm and is perhaps the major driver while developing predictive models, since it will determine

which model to choose. Briefly, it represents how many predictions of the classifier were in fact correct (Henriques Abreu et al., 2016). However, the predictions of a good model should not only be accurate, but also well calibrated: accuracy is not trustful in a misbalanced dataset.

Typically, the receiver operating characteristic (ROC) curve and the calibration plot are respectively used to evaluate discrimination and calibration, two different aspects of the performance of diagnostic or predictive models (Park, 2018). Discrimination accounts for the sensitivity (True Positive rate, i.e. the proportion of positives that are correctly identified) and specificity (True Negative rate, i.e. the proportion of negatives that are correctly identified) of a binary classification algorithm, whereas calibration shows the relation between the true class of the samples and the predicted probabilities.

A ROC curve is a graph representing the performance of a classification model for all classification thresholds. This curve plots the rate of true positives as a function of the rate of false positives (**Figure 14**).

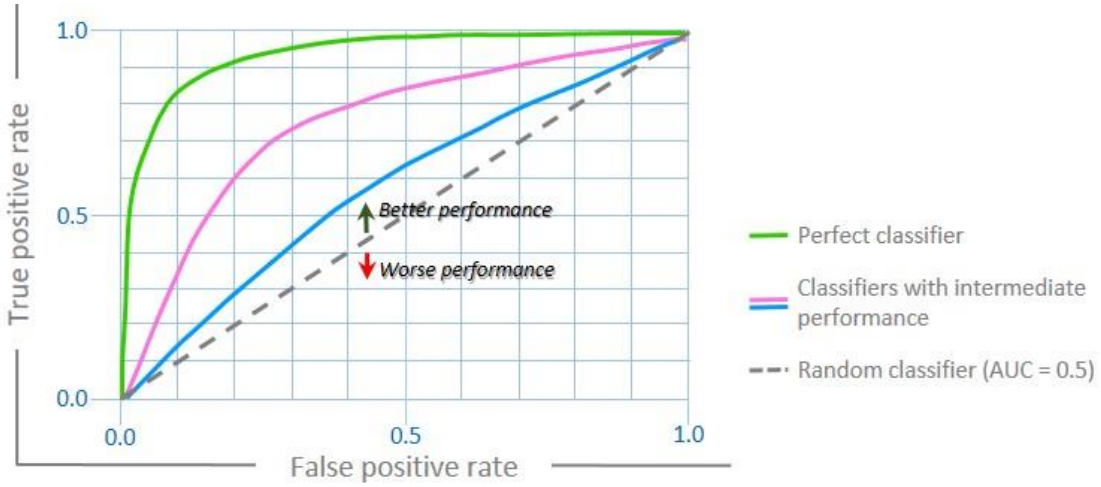


Figure 14. ROC curves. In a random classifier, the distribution of examples is hazardous. "Pink" model shows a more accurate classification than "blue" model. The AUCs of their ROC curves are between 0.5 and 1, which means that these two models rank a random positive example above a random negative example more than 50% of the times.

The area under the ROC curve, or AUC, also known as *C statistic*, measures the entire two-dimensional area under the entire ROC curve (by integral calculations) and provides an aggregate measure of performance ranging from 0 to 1 for all possible classification thresholds. The AUC is a widely employed value that can be interpreted as a measure of the probability that the model will rank a random positive example above a random negative

example. In other words, a model with 100% wrong predictions has an AUC of 0, and a model where all the predictions are correct, has an AUC of 1, meaning that the closer the AUC is to 1, the better the discrimination performance of the diagnostic test. However, in practice, a "perfect" classification model with an AUC of 1 should be suspicious, as it probably results from an error in the model; an example of this situation could be the overfitting of the training data.

Another form to visualize the performance of an algorithm is the confusion matrix (**Figure 15**), where the instances in a predicted class are confronted to the actual class, allowing to easily recognize whether the system is *confusing* two classes (i.e. commonly mislabeling one as another).

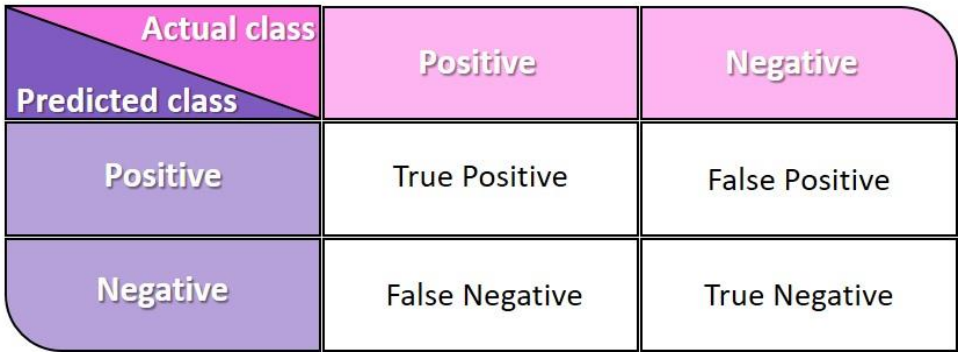


Figure 15. Confusion matrix. By comparing the predicted values with the true values taken by a variable, it is possible to evaluate the precision of the model.

The calibration plot displays the goodness of fit between predicted and real probabilities. In this graph, the samples are clustered in deciles (0-10%, 10-20%, … 90-100%) according to their class probabilities. The mean values of every decile predicted by the pre-trained model are represented on the x-axis and the corresponding event rate (true outcome) on the y-axis. Perfect calibration would be reflected as a 45° line (**Figure 16**).

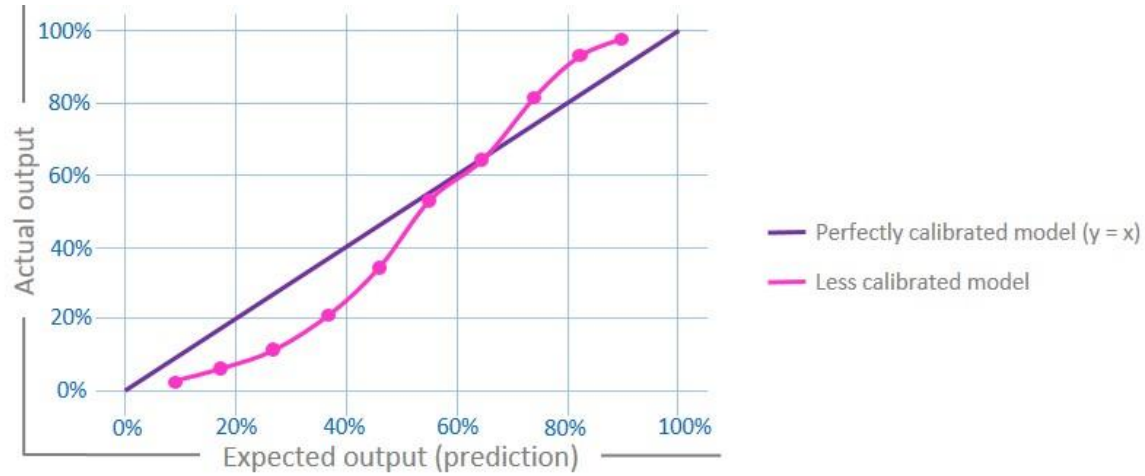


Figure 16. Calibration plot. Example of two calibrated models with the actual class label represented in the y-axis and its predicted probability represented in the x-axis.

Other essential metrics used to ensure a model performs well are *Precision/Recall* and *Dice coefficient (F1-score)*. All three measures distinguish the correct classification of labels within different classes. Recall is a function of its correctly classified examples (true positives) and its misclassified examples (false negatives). Precision is a function of true positives and examples misclassified as positives (false positives). There is a trade-off between precision and recall: when tuning a classifier, improving the precision score often results in lowering the recall score and vice versa.

F-1 score is a measure used to assess the quality of binary classification problems as well as problems with multiple classes. It ranges from 0 to 1, with F1-score = 1 meaning that the model has perfect precision and recall, otherwise said, signifying the greatest similarity between the prediction and the ground truth. F-1 combines precision and recall into a single metric, using the harmonic mean, which is given by the formula: $F1\text{-}score = 2 \times (precision \times recall)/(precision + recall)$ (Goutte & Gaussier, 2005; Sokolova et al., 2006).

Finally, beyond all the mentioned performance metrics, the model validation continues with the demonstration of its value by comparison against the reference standard (the best diagnostic or therapeutic modality for a condition, issued by experts in the domain), and through its effect on patient outcomes. This can be attained by means of clinical trials or well-designed observational outcome research (Park & Han, 2018).

Putting the models to the test: the public *grand challenges*

One of the main difficulties in algorithms validation in the medical field is the lack of big, curated datasets. This drawback started to change with the advent of *challenges*. In these open competitions, researchers are provided with a platform that includes a set of data, to evaluate the performance of their algorithms by answering a particular scientific question within the domain of computational pathology (Hartman et al., 2020; Serag et al., 2019). Putting out image datasets for the public and raising meaningful scientific subjects, challenges are excellent means to compare potentially useful models in medical imaging issues.

Competition topics are various, such as the automatic detection of metastases in lymph nodes of breast cancer patients, exposed in the CAMELYON16 (Ehteshami Bejnordi et al., 2017) and CAMELYON17 challenges (Bándi et al., 2019; Litjens et al., 2018), the quantification of tumor cellularity (Petrick et al., 2021) or the assessment of biomarkers. For the "HER2 challenge contest 2016", based on the automated detection of HER2 status on IHC stained slides, 8 of the 10 top-ranked methods were based on DL (CNN). The results supported the idea that an automated or semi-automated scoring method has a high potential for deployment in daily practice (Qaiser et al., 2018; Qaiser & Rajpoot, 2019). The recent "HEROHE Grand Challenge 2020" raised a similar question but on HE images of invasive breast cancer, where the task was to achieve HER2 status without the corresponding IHC image, searching to exploit morphological characteristics as surrogates for this determination. A cascade of DL classifiers plus multi-instance learning (MIL) were applied to the dataset, showing good efficiency scores for different evaluation metrics (Conde-Sousa et al., 2021; La Barbera et al., 2020). Five challenges focused on the evaluation of mitoses in invasive breast carcinomas: MITOS2012 (Roux et al., 2013), AMIDA13 (Veta et al., 2015), MITOS&ATYPIA14 (https://mitos-atypia-14.grand-challenge.org/), TUPAC16 (Veta et al., 2019) and MIDOG2021 (https://imig.science/midog/) (Aubreville et al., 2022). Another competition has been held in the MICCAI (Medical Image Computing and Computer Assisted Intervention) Conference regarding mitosis detection: the ICPR 2012 Mitosis Detection Competition (Cireşan et al., 2013).

The French Society of Pathology (SFP, *Société Française de Pathologie*) organized its first data challenge in 2020, based on biopsies of uterine cervix (Delaune et al., 2022).

E. Applications

The striking contribution of AI to health professionals is the possibility to convert an abundant assortment of disconnected data into investigated information valuable for decision-making, through the correlation between different biological parameters evidenced by mathematical algorithms (Lecuona & Villalobos-Quesada, 2018). In the pathology sphere, AI could take part to cope with the enormous quantities of data that digital pathology creates, not only aiding in binary classification but also in segmentation, estimation of continuous measurements, and workflow automation. The applicability of AI on breast cancer molecular pathology could also optimize the interpretation of the generated data, improving the recognition of different pathological subgroups that could be correlated with definite outcomes. Moreover, it will allow the pathologists to work at a "higher level" supervising and managing the system, giving them the time needed to focus on problematic cases. AI software tools can diminish their workload by handling time-consuming tasks such as counting mitoses or assisting with case triage (Kayser et al., 2009; Pantanowitz et al., 2020). Algorithms could also provide the pathologists with meaningful knowledge to support them in the daily decision-making, through the purveyance of archive cases similar to the event being examined or prior specimens from the same patient for comparison (Tizhoosh & Pantanowitz, 2018).

Generally speaking, a ML model can be used for numerous purposes in the medical field. Considering only diagnostic applications, AI may assist from the screening phase, selecting high-risk patients, then in the diagnosis construction by backing physicians and improving their accuracy, to the post-diagnosis phase where it could function as a quality check (Liu et al., 2019). DL models are applicable to several imaging modalities, such as CT and MRI data, endoscopy and dermoscopy images, but the histology slides are the support that can carry the bigger density of information, which turns WSI into an appealing resource for these approaches (Echle et al., 2021). As showed by a study conducted by Turkki et al., useful information for outcome prediction can be retrieved by ML even in small samples such as tissue microarrays (Turkki et al., 2019). Applied to the research field, AI can detect patterns that are not noticeable for the human eye, opening the avenue to the so-called imaging biomarkers, i.e. new unknown biomarkers resulting from DL algorithms. The correlation of these patterns with molecular subtypes, treatments responses, and prognosis provides an opportunity to refine the diagnostics in precision medicine (Ektefaie et al., 2021; Rakha et al., 2020).

Digital image analysis tools and AI methods can be applied to a large variety of tissue preparations to identify particular lesions or patterns, to optimize cancer scorings and to quantify immune infiltrates (Aeffner et al., 2018). Typical image analysis tasks in the context of digital pathology include detection, segmentation and tissue classification, as well as quantification and grading. As mentioned above, segmentation consists of delineating precise borders for morphological elements such as nuclei, epitheliums, etc., so that these features can be correctly extracted. Detection refers to merely pinpoint the center of the cell or event of interest (e.g., a lymphocyte or a mitotic figure). Tissue classification is a more complex assignment that can be performed by learning a set of archetypal features of a tissue class via DL. This approach needs annotated image patches with the class label to learn the most representative characteristics for class discrimination. Thus, the accuracy of the classifier relies on rigorous ground truth annotations, performed by an expert in a time-consuming and laborious duty. Optimization of the process of ground truth production demands pixel level precise annotations, and dealing with the representation of three-dimensional (3D) structures in 2D planar images of tissue sections. DL could be the solution itself for the first task, providing high-quality annotated outputs to be verified by pathologists, who will save time focusing only in the correction of errors made by the DL network (Janowczyk & Madabhushi, 2016). From this we can infer that the objective of AI applications, as explained by Robertson et al., "is not to replace the pathologist but to make the diagnostic workflow more efficient and help evaluate and extract the most important information from the images, as well as to detect patterns not visible to the human eye" (Robertson et al., 2018). Some of the current applications of AI in breast cancer pathology are exposed in **Table 5**.

Applications	Comments
Diagnostic applications	
Tumor detection	
Primary tumor detection:	Detection of malignant tumours and differentiation from benign and normal structures using digitalized images of fine-needle aspiration biopsy samples [1]. Quantitative measurements of nuclear shape and size, which could be applied across different tumour subtypes [2].
Metastatic deposits detection in lymph nodes:	Detection of metastatic tumour deposits in the lymph nodes, with a higher diagnostic achievement over 11 pathologists [3].
Breast cancer grading	Breast cancer grade assessment by image analysis with DL [4], objective enumeration of mitotic figures [5], measurements of nuclear shape and size, with automatic detection and segmentation of cell nuclei in histopathology images [6].
Breast cancer subtype	Image analysis with DL to detect breast cancer histologic subtypes [4].
Assessment of tumour heterogeneity and tumour microenvironment	Measurement of intra-tumour and inter-tumour heterogeneity [5,7], identify and quantify non-epithelial cells such as fibroblast, neutrophils, lymphocytes and macrophages [8] and computerized image-based detection and grading of tumour infiltrating lymphocytes (TILs) in HER2+ breast cancer [9].
Receptor status and intrinsic subtype assessment	Quantitative measurements of immunohistochemically stained Ki-67 [10], ER [4], PR and HER2 images [11]. GAN-based approach to provide a virtual immunohistochemistry staining pattern from the H&E stained WSIs [10,12]. Image analysis with DL to predict breast cancer intrinsic subtype [4].
Prognostic Applications	
Prognostic significance of tumour morphological features	Morphological features (nuclear shape, texture and architecture) to predict risk of recurrence and overall survival in patients with ER-positive breast tumours [2].
Prognostic significance of different peri-tumoral elements	AI-based assays to measure the arrangement and architecture of different tissue elements such as TILs within the tumour have demonstrated their value in predicting survival [7] and that the spatial distribution of TILs among tumour cells expression profiling is associated with late recurrence in ER-positive breast cancer [13].
Applications related to predictive values and response to treatment	ML approaches used to correlate the expression of certain markers such as cell cycle and proliferation markers [14] or the presence of certain morphological features in the tumour to the response of specific therapy.

Table 5. Current applications of artificial intelligence in breast pathology (adapted from Ibrahim 2020 The Breast). References: 1- (Osareh & Shadgar, 2010); 2- (Whitney et al., 2018); 3- (Ehteshami Bejnordi et al., 2017); 4- (Couture et al., 2018); 5- (Lu et al., 2016); 6- (Al-Kofahi et al., 2010); 7- (Yuan, 2015); 8- (J. Chen & Srinivas, 2016); 9- (Basavanhally et al., 2010); 10- (Sahiner et al., 2018); 11- (Hossain et al., 2019); 12- (Z. Xu et al., 2019); 13- (Heindl et al., 2018); 14- (Tőkés et al., 2016).

Two main applications that illustrate the use of ML in oncology are *classification* and *regression*, according to the variable related to the prediction problem. Thus, the prediction problems of a qualitative variable are classification problems, while the prediction problems of a quantitative variable (non-categorical) are considered regression problems. Classification or pattern recognition are in general addressed with supervised learning approaches, which bring advantages such as accuracy and reproducibility when assessing the expression of IHC markers, tumor morphology, and spatial distribution of tumor infiltrating lymphocytes (TILs). Regression or survival analysis is a traditional means to assess the prognostic significance of each candidate feature or covariate. Since ML tools are able to detect novel features in intricate input data, they open the gate to obtain prognostic information beyond established classifications and to develop novel predictive models in cancer research, based on complex nonlinear relationships among prognostic factors. The model learns to predict parameter values that will correlate with a better or worst prognosis (Koelzer et al., 2019; Ryad Zemouri et al., 2019) (**Figure 17**).

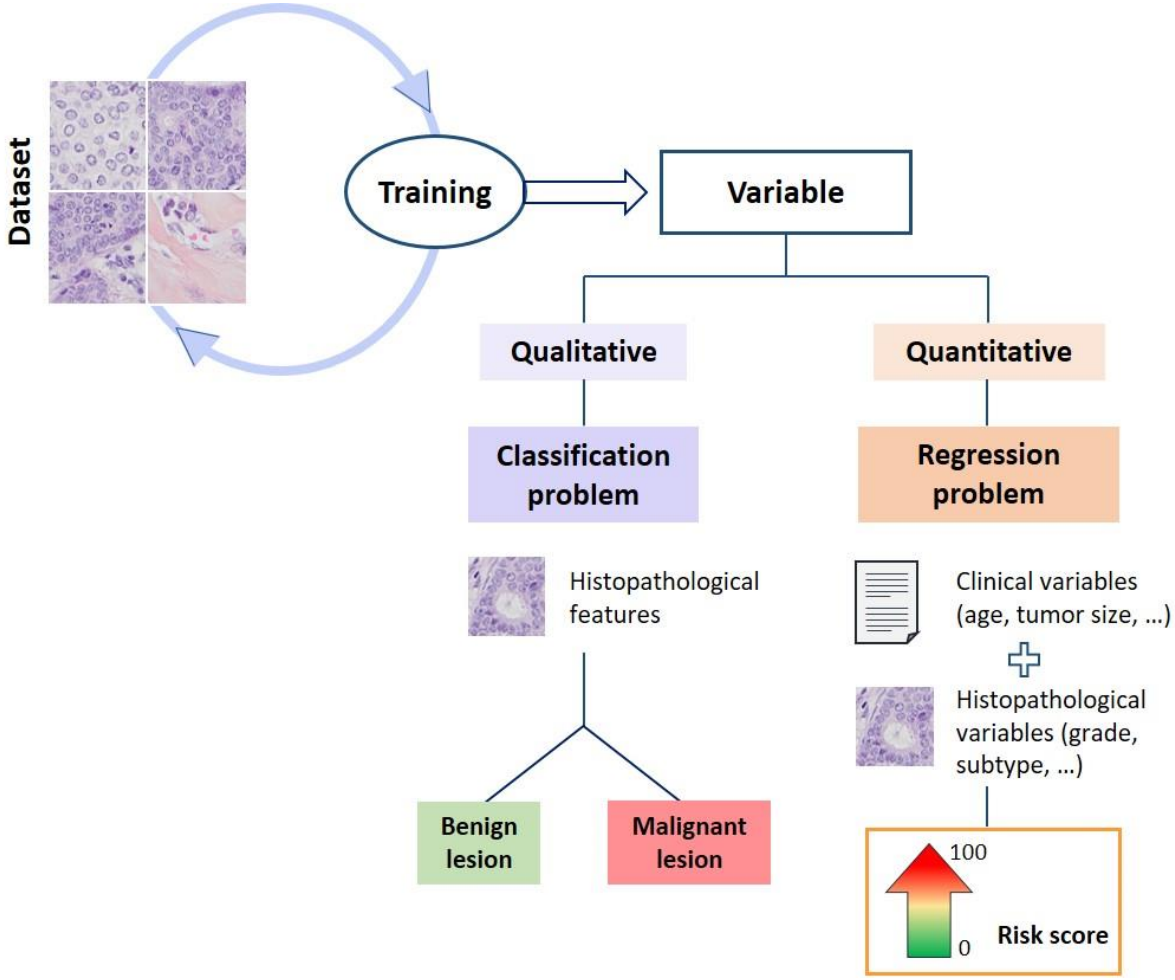


Figure 17. Classification versus regression. After a first step where a model learns on training data, it can be used to solve two types of problems. In classification, it will categorize objects into classes or categories defined by the user. In regression, it will predict an unknown parameter from a sum of elements, that may be employed to construct complex scales that allow stratification, such as risk scores.

Histopathology has found in ML methods a source of appliances useful to improve predictions in two different fields of application: 1) clinico-pathological workflow, reproducing a specific task according to the state of the art, to assist pathologists in diagnostics task and/or to automatize repetitive duties, simplifying patients' stratification, and 2) research, enabling the detection of unknown signals not accessible to classical observation in large datasets coupled to outcome-based labels (Bataillon et al., 2019; Ibrahim et al., 2020). Some of the tasks included in the first point represent a gain of time for pathologists, such as the detection of tumor versus benign lesions/non-pathological tissue, or the diagnosis of lymph node metastases (Ehteshami Bejnordi et al., 2017), while other tasks serve to bring objectivity to the results, such as histological grading (Jaroensri et al., 2022) or the evaluation and quantification

of IHC staining to assist pathologist in biomarker status determination. The second point refers to the "prognostic applications" that will aim to develop novel drugs or to predict patients' response to treatments and evolution, such as TILs quantification or evaluation of the stroma (Courtiol et al., 2019; Dodington et al., 2021; H. Li et al., 2021; Nederlof et al., 2022; Wu et al., 2022).

F. Novel defies posed by AI: shaping the medicine of the future

As digital pathology market is increasingly expanding, laboratories that rely on digital pathology will either start or intensify the use of AI to meet the demands of modern medicine. To accomplish this transformation, the generalization of the approach must be achieved, in order to ease its adoption by different centers, using data from several scanner brands and models, and slides issued from different preparation processes. These considerations are essential in order to avoid, for example, the association of a particular disease with a device employed in the health center where it is prevalent and, as a result, the prediction of this disease every time that the algorithm is confronted to a data coming from the same device.

An already significant improvement in workflow is expected on tedious and time-consuming tasks such as mitosis counting, but it requires also an effort on the part of the pathologists, that must have to be comfortable embracing this technology, participating in preparation tasks such as slides annotation, and in the supervision of the global operation of the model. This labor will translate in an increased accuracy of image-based diagnostics, that will improve with pattern recognition and, in the long run, it will reduce their workload. Hence, even if, at the beginning, the change to a workflow that includes AI and digital pathology may imply a lot of work on understaffed pathology departments that suddenly must handle a vast amount of data, it will be worthy later. As an example, we can look upon the great number of normal biopsies for certain screening procedures: AI can learn what is normal or not, and flag the cases that need further investigation, improving tasks organization.

Advances will go even further. The implantation of AI approaches that, applied directly on H&E sections, can predict biomarker status and molecular subtypes, could compete with IHC or identify patients most likely to benefit from molecular testing, even to refine cancer subtyping, as shown by Islam et al., by integration of multiple data sources (Couture et al., 2018; Jaber et al., 2020; Mohaiminul Islam et al., 2020; Shamai et al., 2019).

In the breast pathology field, the application of AI, apart from improving the pathologist's diagnostic accuracy and biomarker assessment, will also deliver results beyond that can be gain by eyeball assessment of histological characters. DL could thus constitute a cheaper and faster alternative for some of the expensive multigene assays to predict the outcome of breast cancer, if not replacing them completely as mentioned above, or at least as a previous step to the implementation of these onerous tests (Finberg et al., 2007). Promising results obtained by Coudray et al. in lung cancer show that it is possible to predict certain gene mutations from image data alone (Coudray et al., 2018). Determination of mutational status meets the obstacle of the small number of slides containing positive instances (i.e. gene mutations) required to attain good accuracy values. Nonetheless, it is worth directing efforts to better constitute the training datasets: given the impact of these mutations in treatment choices, this DL application may have a crucial role in the personalized medicine field in the near future.

Historically, the characteristics taken into account by pathologists to estimate the prognosis of tumors have been related to epithelial cells, as reflected in the Nottingham grading system. Nonetheless, some stromal features have gained space in histopathological reports (Salgado et al., 2015). As proven by the findings of Beck et al. in DL models research, stromal features are highly associated with overall survival, suggesting that the evaluation of these parameters may refine prognosis assessment, being useful, for instance, to achieve better discrimination between low and high-risk patients in histological grade 2 subgroup of invasive early breast cancer. Even more interesting, the system used in this study was automated with no manual steps (excepting the quality check of the images), demonstrating the utility of DL approaches to find features whose significance was not previously documented (Beck et al., 2011). In a similar vein, and based on computer-aided HE histopathology images analysis, Chen et al. suggested that assessing stromal morphologic features may offer significant help to improve outcome prediction in breast cancer (J.-M. Chen et al., 2015).

Several limitations, some of them already mentioned in the sections above, include drawbacks from the preanalytical phase that can be the cause of artifacts on the final image such as creases, shadows, blurry areas and variation in coloring, to the later steps such as the need of significant storage capacity (~3GB per slide scan) to save all the high-definition images. Regarding the data itself and the DL model application to it, on one side, there is the enormous

number of tissue types and lesions that, combined, engender a never-ending list of samples to be learnt by an algorithm, with the consequence this entails for training: an even longer list containing several examples for every case is required! On the other side, DL strategies should be adapted to the many peculiarities of the medical field, not only fitting the larger size of medical images and "less curated" datasets obtainable from most of the medical studies, but also producing a pertinent result. The classical binary variable classifier should be tailored to the clinical practice, where the answer is not "yes or no" but a more complex result that takes into account cognition, experience and clinical context. Other constraints are the need of annotations to be performed by an expert, and the fact that outcome depends on many factors, which may be present or not on the analyzed images (Robertson et al., 2018; Tizhoosh & Pantanowitz, 2018). With respect to "practical" questions, another limitation is the cost of modernization of outdated IT infrastructures in health care centers. High performance computers are needed to guarantee a rapid data processing, and user-friendly platforms to ease adoption among those who do not come from the computer field.

Pitfalls can be resumed in four key points, as exposed by Xu et al.: 1) too few datasets, 2) variances among equipment used in the different locations (lack of standardization), 3) lack of explanation capacity of current AI methods (algorithm seen as a *black box*), and 4) diagnosis of rare diseases (not enough cases to meet the requirements of AI approaches) (J. Xu et al., 2019). As pointed out by the authors, most (if not all) of these items can be solved by working at the AI model level, i.e. perfecting algorithms that can be adapted to different conditions and size of available information, or resorting to transfer learning strategies (reusing pre-trained CNN networks on other problems, employing open source image libraries) or combining several models to overcome the scarcity of learning samples.

The interpretability problem is of particular importance in pathology. The FDA have already approved in the US more than 60 applications in ML/AI for radiology, cardiology, and internal medicine/general practice so far, but only a few for pathology (Benjamens et al., 2020; Jahn et al., 2020); https://medicalfuturist.com/fda-approved-ai-based-algorithms/; https://www.thelancet.com/cms/10.1016/S2589-7500(20)30292-2/attachment/c8457399-f5ce-4a30-8d36-2a9c835fb86d/mmc1.pdf). This may be due to the fact that with pathological diagnoses comes a great complexity and responsibility and this is hard to accept when sources of decision remain locked. This trend could revert with the development of adaptive algorithms

that can be modifiable while being used.

Since image research cannot be validated without the active involvement of pathologists to confirm certain aspects (influence of pre-analytical variables, WSI quality, algorithm performance) an important step to take is to train young pathologists in computational science basics and in new technologies which, rather than looking to replace them, cannot continue to "learn" without the pathologists being at the heart of the process: it is the pathologist who contribute the pathophysiological knowledge to interpret the generated data. Hence, answering the recurrent question of whether AI will soon supplant pathologists, in Heeke's words, "AI won't replace pathologists, yet pathologists who use AI could replace those who don't" (Heeke et al., 2019).

Objectives

Objectives

The aim of this project is to build clinical decision support tools that could aid in patients' stratification and treatment, applying mathematical models over multi-parameter data with the purpose of expanding health care. Giving accurate predictive information on the behavior of individual malignancies could improve the determination of which patients with an initial diagnosis of breast cancer may be safely observed rather than require further therapies. Thus, the application of image analysis coupled with AI and using information and supports already accessible at the Pathology laboratory, could better predict relapse compared to the currently available methods.

Secondarily, we attaint to elucidate breast cancer biology at a cellular level by the interpretation of results produced by the developed algorithms. To do so, we planned to assess the correlation between the pathological subgroups generated by AI and the risk of recurrence, paying special attention at the features extracted by the model that could be related to different outcomes.

Therefore, our main objectives were to:

1) evaluate whether AI applied to WSI could predict metastatic relapse at five years,
2) evaluate whether the prognostic elements provided by AI adds supplementary information to routine used clinico-pathological prognostic parameters,
3) decipher the features used by AI to estimate risk (interpretability phase).

Materials and Methods

Patients

All materials used in this project have already been described in Garberis & Gaury et al. (see in Results, Part "Materials, Methods, Design"). Herein, I provide additional details.

A. Patient cohorts

The study involves two cohorts: a training cohort and a validation cohort.

The training cohort is anchored at Gustave Roussy Hospital, Villejuif, France, and includes a great part of the population from the GrandTMA study, which nucleated women treated with surgery and/or chemotherapy (anthracyclines + taxanes), endocrine therapy, trastuzumab, radiation therapy, with less than 5 years of hindsight and associated with a complete clinical database. The main objective of this retrospective study was to construct a Tissue MicroArray (TMA) that could be used for translational research projects.

These patients integrated, at the time, the records of a program called "consultation and assessment in one day", during a first consultation in the Breast Pathology Department. The objective of this consultation, motivated by an abnormality detected in breast/s, is to try to specify the nature of this anomaly within one day. To this prospect, the patient is taken in care by a multidisciplinary medical team (oncologists, surgeons, radiologists, pathologists, all breast specialists) who will together take the necessary decisions to establish a diagnosis, to perform adequate additional radiological explorations and, if necessary, offer a treatment, depending on the different possible orientations after this consultation and the delay of additional exams.

The cohort comprises 1850 women aged 22–95 years, diagnosed with early stage breast cancer (including all types of invasive adenocarcinomas), surgically treated at Gustave Roussy, between October 2005 and December 2013. Patients had a representative tumor specimen (formalin-fixed, paraffin-embedded tumor sample) used for diagnostic purposes and available for further analysis. They were all followed at Gustave Roussy according to the clinical protocol (planned by the study). Of note, most of the patients were transferred to another facility for follow-up after five years; therefore, we decided to assess the risk only at five years.

The validation cohort was extracted from a dataset from CANTO, a French observational and prospective study. Additional information can be retrieved in (Vaz-Luis et al., 2019).

B. Privacy considerations

As this is a retrospective study relating solely to data usually collected for healthcare, this work was carried out in accordance with the provisions of the Public Health Code applicable to research not involving the human person (Public Health Code - Article R1121-1 amended by Decree no. 2017-884, May 9th, 2017) and therefore it does not come under the jurisdiction of a Committee for the Protection of Persons. It obtained the favorable opinion of the expert Committee for Research, Studies and Evaluations in the field of breast pathology, as well as of the Ethics Committee (Data Protection Office, Gustave Roussy). It has been submitted to the National Commission for Computing and Liberties (CNIL) under reference 2225201 v 0 and has been declared in accordance with the reference methodology MR-004. The patients involved were informed of the research via an information letter distributed by post with the possibility of opposing the study within three weeks from the date of dispatch.

Methods

All methods used in this project have already been described in Garberis & Gaury et al. (see in Results, Part "Materials, Methods, Design"). Herein, I provide additional details.

A. Slide scanning

All glass slides selected for this study were stocked at the Pathology Department, as part of routine clinical care and are thus of diagnostic quality. As a quality-control measure, all slides were inspected manually before (for cleaning and re-mounting when necessary) and after scanning (for re-scanning of out-of-focus images in order to avoid issues that might affect subsequent image analysis).

Scanning was performed on an Olympus V120 scanner (Olympus Corp., Shinjuku, Tokyo, Japan). Slides were scanned at 20x magnification. Automated scanning processes (selection of scanning area, placement of focus points) were quality checked and repeated manually where necessary.

Results

Research Article

"Deep Learning Assessment of Metastatic Relapse Risk from Digitized Breast Cancer Histological Slides".

Manuscript under editorial consideration for publication in Nature Medicine. Submission Date: 7th Feb 23.

Article Type: **Original Article**

Title: Deep Learning Assessment of Metastatic Relapse Risk from Digitized Breast Cancer Histological Slides

Authors: I. Garberis[1*], V. Gaury[2*], C. Saillard[2], D. Drubay[3], K. Elgui[2], B. Schmauch[2], A. Jaeger[2], L. Herpin[2], J. Linhart[2], M. Sapateiro[4], F. Bernigole[4], A. Kamoun[2], E. Bendjebbar[2], A. de Lavergne[2], R. Dubois[2], M. Auffret[2], L. Guillou[2], I. Bousaid[5], M. Azoulay[5], J. Lemonnier[6], M. Sefta[2], A. Jacquet[6], A. Sarrazin[2], J-F Reboud[2], F. Brulport[2], J. Dachary[2], B. Pistilli[7], S. Delaloge[7], P. Courtiol[2], F. André[1,7], V. Aubert[2], M. Lacroix-Triki[4]

Affiliations:

1. INSERM U981, Gustave Roussy, Paris-Saclay University, Villejuif, France.
2. Owkin, Paris, France.
3. Gustave Roussy, Office of Biostatistics and Epidemiology, Université Paris-Saclay, Villejuif, France
   Inserm, Université Paris-Saclay, CESP U1018, Oncostat, labeled Ligue Contre le Cancer, Villejuif, France
4. Department of Pathology, Gustave Roussy, Paris-Saclay University, Villejuif, France.
5. Department of digital transformation and information systems (DTNSI), Gustave Roussy, Villejuif, France.
6. Unicancer R&D, Unicancer, Paris, France.
7. Department of Cancer Medicine, Gustave Roussy, Paris-Saclay University, Villejuif, France.

* These authors contributed equally to this work.

Corresponding author: Dr Ingrid Garberis. INSERM U981, Gustave Roussy, Paris-Saclay University, 114 rue Édouard Vaillant, 94805 Villejuif, France, Te. +33 01 42 11 32 58, Email: ingrid-judith.GARBERIS@gustaveroussy.fr

ABSTRACT

Adapted cancer treatment strategy requires accurate stratification. We developed an artificial intelligence (AI)-based tool that uses digitized tumor slides to assess the 5-years metastasis-free survival (MFS) of patients with estrogen receptor-positive, HER2-negative (ER+/HER2-) early breast cancer (EBC). We developed a Deep Learning model (RlapsRisk™ BC) that independently predicts MFS and added significant prognostic information to clinico-pathological variables (c-index in the validation set 0.80 versus 0.76 for clinico-pathological factors alone, p-value<0.05). A threshold corresponding to a probability of MFS event of 5% at 5 years was applied to dichotomize patients into low or high-risk groups. After dichotomization, combining RlapsRisk BC and clinico-pathological factors showed a higher cumulative sensitivity on the validation dataset (0.76 vs 0.61) for an equal dynamic specificity (0.76) in comparison with the clinical score alone. Expert characterization of the most predictive tissue tiles according to RlapsRisk BC revealed well-known morphological features in prognostic determination and potential new morphological features to be further explored.

INTRODUCTION

Despite significant progress in classification and treatment over the past two decades, breast cancer (BC) remains the leading cause of cancer death for women worldwide (1). Proposing an optimal therapeutic strategy to each patient requires systematic and accurate characterization of each disease. Specifically, estrogen receptor positive (ER+), HER2-negative (HER2-) invasive breast cancer, which accounts for approximately 70% of all invasive BC, is associated with a wide spectrum of outcomes and treatment requirements. For many of these women, a key question remains whether adjuvant chemotherapy with the burden of acute side effects and the potential long-term persistent quality of life (QoL) deterioration (2) can be safely avoided. Furthermore, women with a predicted high risk of metastatic relapse despite current standard treatment could be offered more intensive or extended adjuvant strategies, including the addition of a CDK4/6 inhibitor, or other approaches (3),(4).

Prognosis definition has been traditionally based on clinical and histopathological factors, such as the patient's age and the histological classification and grade. Biomarker assessment mainly by immunohistochemistry (ER, progesterone receptor - PR-, HER2 and the proliferation marker KI67) was added to this estimation and later refined with the inclusion of molecular signatures. Results from the TAILORx trial showed that Oncotype DX®, a 21-gene expression molecular test that assesses the 10-year metastasis-free survival (MFS), could spare up to 85% of women with early estrogen receptor-positive, HER2-negative (ER+/HER2-) breast cancer (EBC) unnecessary adjuvant chemotherapy, without impacting patient outcomes (5),(6),(7). Currently, several gene expression signatures assessed on the primary tumor material are endorsed by international guidelines to support clinicians in refining the prognosis of patients with EBC and taking adjuvant treatment decisions (8). Beside this molecular characterization, prognostic tools using classical factors and embedded into publicly available websites may be used as an aid in clinical decision making. For instance, Predict Breast Cancer, a widely used online prognostication software, uses known prognostic factors such as tumor size, KI67 index, tumor grade and lymph node status to predict overall survival at 5 and 10 years (9),(10). However, sensitive markers such as Ki67 may be subject to reproducibility and expertise biases (11),(12),(13),(14),(15).

The characterization and prognostication of diseases has evolved over time and recently has expanded to include more sophisticated instruments and methods from the computational field. Artificial intelligence (AI), particularly machine learning (ML) approaches, are increasingly being developed to answer biological and clinical questions. Notably, recent studies have shown the potential of deep learning (DL) models applied to histopathological whole slide images (WSI) to predict patient outcome and unveil features correlated with prognosis in different malignancies, such as brain tumors (16), mesothelioma (17), colorectal cancer (18) and breast cancer (19).

In this study, we aimed to investigate whether Artificial Intelligence (AI) applied on tumor WSI could: (i) identify patients who have a substantial risk of metastatic relapse despite receiving standard treatments, and (ii) provide additional prognostic information beyond clinico-pathological prognostic criteria. The ultimate goal was to develop an AI-based digital pathology tool to allow assessment of risk of metastatic relapse.

RESULTS

The primary objective of this study was to evaluate the additional 5-years MFS prognostic value of RlapsRisk BC score relative to that of the current clinico-pathological criteria, in patients with ER+/HER2- early breast cancer. The secondary objective consisted in comparing the capacity of a model combining standard clinico-pathological criteria and RlapsRisk BC to dichotomize patients between high risk and low risk of developing 5-years MFS events to that of a model based on clinical factors only. This comparison was assessed on the entire population of the validation cohort and in different subgroups of clinical interest (histological grade 2, clinical intermediate risk of relapse as defined in supplementary table 2, pre- versus post-menopausal status, with or without lymph node invasion, treated with or without adjuvant chemotherapy).

Model development and datasets

To build our model, we used the GrandTMA cohort as a discovery dataset. This cohort has been collected retrospectively from patients diagnosed in the “One-stop breast

clinic" program and treated at Gustave Roussy Cancer Center (Villejuif, France) between October 10th 2005 and February 7th 2013 (20). The cohort comprised 1802 patients diagnosed with early invasive BC (1429 ER+/HER2-, 110 ER+/HER2+, 70 ER-/HER2+ and 193 ER-/HER2- tumors) who underwent surgery as first treatment and had at least one available hematoxylin-eosin-saffron (HES)-stained tumor slide from surgery specimen. For the purpose of a one-shot external validation of our model, we used a dataset from the French observational and prospective CANTO cohort (NCT01993498) (21). Out of 14,000 patients accrued in CANTO so far, 1703 ER+/HER2- EBC patients had a minimum follow-up of 5 years and were eligible for the present study. HES slides and full clinico-pathological features were exploitable from 889 patients. None of these patients were also included in the GrandTMA cohort. See supplementary methods for additional information on the two cohorts.

Our approach consisted in developing an algorithm that learned from the WSIs to predict the 5-years metastasis-free survival (MFS) without any local annotation provided by pathologists. To build this model we used a method composed of three steps: i) tissue tiling, ii) feature extraction, iii) creation of a risk score (Figure 1). All these steps are detailed in the supplementary methods.

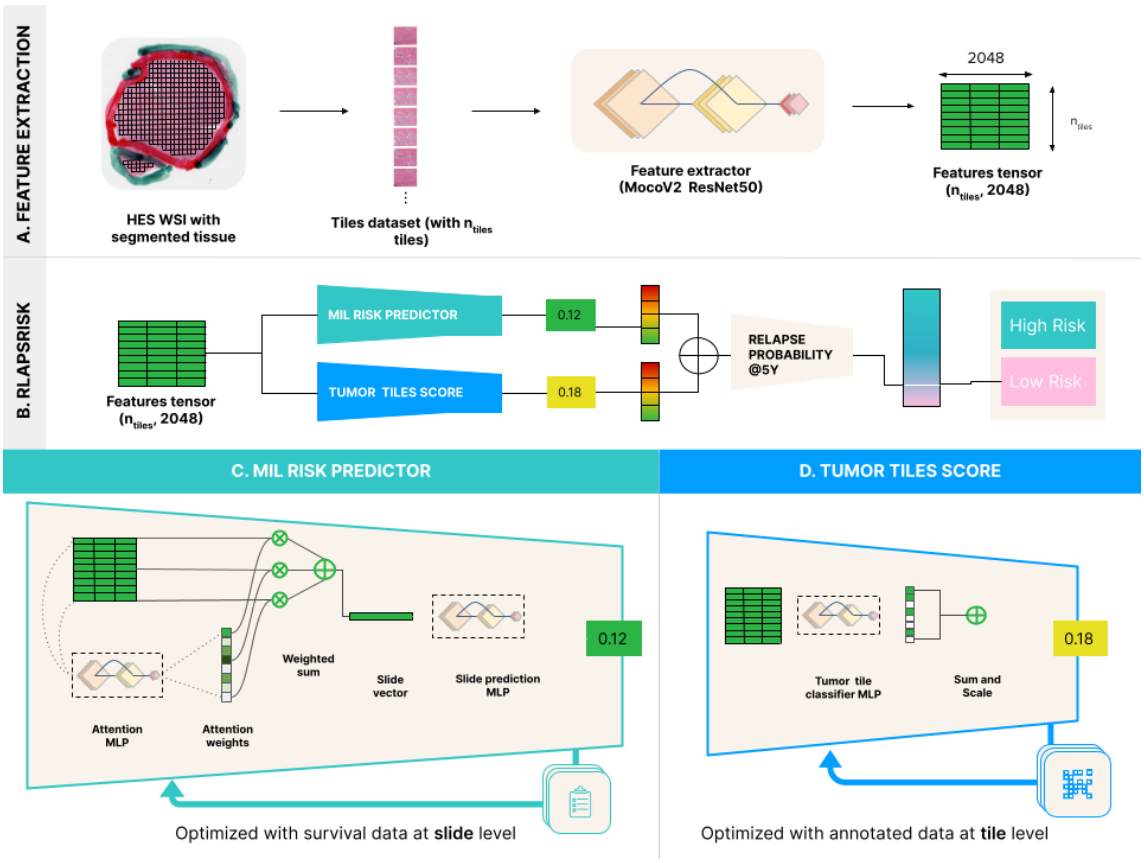


Figure 1: RlapsRisk BC algorithm overview

We then developed a clinical score to predict the 5-year MFS from a multivariable Cox model trained using the discovery dataset (hereafter the Clinical Score), based on the following clinical variables: age, tumor size, histological grade, lymph node invasion and Ki67 expression. We considered this score as our baseline reference that we compared to a score based on a Cox Model adjusted for the clinical factors and RlapsRisk BC score to assess the additional predictive value of the RlapsRisk BC score to the standard clinical factors (hereafter the Combined Score). Adjuvant treatment regimen was not a prognostic factor in multivariable analyses (both with or without RlapsRisk BC score) and was therefore removed from all prognostic models.

RlapsRisk BC score and Prognosis

On both GrandTMA and CANTO cohorts, RlapsRisk BC score was an independent prognostic factor of MFS (Table 1) when integrated into a multivariable Cox model

including histological grade, age, lymph node invasion, tumor size, and Ki67 expression. All variables, except histological grade, were considered continuous.

Grand TMA	Variable	Cox Model without RlapsRisk BC		Cox Model with RlapsRisk BC	
		Unit HR (95% CI)	***P***	**Unit HR (95% CI)**	***P***
Multivariable Cox model	Histological grade 1	1 (ref)	N.A	1 (ref)	N.A
	Histological grade 2	1.85 (1.05–3.26)	0.03	1.90 (1.06–3.40)	0.03
	Histological grade 3	2.13 (1.13–4.00)	0.02	2.19 (1.15–4.18)	0.02
	Age	1.40 (1.13–1.73)	<0.005	1.39 (1.13–1.71)	<0.005
	Lymph Node Invasion	1.13 (1.08–1.15)	<0.005	1.12 (1.08–1.16)	<0.005
	Tumor Size	1.33 (1.16–1.51)	<0.005	1.18 (1.00–1.40)	0.05
	Ki67 expression	1.56 (1.32–1.85)	<0.005	1.39 (1.16–1.66)	<0.005
	RlapsRisk score	N.A.	N.A.	1.38 (1.19–1.59)	<0.005
	C-index (cross-validation)	0.78 (+/- 0.04)		0.80 (+/- 0.04)	

Table 1A: Multivariable Cox proportional hazard models estimating the contribution of several prognostic variables on MFS on GrandTMA. P columns contain the probability of observing these results if the Unit HR were equal to 1.

CANTO	Variable	Cox Model without RlapsRisk BC		Cox Model with RlapsRisk BC	
		Unit HR (95% CI)	*P*	Unit HR (95% CI)	*P*
Multivariable Cox model	Histological grade1	1 (ref)	N.A	1 (ref)	N.A
	Histological grade 2	1.14 (0.80–1.63)	0.47	1.16 (0.81–1.66)	0.41
	Histological grade 3	1.39 (0.90–2.14)	0.14	1.28 (0.83–1.97)	0.27
	Age	1.03 (0.87–1.21)	0.77	1.03 (0.87-1.22)	0.33
	Lymph Node Invasion	1.13 (1.05–1.21)	<0.005	1.11 (1.04–1.18)	<0.005
	Tumor Size	1.17 (1.02–1.36)	0.03	1.13 (0.98–1.31)	0.10
	Ki67 expression	1.14 (0.98–1.33)	0.08	1.24 (0.98–1.33)	0.09
	RlapsRisk BC score	N.A	N.A	1.26 (1.08–1.46)	<0.005
	C-index (external validation)*	0.76 (+/- 0.037)		0.80 (+/- 0.035)	

Table 1B: Multivariable Cox proportional hazard models estimating the contribution of several prognostic variables on MFS on CANTO. P columns contain the probability of observing these results if the Unit HR were equal to 1.
** C-index were computed from the Cox multivariable models trained on Grand TMA.*

The discrimination power of these two scores was first compared using the Harell C-index (22) on the discovery dataset with a stratified fivefold cross-validation strategy, and three repeats. Due to the small number of events, we stratified this cross-validation on the events to preserve a minimal number of events per fold. Data from the CANTO cohort were held out from the training series and were used only for a one-shot external validation and the assessment of the discrimination of each model. RlapsRisk BC increased the model discrimination when added to the clinical factors in the CANTO validation dataset, with a Harell C-index of 0.80 compared to 0.76 for the clinical factors alone (+0.04, p-value <0.05).

We then assessed the prognostic performance of RlapsRisk BC score and the clinical factors in patients subgroups defined by standard clinico-pathological factors, by the

clinical risk groups (Supplementary Table 2) and by adjuvant treatment regimen. The assessment of potential heterogeneities in these subgroups was conducted by Cox regression analyses. When a factor was used to build a subgroup it was removed from the associated Cox multivariable model. A higher RlapsRisk BC score was associated with an increased risk of distant recurrence in all subgroups except for patients with pN2, pN3 and pT3 TNM stages (see Figure 2), which were subgroups with limited sample sizes.

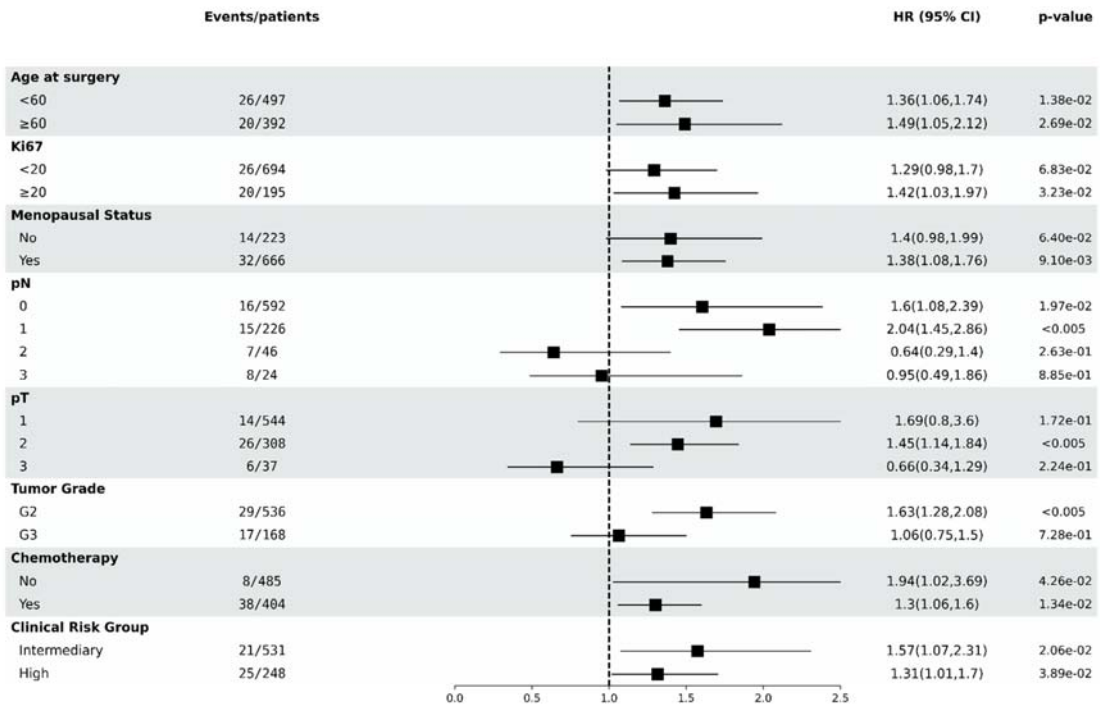


Figure 2: Forest plot of the adjusted RlapsRisk BC score HRs on prediction of 5-year metastasis-free survival. Each square of the forest plot represents the HR of the adjusted RlapsRisk BC score (a continuous variable) + clinical factors in the subgroup of patients defined by the variable category in the first column of the table. The 95% HRs confidence intervals are represented by the horizontal lines. Histological tumor grade 1 and low clinical risk Group were removed as no MFS events were recorded in these subgroups. P columns contain the probability of observing these results if the Unit HR were equal to 1.

Clinical validity of RlapsRisk BC

To compare the capacity of the models to dichotomize patients between high risk and low risk of developing 5-years MFS events, thresholds corresponding to a probability of MFS event of 5% at 5 years were set for each risk score in the training set and prespecified accordingly for validation (see supplementary methods for details). When

dichotomized, the scores were then referred to as “classifiers” (e.g. combined model classifier).

After applying the MFS risk stratification according to the previously defined classifiers, Kaplan–Meier analyses showed significant differences in distant recurrence events between low-risk and high-risk patients both in the discovery and validation datasets, as summarized on Table 2 and highlighted on Figure 3.

	CANTO validation cohort (N=889)		
	RlapsRisk BC Classifier	**Clinical score classifier**	**Combined model classifier**
Number at Low Risk*	562	663	658
Number at High Risk**	327	226	231
% of patients with MFS events in the Low Risk group	1,42%	1,96%	1,22%
% of patients with MFS events in the High Risk group	7,95%	9,29%	11,26%
Kaplan Meier’s Hazard Ratio	4.36	4.25	6.99
95% CI	2.32–8.18	2.36-7.64	3.73-13.09
p-value	<0.005	<0.005	<0.005

Table 2: Classification of Patients according to each classifier (RlapsRisk BC, Clinical score, and Combined)
** Predicted probability of 5-years MFS event ≤5%*
***Predicted probability of 5-years MFS event >5%*

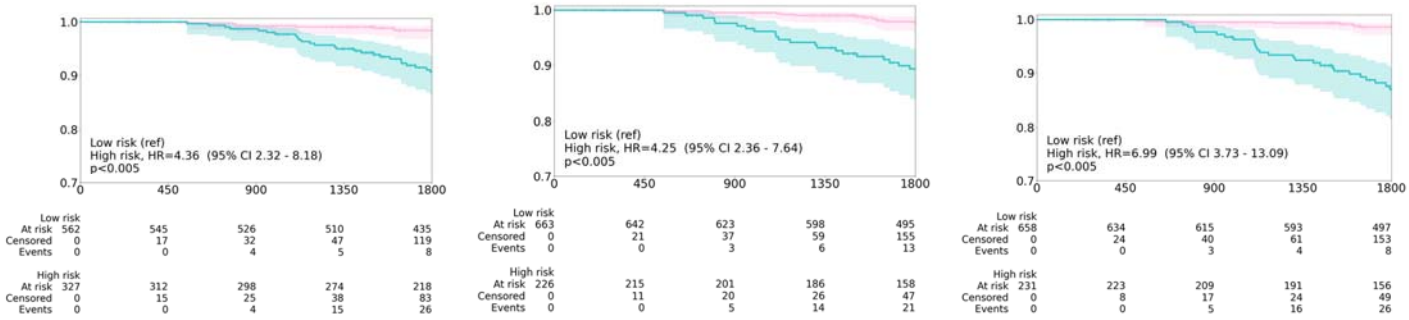


Figure 3: Metastases-free survival of patients stratified according to RlapsRisk BC classifier (left) Clinical score classifier (middle) and RlaspRisk + clinical score combined model classifier (right) among patients from the CANTO validation cohort. Numbers in parentheses indicate the 95% CI of the

HR.

We also assessed the performance of our tool across groups of patients at different risk of recurrence, as defined according to known prognostic factors, such as menopausal status, presence of lymph node invasion vs. not, treatment by chemo-endocrine therapy vs. endocrine therapy alone. The discriminative power of the combined model classifier in those subgroups (Table 3) suggests that, when combined with the current clinical factors RlapsRisk BC could be used as an additional layer of information for better defining the risk of recurrence of each patient. Indeed, on the full CANTO dataset, with a common dynamic specificity of 0.76, the combined model Classifier increased the cumulative sensitivity by 15 points in comparison to the clinical score classifier alone. When looking at the histological grade 2 subgroup and the "intermediate clinical risk" subgroup (defined in the supplementary table 2), with a dynamic specificity equivalent and a gain of 28 and 26 points respectively in cumulative sensitivity, the combined model classifier largely outperformed the clinical score classifier. These figures illustrate the benefit of adding RlapsRisk BC score to the current clinical factors for a better estimation of prognosis within subgroups with a difficult prognosis estimation.

Interestingly, the score combining RlapsRisk and the clinical variables is the only model achieving a stable performance for both pre- and post-menopausal patients.

Subgroup		RlapsRisk BC classifier		Clinical score classifier		Combined model classifier	
	N	Cumulative sensitivity	Dynamic specificity	Cumulative sensitivity	Dynamic specificity	Cumulative sensitivity	Dynamic specificity
Full Population	889	0.77 [0.61 – 0.87]	0.67 [0.63 – 0.70]	0.61 [0.45 – 0.75]	0.76 [0.72 – 0.79]	0.76 [0.61 – 0.87]	0.76 [0.73 – 0.80]
Patients with histological grade 2 carcinoma	534	0.80 [0.59 – 0.91]	0.68 [0.63 – 0.72]	0.46 [0.28 – 0.66]	0.79 [0.74 – 0.83]	0.74 [0.53 – 0.87]	0.79 [0.75 – 0.83]
Intermediate clinical risk group	529	0.84 [0.58 – 0.95]	0.73 [0.68 – 0.79]	0.41 [0.20 – 0.66]	0.82 [0.78 – 0.86]	0.67 [0.41 – 0.85]	0.84 [0.80 – 0.87]
Patients with Node positive disease	296	0.71 [0.53 – 0.85]	0.5 [0.43 – 0.57]	0.71 [0.52 – 0.85]	0.62 [0.55 – 0.68]	0.8 [0.61 – 0.90]	0.56 [0.49 – 0.63]
Patients with N0 disease	593	0.90 [0.62 – 0.98]	0.74 [0.70 – 0.78]	0.38 [0.17 – 0.65]	0.82 [0.78 – 0.85	0.69 [0.40 – 0.88]	0.85 [0.81 – 0.88]
Pre-Menopausal patients	223	0.81 [0.52 – 0.94]	0.57 [0.65 – 0.79]	0.50 [0.25 – 0.75]	0.83 [0.77 – 0.88]	0.72 [0.43 – 0.90]	0.79 [0.72 – 0.84]
Post-Menopausal patients	666	0.75 [0.56 – 0.87]	0.70 [0.66 – 0.73]	0.66 [0.47 – 0.80]	0.73 [0.69 – 0.77]	0.78 [0.60 – 0.90]	0.75 [0.71 – 0.79]
Patients who received adjuvant CT	405	0.75 [0.57 – 0.87]	0.48 [0.43 – 0.54]	0.66 [0.49 – 0.80]	0.60 [0.54 – 0.65]	0.78 [0.61 – 0.89]	0.57 [0.51 – 0.62]
Patients who did not received adjuvant CT	484	0.85 [0.52 – 0.97]	0.80 [0.76 – 0.84]	0.41 [0.16 – 0.71]	0.88 [0.84 – 0.91]	0.70 [0.38 – 0.90]	0.91 [0.87 – 0.93]

Table 3: Performance of the Classifiers in different subgroups of the CANTO cohort (external validation)

Additional exploratory Kaplan–Meier analyses highlighted significant differences between high-risk and low-risk in all these subgroups (see supplementary material Figures 3 to 7).

Model interpretability: features assessment

To identify and characterize areas of slides that most impact the overall risk score given by the model, we computed the marginal contribution of each tile to the overall risk score to assess its positive or negative contribution to the final risk score predicted by the final multilayer perceptron (MLP). The extremal tiles were further reviewed and annotated by two expert pathologists blinded to the predicted outcome (see supplementary methods for more details). The most predictive features of high risk of relapse according to RlapsRisk BC Classifier (Figure 4B) were the presence of high tumor cell content ($P < 0.0001$), high degree of nuclear pleomorphism ($P < 0.0001$), massive architecture ($P < 0.0001$) with low tubule formation ($P < 0.0001$) and trabecular structures ($P < 0.0001$) as well as mitotic activity ($P < 0.0001$). Multiple

features were associated with low risk of relapse such as fibrosis (P < 0.0001), presence of vascular structures (P < 0.0001) and isolated tumoral cells (P = 1.2E-03). Interestingly, tiles associated with low risk (Fig. 4A) were not only located in tumor stroma (P < 0.0001), but also in the adjacent normal tissue (P < 0.0001). Moreover, the model was also able to capture complex cell interactions, as spatial mixing of fibroblasts and tumor infiltrating lymphocytes (TILs) were also retrieved in low risk tiles (P = 2.5E-04), a feature of good prognosis highlighted in a recent study (25). These results showed that our model was able to capture well established prognostic factors as well as complex and novel patterns to predict patient outcomes.

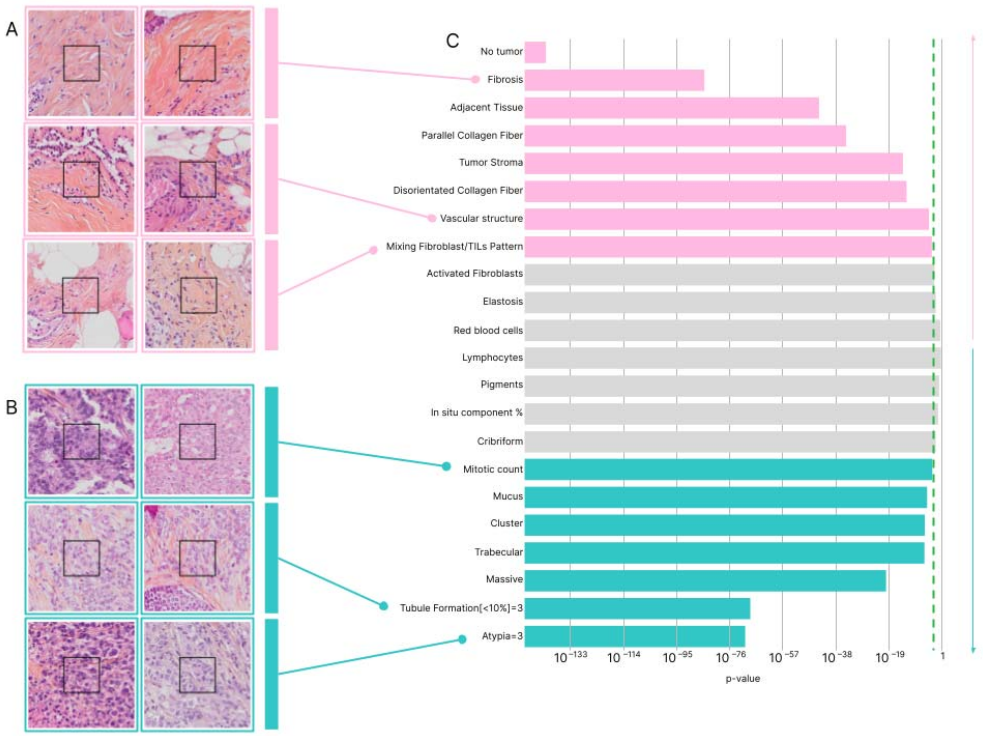


Figure 4: Model Interpretability. A. Tiles exhibiting features associated with a low risk of relapse; B. Tiles exhibiting features associated with a high risk of relapse; C. Annotated histological features.

DISCUSSION

The integration of digital pathology is increasing in everyday's practice, and workflows that include digitization of glass slides are no longer an exception in pathology laboratories. This ongoing transition is paving the way for the implementation of AI-based digital pathology medical devices in clinical practice.

In this study, we developed and validated a new digital pathology score, which predicts distant recurrence at 5 years after surgery in adequately-treated early-stage breast cancer patients, bearing ER-positive and HER2-negative tumors. Our method uses just a standard-stained and scanned tumor slide already available for diagnostic purposes in the pathology laboratory.

RlapsRisk BC score has a strong, validated, prognostic value that is independent of established clinicopathologic factors. It provides additional information beyond classical clinico-pathological factors. When combined together, clinico-pathological factors and RlapsRisk BC were able to dichotomize patients into low-risk and high-risk groups with a strong discriminative power, both on the entire population and within subgroups of interest.

Contrary to other digital pathology approaches predicting morphological features, such as histological grade or KI67 index that are ultimately indirectly linked to prognosis (26),(27),(28),(29), we trained our Deep Learning model to directly predict the 5-year MFS. Not only did we not require local annotations to train the algorithm which takes as input the entire WSI, but our prediction task was directly addressing our clinical question, i.e. prediction of 5-years MFS.

A different approach presented by Wang et al. (30) attained a risk prediction from Nottingham histological grade through the re-stratification of the intermediate category (NHG2). NHG2, which encompassed the largest group of patients, was dichotomized in a low-grade subgroup and a high-grade subgroup, achieving a HR of 2.94 (95% CI 1.24-6.97, P = 0.015) for the stratification between the two groups according to Recurrence Free Survival (RFS). While the presented tool had an important prognostic implication, it was centered on the intermediate risk group and it addressed the question of recurrence risk in an indirect manner (30). However, the strength of their approach is served by the straightforward explicability of the model. Highlights of simple interpretability features will ensure the pathologist that a given classification was performed according to expected criteria: tubule formation, nuclear pleomorphism and mitotic activity, in the case of histological grading. Similarly, in our study, RlapsRisk BC Classifier achieved on the validation cohort a high discriminative power on NHG2

patients with a sensitivity of 0.8 and a specificity 0.68, respectively, and illustrated by an HR of 4.15 (95% CI 1.88-9.15, P < 0.005) for MFS (see supplementary material) despite the use of a different endpoint which excludes locoregional relapse, contralateral tumors and death contrary to RFS.

Interpretability is indeed a valid concern for the use of AI in digital pathology, presented as a crucial challenge by Duggento et al. (31). In our study, we aimed to investigate the features that supported a low-risk or high-risk prediction from RlapsRisk BC. Analyzing the most predictive tiles, we confirmed that the model retrieves well-established criteria alongside complex and novel histological patterns that have an impact on patient outcomes, such as high atypia or mitotic activity, without the need for locally annotated data. In this manner, we propose a natively interpretable model to step out of the black box paradigm, a major hurdle toward a broad adoption of any AI solution in clinical practice to help clinicians tackle medical challenges (32).

Currently, one of these challenges in ER+/HER2- EBC management is the adaptation of the treatment according to the risk of the patient. Reducing the number of unnecessary adjuvant chemotherapies or even endocrine therapies to improve quality of life while maintaining an equivalent survival rate of the patients is the key challenge. This group is currently the target population for molecular signatures, where the issue is to use the genomic score to decide on an adjuvant chemotherapy or to decide on a de-escalation in certain patients. However, the restrictive indications for use limit the eligible population of molecular tests, their use is not generalized, and they are not covered by the health care system in many countries (33). These facts expose a gap where a simpler, less-expensive, routinely available tool, could support decision-making, replace molecular signatures or, at least, work as a pre-screening test for ulterior use of onerous molecular determinations in a subset of cases. This approach has already been explored in some studies (34), and a comparison study between molecular signatures and RlapsRisk in a similar setting to the TRANSATAC study (35) would provide helpful insights on this matter.

Even though our study presents an innovative and useful tool for patient stratification, there are some limitations.

Regarding the preanalytical phase, our model was fitted for HES-stained tumor slides,

and has not yet been validated on HE-stained histopathological slides, despite it being the staining of choice in pathology laboratories outside of France. Adaptations of the model for an optimal application on HE-stained slides and other routine stainings are currently under development. As for the scanning of the slides, only two different scanners were used for the digitization.

In addition, the population of the validity set derives for the majority from the same center, where data for the training part were collected. On those issues, assorted data from novel external centers is being collected and included in future validation cohorts, which will allow us to confirm the medical utility of RlapsRisk BC, as recommended in (36).

In conclusion, RlapsRisk BC™ resulted to be an independent prognostic factor of MFS and added significant prognostic information to clinico-pathological variables. After patients dichotomization into low-risk and high-risk groups, RlapsRisk combined with classical clinico-pathological risk factors showed higher discrimination power compared to clinico-pathological risk factors alone. A prospective observational study comparing RlapsRisk BC to molecular prognostic signatures is currently ongoing and will allow to determine the impact of the implementation of this AI-based tool into the practice workflow. Future extensions of our research include the development of novel algorithms, adapted to a broader variety of inputs. To deepen interpretability issues, an exhaustive analysis of tiles is contemplated with a focus on the spatiality notion that could provide novel insights in tumor biology.

Data Availability

The GrandTMA dataset that supports the findings of this study is available from Gustave Roussy but restrictions apply to the availability, which were used with permission for the current study, and so are not publicly available. The CANTO dataset used for external validation is available from UNICANCER but restrictions apply to the availability of data, which were used with permission for the current study, and so are not publicly available. The datasets, or a test subset, may be available from Gustave Roussy or UNICANCER subject to ethical approvals.

Code Availability

The code used for training the models has a large number of dependencies on internal tooling and its release is therefore not feasible. However, all experiments and implementation details are described thoroughly in the Online Methods section so that it can be independently replicated with non-proprietary libraries.

ACKNOWLEDGEMENTS

We thank the patients who participated in both studies and who did not oppose this additional research. We would like to acknowledge the support provided by Région Île de France and the impulsion given to this study by organizing the AI for Health Data Challenge in 2019.

FUNDING

The first phase of this work was supported by funding from the Region Ile-de-France in the framework of “AI for Health Data Challenge 2019”.

DISCLOSURE

VG, CS, KE, BS, AJ, LH, RD, MA, LG, MS, AS, JR, FB, JD, VA are employees of Owkin Inc.

SD reports grants and non-financial support from Pfizer, grants from Novartis, grants and non-financial support from AstraZeneca, grants from Roche Genentech, grants from Lilly, grants from Orion, grants from Amgen, grants from Sanofi, grants from Exact Sciences, grants from Servier, grants from MSD, grants from BMS, grants from Pierre Fabre, grants from Exact Sciences, grants from Besins, grants from European Commission grants, grants from French government grants, grants from Fondation ARC grants, grants from Taiho, grants from Elsan, outside the submitted work.

FA declares institutional financial interests, research grants with Novartis, Pfizer, AstraZeneca, Eli Lilly, Daiichi, Roche, Sanofi.

BP reports Consulting fees from Astra Zeneca (institutional), Seagen (institutional), Gilead (institutional), Novartis (institutional), Lilly (institutional), MSD (institutional),

Pierre Fabre (personal), Daiichi-Sankyo (institutional/personal); Research funding (institutional) from Astra Zeneca, Daiichi-Sankyo, Gilead, Seagen, MSD, Fondation ARC.

Travel support: Astra Zeneca; Pierre Fabre; MSD ; Daiichi-Sankyo.

MLT reports Consulting fees from Astra Zeneca (institutional/personal), Seagen (personal), Lilly (personal), MSD (institutional/personal), Pierre Fabre (personal), Daiichi-Sankyo (institutional/personal), Myriad Genetics (personal), Exact Sciences (personal), Roche Diagnostics ((institutional/personal); Research funding (institutional) from Roche Diagnostics, Daiichi-Sankyo and Pierre Fabre

Travel support: Astra Zeneca, Seagen, Daiichi-Sankyo.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. mai 2021;71(3):209-49.
2. Ferreira AR, Di Meglio A, Pistilli B, Gbenou AS, El-Mouhebb M, Dauchy S, et al. Differential impact of endocrine therapy and chemotherapy on quality of life of breast cancer survivors: a prospective patient-reported outcomes analysis. Ann Oncol Off J Eur Soc Med Oncol. 1 nov 2019;30(11):1784-95.
3. Mastro LD, Mansutti M, Bisagni G, Ponzone R, Durando A, Amaducci L, et al. Extended therapy with letrozole as adjuvant treatment of postmenopausal patients with early-stage breast cancer: a multicentre, open-label, randomised, phase 3 trial. Lancet Oncol. 1 oct 2021;22(10):1458-67.
4. Harbeck N, Rastogi P, Martin M, Tolaney SM, Shao ZM, Fasching PA, et al. Adjuvant abemaciclib combined with endocrine therapy for high-risk early breast cancer: updated efficacy and Ki-67 analysis from the monarchE study. Ann Oncol. 1 déc 2021;32(12):1571-81.
5. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. N Engl J Med. 19 nov 2015;373(21):2005-14.
6. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. N Engl J Med. 12 juill 2018;379(2):111-21.
7. Sparano JA, Gray RJ, Ravdin PM, Makower DF, Pritchard KI, Albain KS, et al. Clinical and Genomic Risk to Guide the Use of Adjuvant Therapy for Breast Cancer. N Engl J Med. 20 juin 2019;380(25):2395-405.
8. Oncotype DX Tests [Internet]. [cité 24 oct 2022]. Disponible sur: https://www.breastcancer.org/screening-testing/oncotype-dx
9. Wishart GC, Bajdik CD, Dicks E, Provenzano E, Schmidt MK, Sherman M, et al. PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. Br J Cancer. août 2012;107(5):800-7.
10. Wishart GC, Bajdik CD, Azzato EM, Dicks E, Greenberg DC, Rashbass J, et al. A

population-based validation of the prognostic model PREDICT for early breast cancer. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol. mai 2011;37(5):411-7.
11. Polley MYC, Leung SCY, McShane LM, Gao D, Hugh JC, Mastropasqua MG, et al. An international Ki67 reproducibility study. J Natl Cancer Inst. 18 déc 2013;105(24):1897-906.
12. Casterá C, Bernet L. HER2 immunohistochemistry inter-observer reproducibility in 205 cases of invasive breast carcinoma additionally tested by ISH. Ann Diagn Pathol. avr 2020;45:151451.
13. Gown AM. Current issues in ER and HER2 testing by IHC in breast cancer. Mod Pathol Off J U S Can Acad Pathol Inc. mai 2008;21 Suppl 2:S8-15.
14. Predict Breast [Internet]. [cité 24 oct 2022]. Disponible sur: https://breast.predict.nhs.uk/
15. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res BCR. 2010;12(1):R1.
16. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A. 27 mars 2018;115(13):E2970-9.
17. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med. oct 2019;25(10):1519-25.
18. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. Npj Digit Med. 19 avr 2021;4(1):1-13.
19. Ibrahim A, Gamble P, Jaroensri R, Abdelsamea M, Mermel C, Chen PH, et al. Artificial Intelligence in Digital Breast Pathology: Techniques and Applications. The Breast. 1 déc 2019;49.
20. Delaloge S, Bonastre J, Borget I, Garbay JR, Fontenay R, Boinon D, et al. The challenge of rapid diagnosis in oncology: Diagnostic accuracy and cost analysis of a large-scale one-stop breast clinic. Eur J Cancer Oxf Engl 1990. oct 2016;66:131-7.
21. Vaz-Luis I, Cottu P, Mesleard C, Martin AL, Dumas A, Dauchy S, et al. UNICANCER: French prospective cohort study of treatment-related chronic toxicity in women with localised breast cancer (CANTO). ESMO Open. 2019;4(5):e000562.
22. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 14 mai 1982;247(18):2543-6.
23. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol. oct 2017;14(10):595-610.
24. Uno H, Cai T, Tian L, Wei LJ. Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models. J Am Stat Assoc. 2007;102(478):527-37.
25. Nederlof I, Hajizadeh S, Sobhani F, Raza SEA, AbdulJabbar K, Harkes R, et al. Spatial interplay of lymphocytes and fibroblasts in estrogen receptor-positive HER2-negative breast cancer. Npj Breast Cancer. 28 avr 2022;8(1):1-9.
26. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer. 2018;4:30.
27. Koopman T, Buikema HJ, Hollema H, de Bock GH, van der Vegt B. Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement. Breast Cancer Res Treat. 1 mai 2018;169(1):33-42.
28. Stålhammar G, Robertson S, Wedlund L, Lippert M, Rantalainen M, Bergh J, et al.

Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. Histopathology. mai 2018;72(6):974-89.
29. Jaroensri R, Wulczyn E, Hegde N, Brown T, Flament-Auvigne I, Tan F, et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. Npj Breast Cancer. 4 oct 2022;8(1):1-12.
30. Wang Y, Acs B, Robertson S, Liu B, Solorzano L, Wählby C, et al. Improved breast cancer histological grading using deep learning. Ann Oncol Off J Eur Soc Med Oncol. janv 2022;33(1):89-98.
31. Duggento A, Conti A, Mauriello A, Guerrisi M, Toschi N. Deep computational pathology in breast cancer. Semin Cancer Biol. juill 2021;72:226-37.
32. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 1 févr 2019;267:1-38.
33. SENORIF_Référentiel_francilien_de_pathologie_mammaire_en_collaboration_avec_l'institut_Curie_et_Gustave_Roussy [Internet]. calameo.com. [cité 24 oct 2022]. Disponible sur: https://www.calameo.com/read/004021827f069bd672789
34. Whitney J, Corredor G, Janowczyk A, Ganesan S, Doyle S, Tomaszewski J, et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. BMC Cancer. 30 mai 2018;18(1):610.
35. Sestak I, Buus R, Cuzick J, Dubsky P, Kronenwett R, Denkert C, et al. Comparison of the Performance of 6 Prognostic Signatures for Estrogen Receptor–Positive Breast Cancer: A Secondary Analysis of a Randomized Clinical Trial. JAMA Oncol. 1 avr 2018;4(4):545-53.
36. Use of Archived Specimens in Evaluation of Prognostic and Predictive Biomarkers | JNCI: Journal of the National Cancer Institute | Oxford Academic [Internet]. [cité 25 nov 2022]. Disponible sur: https://academic.oup.com/jnci/article/101/21/1446/964215

Methods

Datasets description

GrandTMA

To build our models we used a discovery dataset collected retrospectively from patients treated at Gustave Roussy in France and included in the “GrandTMA” cohort. This cohort comprises all patients newly diagnosed with a breast carcinoma as part of the “One-stop breast clinic” program at Gustave Roussy between October 10th 2005 and February 7th 2013 (20). The inclusion criteria for the present study were: i) diagnosis of invasive breast carcinoma, with or without associated in situ carcinoma, ii) any type of treatment but neoadjuvant chemotherapy, iii) availability of a surgical specimen with a formalin-fixed, paraffin-embedded (FFPE) tumor sample available, iv) complete clinical and therapeutic data, v) follow-up over at least 4 years and updated annually. The exclusion criteria were: i) exclusive non-invasive tumors, ii) cytology-only available cases, iii) absence of follow-up, iv) other non-adenocarcinomatous lesions of the breast. This led to the inclusion of 1802 patients diagnosed with early invasive BC (1429 ER+/HER2-, 110 ER+/HER2+, 70 ER-/HER2+, 193 ER-/HER2-), with at least 1 available hematoxylin-eosin-saffron (HES)-stained tumor slide from the surgical specimen at the Pathology Department (details in Supplementary Figure 1). Biomarker status (ER, PR, and Ki67 immunohistochemistry (IHC) expression, and HER2 protein expression/gene amplification) was defined and determined locally according to the current recommendations of the College of American Pathologists and the American Society of Clinical Oncology (CAP/ASCO (37),(38)), and the French recommendations of the Study Group on Immunohistochemical Prognostic Factors in Breast Cancer (GEFPICS (39)). ER and PR expression positivity was defined as an IHC staining of at least 10% of tumor cells, as is standard in several European countries. Lesions were considered positive for HER2 (score 3+) if the number of tumor cells with a complete and intense membrane IHC staining exceeded 10% of the whole invasive tumor cells population; equivocal (score 2+) if the number of tumor cells showing a complete and moderately intense membrane IHC staining, or an incomplete basolateral membrane IHC staining of moderate to severe intensity exceeded 10% of the total infiltrating tumor population; and negative in the remaining cases (scores 0 and 1+). When

dichotomized, Ki67 cut-off was defined according to the current recommendations (i.e. cut-off set at 20%) (38), (40).

Image acquisition was performed using an Olympus VS120 slide scanner at 20x magnification. In order to avoid scanning issues that might affect subsequent image analysis, all slides were checked by a pathologist after digitization to discard slides with insufficient quality and rescanned when necessary (blurred images, need of slides re-mounting when coverslip was damaged).

This work was carried out in accordance with the provisions of the Public Health Code applicable to research not involving the human person (Public Health Code - Article R1121-1 amended by Decree no. 2017-884, May 9th, 2017) and therefore it does not come under the jurisdiction of a Committee for the Protection of Persons. It obtained the favorable opinion of the expert Committee for Research, Studies and Evaluations in the field of breast pathology, as well as of the Ethics Committee (Data Protection Office, Gustave Roussy). It has been submitted to the National Commission for Computing and Liberties (CNIL) under reference N° F20220121170839 and has been declared in accordance with the reference methodology MR-004. The patients involved were informed of the research via an information letter distributed by post mail with the possibility of opposing the study.

CANTO

For the purpose of an external validation of our model, we used a dataset from the French observational and prospective CANTO cohort (NCT01993498) (21). In this cohort, patients were included at diagnosis of their invasive breast cancer, before any treatment, following the given criteria: i) women only, ii) aged over 45 years old, iii) HER2- and ER+ (same definition as for the GrandTMA cohort), iv) with a histologically invasive breast cancer diagnosed, v) with no clinical evidence of metastasis at the time of inclusion. Out of 14,000 patients accrued in CANTO so far, 1703 ER+/HER2- EBC patients had a minimum follow-up of 5 years and were eligible for the present study (708 patients from Gustave Roussy and 997 patients from other cancer centers of the UNICANCER group). None of these patients were also included in the GrandTMA cohort. Thirty-one patients from Gustave Roussy had incomplete data and were excluded from the study, resulting in 675 HES slides available with clinical data. From

the other centers of the CANTO cohort, we had access to 214 HES slides from resection used for primary diagnosis (details in Supplementary Figure 2). In total 889 patients had exploitable HES slides, together with full clinico-pathological features (described in Supplementary Table 1).

All data collections were performed in the framework of the CANTO clinical trial (NCT01993498), in compliance with all legal requirements. All patients included in the study were informed through the website https://mesdonnees.unicancer.fr/ on the reuse of their data for a separate objective with the possibility of opposing the study.

Endpoint

The chosen endpoint for survival data analysis was metastasis-free survival (MFS) at five years, defined as the time from initial surgery to occurrence of metastatic event or death before five years. Local relapse or axillary lymph node recurrence events were ignored. Patient's follow-up was censored at the time of contralateral breast cancer, second non-breast primary cancer or last available date of follow-up.

Model Description

To develop our risk score from histology slides we used a method composed of three steps: i) tissue tiling, ii) feature extraction, iii) creation of a risk score. The transformation of the score into a probability of occurrence of a MFS event before five years and the selection of a threshold are two additional steps, detailed in Statistical Analysis.

Tissue segmentation and tiling

Each of the Whole Slide Images (WSI) was first divided into small squares, 76 × 76 micrometers in size (224 × 224 pixels) called "tiles". This tiling was performed by first segmenting the tissue, using a pre-trained U-Net neural network (41) that discarded the background, and artifacts of scanning or preparation. This segmented tissue was then divided into N (ranging from 10,000 to 75,000) tiles.

Feature Extraction

The N tiles were embedded into D-dimensional feature vectors using a pre-trained CNN (Figure 1A). We implemented Momentum Contrast v2 (42), a self-supervised

learning algorithm that improved performance for various prediction tasks in previous studies (43) trained on the Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) dataset (44)). Multiple data augmentations were applied while the model was optimized for 200 epochs (approximately 30 hours) on 16 NVIDIA Tesla V100 Graphics processing units (GPU). This frozen pre-trained algorithm was then used to extract features during training and inference.

Risk prediction using Multiple instance learning

The N feature vectors were then aggregated using a multiple instance learning (MIL) model trained to predict MFS at five years (Figure 1C). We reimplemented the attention based model called DeepMIL proposed by Ilse et al. (45). A linear layer with L neurons (L = 256 here) was applied to the embedded features followed by a Gated Attention layer with L hidden neurons. A multilayer perceptron (MLP) with 128 neurons was then applied to the output. To speed-up training, only a random subset of 8000 tiles per WSI was used, while all tiles of a slide are processed for inference. DeepMIL was trained using an extension of the standard cross-entropy loss used to train survival prediction models with right-censored data (46).

Integrating a tumor-related feature in the algorithm

Our preliminary analyses highlighted that the number of tumor tiles contained in each slide was associated with distant relapse and yet was not captured by the model. We therefore incorporated this feature by classifying all the tiles of a slide as tumor vs non-tumor (Figure 1D). An ensemble of four MLPs trained in a patch-based supervised learning approach was used.

The combination of the predicted DeepMIL risk score and the tumor count score was done by scaling and summing both features, forming the RlapsRisk BC score (Fig. 1B).

Thresholds determination

For the RlapsRisk BC score, as well as the clinical and combined risk scores, we fitted a Weibull AFT (Accelerated Failure Time) model on the discovery dataset to transform risk scores into probabilities of occurrence of MFS event before 5 years. This step was used to identify the threshold of each risk score corresponding to a probability of 5-years MFS event of 5% defined by the Weibull Models. This 5% MFS rate threshold corresponds to the 5-year interpolation of an exponential model from the 10-year MFS

of 10%, which is the most common output of the molecular signatures currently used in clinical practice (23).

Method for interpretability features assessment

Training an AI model on digital slides from diagnosis to predict metastatic relapse is an original approach compared to recent research works in Digital Pathology, that generally predict well known pathological features (e.g. histological grading or KI67 index). However, bypassing human knowledge in the training phase requires even more explanations of the functioning of the model. We detail herein our method to identify and characterize typical areas on a well defined set of slides that had extreme risk scores. Interpretability relies on the possibility to access the relevant information for our model that is contained in input data or learned by the model itself.

In the model, each tile was associated with an attention score that was used in the final weighted average to obtain the input vector for the risk predictor (see Figure 4). However, this score did not provide information about the impact on prognosis of the tile. To overcome this limitation, we computed the Shapley value (47) associated with each tile, which measures the marginal contribution of each tile to the overall risk score to assess its positive or negative contribution to the final risk score predicted by the final MLP.

For 20 slides of the validation dataset (10 classified with highest RlapsRisk BC scores, 10 classified with lowest RlapsRisk BC scores by our model), we computed the Shapley values of each tile and extracted those with the 10 highest computed contributions (in the case of the highest RlapsRisk BC scores, those tiles constituting the high risk contribution group) and 10 lowest ones (in the case of the lowest RlapsRisk BC scores, constituting the low risk contribution group) for each slide. Those tiles were further reviewed and annotated by two expert pathologists blinded to the predicted outcome.

Fifty-six histological features were recorded, encompassing tumor architecture patterns, stroma features, presence of different cell types and tiles' localization. Only twenty-two histological features were kept in the analysis after excluding those with low agreement between the two experts (Cohen's kappa < 0.21, (48)). The proportions

of appearance of each feature in the highest and lowest contribution groups were compared with a two proportion Z-test, statistical significance was calculated using the Bonferroni adjustment, resulting in a corrected alpha value of 1E-3.

Statistical Analysis

Performance assessment though Harrell's c-index and Kaplan-Meier analyses were performed with uni- and multivariable Cox proportional hazards models implemented in the lifelines (0.27.4) package of Python, cumulative sensitivities and dynamic specificities were computed using the scikit-survival package (0.19.0). We performed Kaplan-Meier analyses to assess the association of each classifier with MFS (the presented HR and their 95%CI corresponding to the related univariable Cox models), and used Log-rank tests to compare survival distributions between stratification subgroups. In order to evaluate the prediction performance of these classifiers in terms of discrimination of the risks of metastatic events before five years, we computed the cumulative sensitivity as well as the dynamic specificity associated with each model. These are natural extensions of the so-called sensitivity/specificity to the particular setting of time-to-event outcomes that may be censored (24). P-values to compare the performance in c-index of the different models were based on permutation tests. Confidence intervals were computed using bootstrapping with nonparametric, unstratified resampling. All tests were two-tailed, and P values < 0.05 were considered statistically significant. We followed the recent MI-Claim checklist to improve the reporting of our ML methods. The checklist is available in the Supplementary Materials.

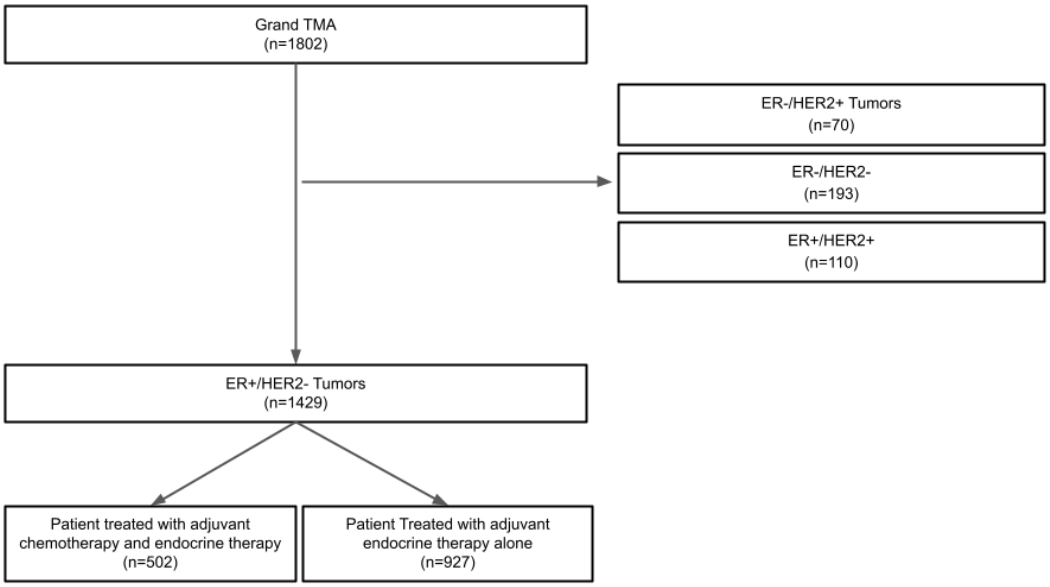
Methods References

37. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update | Journal of Clinical Oncology [Internet]. [cité 17 nov 2022]. Disponible sur: https://ascopubs.org/doi/10.1200/JCO.2018.77.8738
38. Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, et al. Assessment of Ki67 in Breast Cancer: Recommendations from the International Ki67 in Breast Cancer Working Group. JNCI J Natl Cancer Inst. 16 nov 2011;103(22):1656-64.
39. Franchet C, Djerroudi L, Maran-Gonzalez A, Abramovici O, Antoine M, Becette V, et al. Mise à jour 2021 des recommandations du GEFPICS pour l'évaluation du statut HER2 dans les cancers infiltrants du sein en France. Ann Pathol. 1 nov 2021;41(6):507-20.

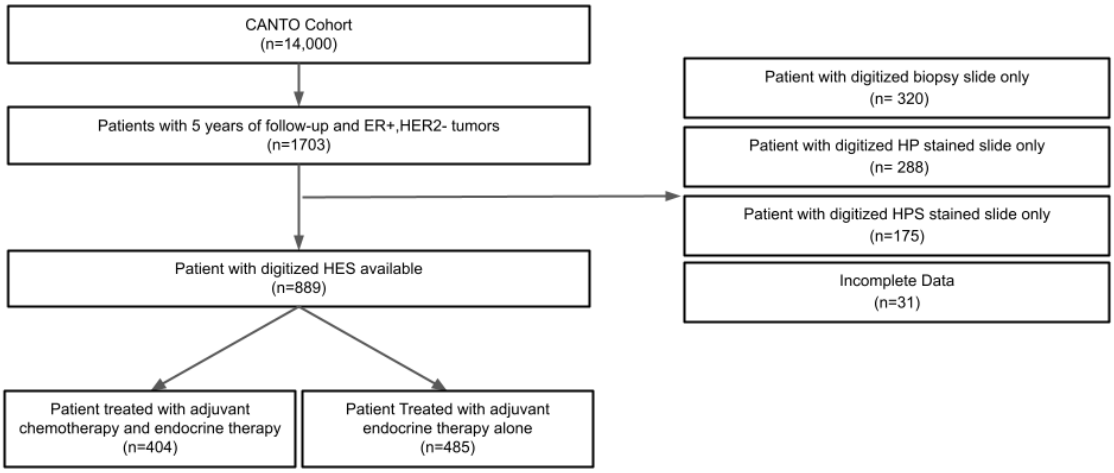
40. Tashima R, Nishimura R, Osako T, Nishiyama Y, Okumura Y, Nakano M, et al. Evaluation of an Optimal Cut-Off Point for the Ki-67 Index as a Prognostic Factor in Primary Breast Cancer: A Retrospective Study. PLOS ONE. 15 juill 2015;10(7):e0119565.
41. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [Internet]. arXiv; 2015 [cité 24 oct 2022]. Disponible sur: http://arxiv.org/abs/1505.04597
42. Chen X, Fan H, Girshick R, He K. Improved Baselines with Momentum Contrastive Learning [Internet]. arXiv; 2020 [cité 24 oct 2022]. Disponible sur: http://arxiv.org/abs/2003.04297
43. Saillard C, Dehaene O, Marchand T, Moindrot O, Kamoun A, Schmauch B, et al. Self-supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. In: Proceedings of the MICCAI Workshop on Computational Pathology [Internet]. PMLR; 2021 [cité 24 oct 2022]. p. 191-205. Disponible sur: https://proceedings.mlr.press/v156/saillard21a.html
44. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015;19(1A):A68-77.
45. Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning [Internet]. arXiv; 2018 [cité 24 oct 2022]. Disponible sur: http://arxiv.org/abs/1802.04712
46. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLOS ONE. 17 juin 2020;15(6):e0233678.
47. Shapley LS. 17. A Value for n-Person Games. In: 17 A Value for n-Person Games [Internet]. Princeton University Press; 2016 [cité 24 oct 2022]. p. 307-18. Disponible sur: https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html
48. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 15 oct 2012;22(3):276-82.

SUPPLEMENTARY MATERIAL

Flow Charts



Supplementary Figure 1: Flow Diagram GrandTMA Cohort



Supplementary Figure 2: Flow Diagram CANTO Cohort

Clinical variables

		TMA		CANTO	
Clinical Variable	Group	*N*	%	*N*	%
MFS events	None	**1379**	**92,61%**	**843**	**94,83%**
	before 5 years	**57**	**3,83%**	**34**	**3,82%**
	after 5 years	**53**	**3,56%**	**12**	**1,35%**
Follow-up	Mean (sd)	**77 months(41)**	**NA**	**69 months(24)**	**NA**
Age	Mean (sd)	**61.2 (12.4)**	**NA**	**57.7 (12.2)**	**NA**
Menopausal Status	Post-menopausal	**1070**	**74.88%**	**667**	**74.86%**
	Pre-menopausal	**359**	**25.12%**	**224**	**25.14%**
pT	1	**979**	**68.51%**	**592**	**66,44%**
	2	**398**	**27.85%**	**265**	**29,74%**
	3	**50**	**3.40%**	**33**	**3,70%**
	4	**3**	**0.21%**	**0**	**0,00%**
	NA	**2**	**0.14%**	**1**	**0,11%**
pN	0	**950**	**66.48%**	**593**	**66,55%**
	1	**359**	**25.12%**	**227**	**25,48%**
	2	**72**	**5.04%**	**46**	**5,16%**
	3	**48**	**3.36%**	**24**	**2,69%**
Lymph node status	N0	**950**	**66.48%**	**753**	**84,51%**
	N+	**479**	**33.52%**	**138**	**15,49%**
Tumor histological grade	G1	**428**	**29.95%**	**185**	**20,76%**
	G2	**736**	**51.50%**	**537**	**60,27%**
	G3	**263**	**18.40%**	**169**	**18,97%**
	NA	**2**	**0.14%**	**0**	**0,00%**
Ki67	<20%	**1150**	**80,48%**	**756**	**84,85%**
	>20%	**279**	**19,52%**	**135**	**15,15%**
Radiation therapy	YES	**1217**	**85.16%**	**475**	**53,31%**

	NO	**212**	**14.84%**	**516**	**57,91%**
Endocrine therapy	YES	**1387**	**97.06%**	**856**	**96,07%**
	NO	**42**	**2.94%**	**35**	**3,93%**
Chemotherapy	YES	**505**	**35.34%**	**485**	**54,43%**
	NO	**924**	**64.66%**	**406**	**45,57%**

Supplementary Table 1 : Patients characteristics in the discovery and validation datasets

Clinical risk groups

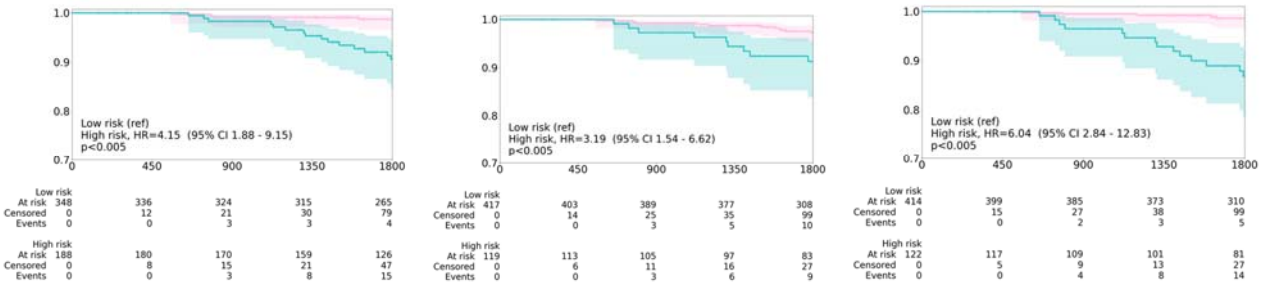
High Risk (At least one criterium)	Intermediate	Low Risk (All Criteria)
• pT3 • pN1 & pre-menopause • pN2 & post-menopause • Grade 3	Other Situation	• pT1 • pN0/i+/mi • Grade 1 • KI67<15-20% • No lymphovascular emboli

Supplementary Table 2 : Clinical risk groups definition.

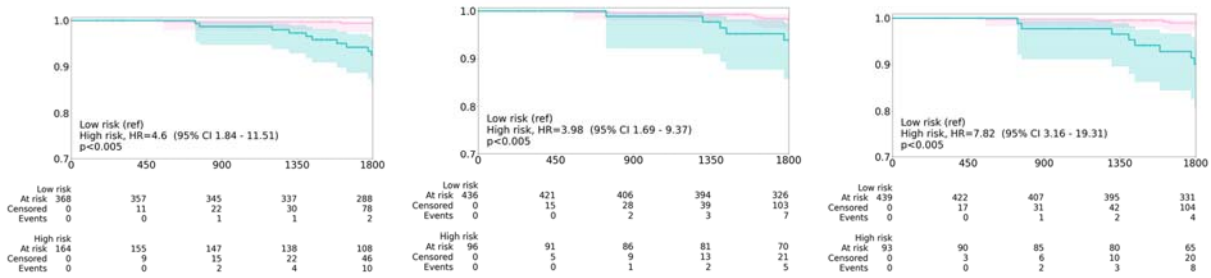
Integrating a tumor-related feature in the algorithm

To train the tumor classification algorithm, we had at our disposal the 1759 slides of the GrandTMA database whose tumor regions had been contoured by two expert pathologists. The feature extraction pipeline presented in the Model Description was re-used and the multiple instance learning model was replaced by a binary classifier composed of a linear layer with 2048 neurons that classified each tile as tumor vs. non tumor.

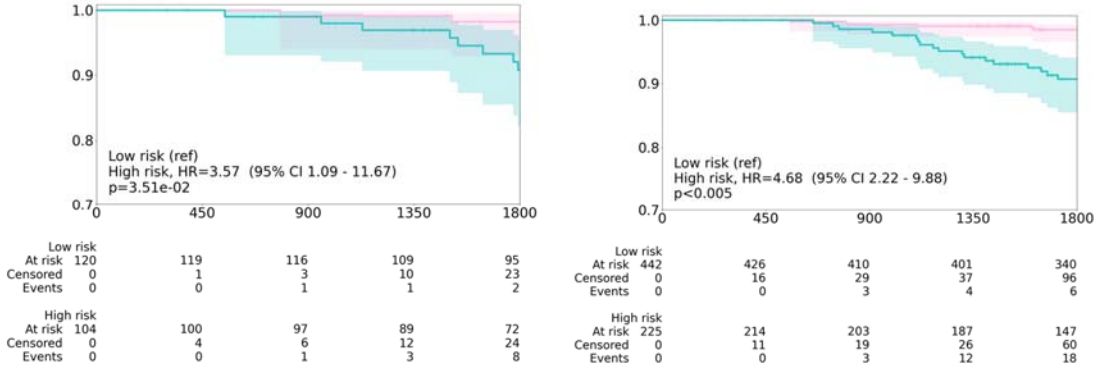
KM Analyses



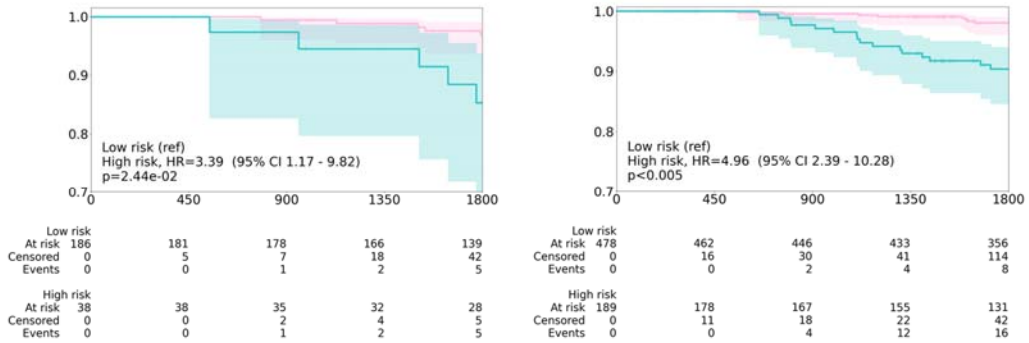
Supplementary Figure 3: Stratification performed by RlapsRisk BC Classifier (left), Clinical Score Classifier (middle), Combined Model Classifier (right) on Grade 2 patients from the CANTO cohort.



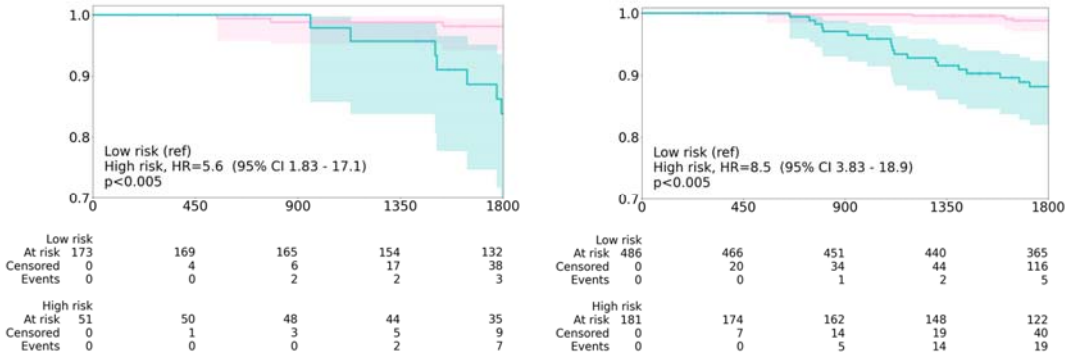
Supplementary Figure 4: Stratification performed by RlapsRisk BC Classifier (left), Clinical Score Classifier (middle), Combined Model Classifier (right) on "intermediate risk" patients (defined in Table 1) from the CANTO cohort.



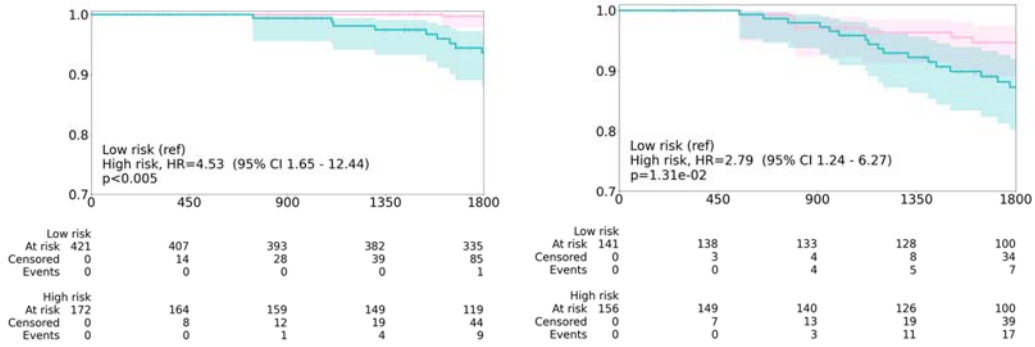
Supplementary Figure 5A: Stratification performed by RlapsRisk BC Classifier on pre-menopausal patients of CANTO (left), post-menopausal patients (right)



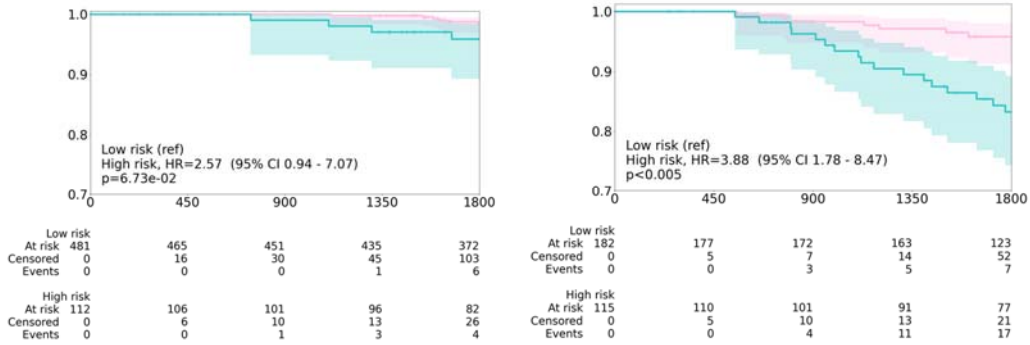
828
829 *Supplementary Figure 5B: Stratification performed by the Clinical Score Classifier on pre-menopausal*
830 *patients of CANTO (left), post-menopausal patients (right)*
831
832



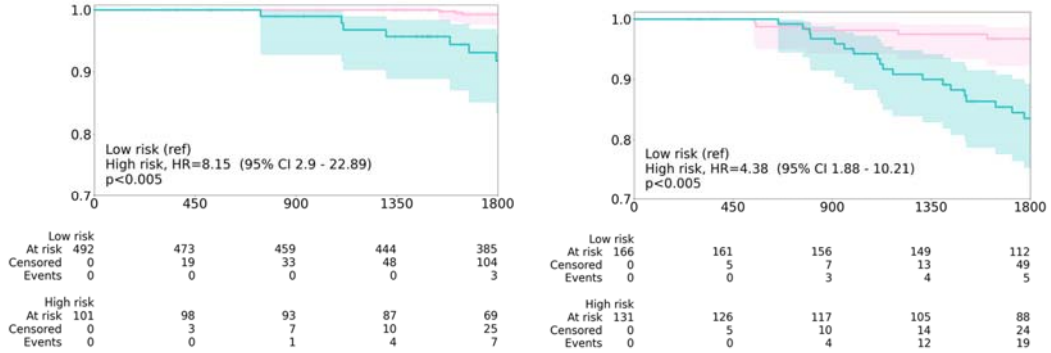
833
834 *Supplementary Figure 5C: Stratification performed by the Combined Model Classifier on pre-*
835 *menopausal patients of CANTO (A), post-menopausal patients (B)*
836
837
838
839
840
841
842
843



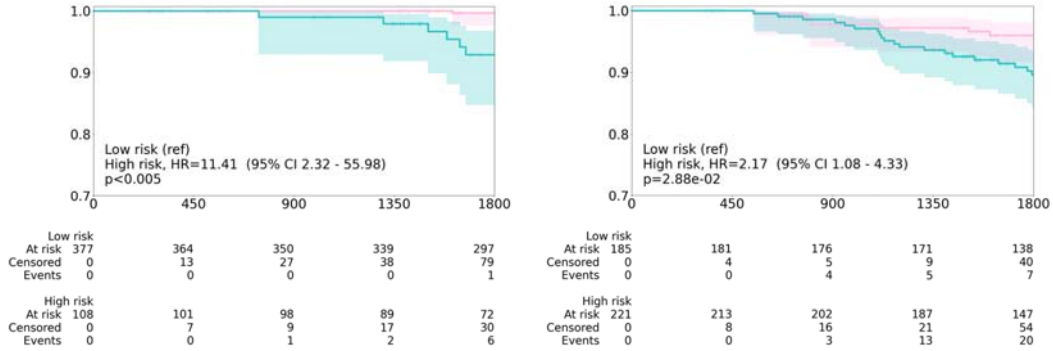
844
845 *Supplementary Figure 6A: Stratification performed by RIapsRisk BC Classifier on patients without lymph*
846 *node invasion of CANTO (left), patients with lymph-node invasion (right).*
847



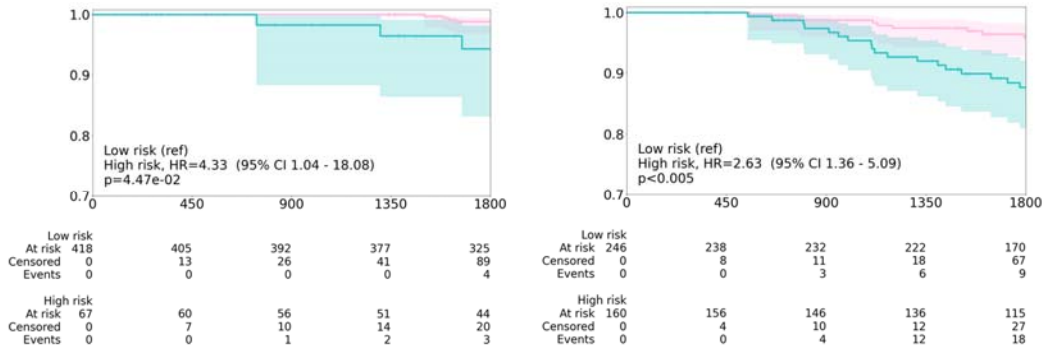
Supplementary Figure 6B: Stratification performed by the Clinical Score Classifier on patients without lymph node invasion of CANTO (left), patients with lymph-node invasion (right).



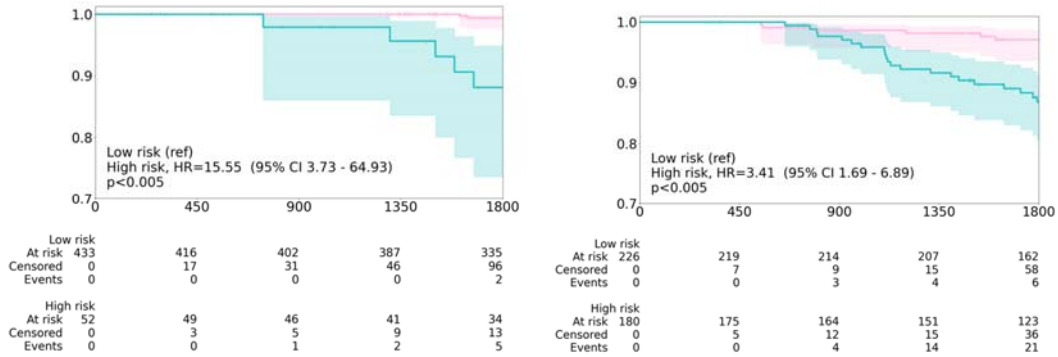
Supplementary Figure 6C: Stratification performed by the Combined Model Classifier on patients without lymph node invasion of Canto (left), patients with lymph-node invasion (right).



Supplementary Figure 7A: Stratification performed by RlapsRisk BC Classifier on patients treated with endocrine therapy alone in CANTO (left), patients treated with chemo-endocrine therapy (right)



Supplementary Figure 7B: Stratification performed by the Clinical Score Classifier on patients treated with endocrine therapy alone in CANTO (left), patients treated with chemo-endocrine therapy (right)



Supplementary Figure 7C: Stratification performed by the Combined Model Classifier on patients treated with endocrine therapy alone in CANTO (left), patients treated with chemo-endocrine therapy (right)

Discussion and Conclusions

Discussion and Conclusions

A. Discussion

1. General considerations

Breast cancer is naturally positioned as a discipline of choice for the application of AI due to its high incidence and the major oncotheranostic issues that it carries.

This is evidenced by the great interest that the association of AI and medicine is arousing in oncology community, with the increase of presented abstracts at high-level congresses and symposiums; particularly, our work was selected for proffered paper presentation at the ESMO Congress 2021 and then the validation phase was designated for poster presentation with an award at the last ESMO Congress 2022 (see Annexes section).

The new requirements of patients' stratification can be met by a combination of computational pathology and artificial intelligence. Mathematical models seem to be very attractive tools with which to elucidate diverse scientific questions such as the response to a certain treatment or the prediction of risk of relapse (Pagès et al., 2018). The benefit of chemotherapy reducing the risk of recurrence in high-risk tumors is well documented, but in low-risk tumors, the side effects of such treatment would most likely outweigh the profits (Penault-Llorca, 2021). In the intermediate group, and essentially in the early stage diseases, the risk stratification is difficult and additional tools are necessary to aid in the decision about whether chemotherapy would add extra benefit compared to endocrine therapy alone. This is even more important because, due to the increase in the proportion of breast cancers with a good prognosis, as a result of screening, early diagnosis and faster implementation of treatments, novel strategies are required targeting this particular population. The current objective is therefore no longer just to prevent recurrence, but also to improve the quality of life of patients, which means to reduce toxicity as much as possible.

Thus, the chosen strategy for our work was to develop a DL model and to combine it with histopathological features present on WSI and clinico-pathological data to assess the risk of relapse in early breast cancer. We tested this method on an external cohort and we showed that it produced reliable data that validated its generalization. We also tried to identify which morphological characteristics, along with their location on the tissue, were most important for

the algorithm and best predictors of patients' outcome.

Supplementary corroboration of our method come to the fact that results are consistent with other approaches already validated for clinical practice. In the same line of molecular signatures and other prediction tools such as Predict Breast and CTS0, the prognostic performance was slightly weaker in pre-menopausal patients than in post-menopausal patients. Regarding lymph node status, our method shows an advantage over the other predictors with an equivalent c-index for prognosis in patients with positive or negative lymph nodes, contrary to the other cited methods that, although still significant, are a little less performant in the node negative setting (Sestak et al., 2018). It should be contemplated that besides all the performance analyses, molecular signatures are still expensive and not available in less developed regions in the world, while our method requires a single routine digitized slide only, already available for diagnosis.

Distinct characteristics of the considered disease have an influence in the study design. Since HR+ breast cancers can show a long period before the recurrence appears, two considerations must be mentioned. First, the careful endpoint selection that must be associated with a specific period for event prediction. This endpoint should be extensive enough to cover a sufficient quantity of events, taking into account other points such as, in our case, the transfer of patients' controls after five years that can result in a loss of follow-up. Second, and as exposed in the review by Abreu et al., in this particular population there exist the imbalanced class distribution, that is, the "recurrence" class is represented by a shorter number of examples than the "no-recurrence" class, which is inherent in survival analyses, however (Henriques Abreu et al., 2016). The class imbalance may deteriorate the performance of the model if this is not considered and corrected, since one of the classes will not have enough elements for the classifier to learn about it. In HR+ breast cancer recurrences, where the events tend to be scarce in the short term, this is particularly dangerous, since a false negative would be traduced in no treatment for a recurrent cancer because the disease is classified as in remission. Nevertheless, we did not employ any particular strategy to manage this problem in our final model, since we had tested some traditional methods of oversampling patients with recorded relapses, but without obtaining significant improvements in the results.

Interpretability is one of the big issues in the field of AI applied to medicine. However, the recognition of "hidden" features by deep learning models, as we found in our study, can

provide novel information regarding prognosis. Once these features identified, they can be deciphered and comprehended for ulterior integration in histopathological assessment, which may be more easily accepted by the medical community than a "closed" method.

2. **Limitations**

Firstly, pre-analytical factors such as fixation procedures, slide preparation (cutting, staining quality) and storage in every laboratory can affect the results of the evaluation and should be adjusted to the recommendations of use of our AI-based tool. Scanning problems involving unfocused images or bad quality WSIs should be also taken into consideration. A pathology quality assessment is imperative as an initial step to minimize the effect of the pre-analytical variables.

Another major concern is the availability of big, clean and updated cohorts in the medical domain, with complete clinical and histopathological data, and associated good quality WSI suitable for digital image analysis. AI-based approaches are dependent on both the quality and the quantity of input data. The absence of any of these conditions translates in insufficient amount of training and/or test data, which, in turn, difficults the validation of experimental AI models. Unfortunately, this limitation is quite frequent, as seen in the interesting work carried out by Klimov et al. to predict the risk of relapse in ductal carcinoma in situ (Klimov et al., 2019). Contrary to the aforementioned work, even if our sample size was limited when compared with other studies of the genre, a blind validation of our model was performed on an independent cohort, with promising results, and efforts are currently ongoing to enlarge the test cohorts in order to increase the robustness of our algorithm.

As highlighted by (Mobadersany et al., 2018), DL models applied to histology raise a challenge: the features in WSI are, in the end, pixels with a signification that depend entirely on context and, finally, in the interpretation, without an intrinsic meaning. The result is commonly a complex system, laborious to grasp and thus more prone to be rejected. In any case, and regardless of the sophisticated tools that can be deployed, the role of the pathologists is essential as a guarantor of the process and only their expertise can provide sense to the results and verify the analyses carried out on the tissue sections.

In a wider angle, the application of medical decision support systems to the field of medicine require the overcoming of some major obstacles. First of all, databases that are rapidly expanding must be held to appropriate standards for the ulterior use in a workflow that

includes digital pathology and/or AI. Second, there must be a change in the degree of acceptance to the new technological solutions, to make easier their implementation after apprehending their functioning and the benefits that they have to offer for patients and for the medical workflow in general. Third, the establishment of multidisciplinary teams must be encouraged and facilitated, including health care delivery personal, database, statistics and computational experts, in order to create a gradual transition to the health care of the future.

3. **Perspectives**

Additional applications of our work include the use of our DL-based approach to predict biomarkers status from HES WSI in an accurate and objective manner (see in *Annexes* section) (poster USCAP). We found that biomarker status could be predicted with 91%, 76%, 94% and 85% accuracy for ER, PR, HER2 and Ki67 respectively, suggesting that it could be a viable option for patients' stratification and as a support in therapeutic decisions in the low-resource setting, where biomarker testing by IHC is not routinely performed. Future work should extend this approach through the validation on larger cohorts.

We are currently focused on the prospective phase of our study, which includes the deployment of RlapsRisk, our AI-based prediction tool, in the routine workflow, and aims at comparing RlapsRisk score to the current routine molecular signatures scores (Oncotype DX®). This trial period consists of scanning new tumor slides and quality check performed by pathologists to feed the algorithm, which will produce a .pdf document with the result of prediction (**Figure 18**), as the model displayed on the Annexes section. This report will then integrate the medical report of the patient, as is presently the case with the pathological report, the radiological studies or the results of molecular tests. We believe that external validation as performed in our work is a great advantage over other AI-based studies, as pointed by (Kim et al., 2019), and that this fundamental prospective phase will contribute with the design features that are recommended for robust validation of the real-world clinical performance of AI algorithms. This study will allow us to directly compare our AI-based RlapsRisk to the current molecular signatures (Oncotype DX®).

We also plan to increase the retrospective validation of the RlapsRisk tool in a larger clinical trials that were used to validate molecular signatures, such as the TransATAC study. Moreover, it would be very important to explore whether the RlapsRisk device is able to predict late recurrence between five and ten years (an aspect that we could not explore in our

GrandTMA cohort due to the decentralized follow-up of the patients after 5 years). Ongoing studies are in process on national and international plan to: 1) retrospectively validate our results in other centers (France, Italy, Netherlands), 2) extend validation to HE-stained slides and 3) generalize the validation of RlapsRisk to other scanner devices.

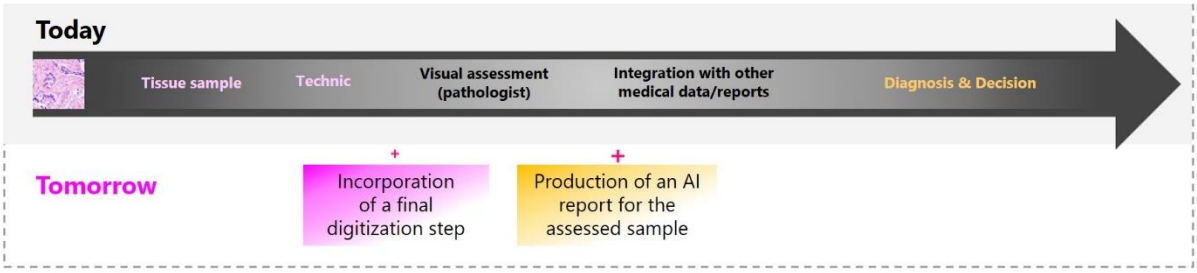


Figure 18. Routine workflow at the Pathology laboratory with the implementation of RlapsRisk.

B. Conclusions

In conclusion, we validated an AI-based digital pathology medical device for use on diagnostic tissue slides. The technique enabled us to provide prognostic information from HES-stained tumor WSI in the final form of a report containing a risk of relapse score and detailed areas of the slides that contributed to the prediction, in order to support interpretability.

Finally, we expect that our efforts pave the way for future biomarker research, contributing to plant the foundations for the development of new personalized treatments, taking advantage of the capacity of deep learning to discover previously unrecognized morphological characteristics from tissue sections that could have clinical relevance. This convergence of digital pathology, automation, and powerful analytics like DL and AI in healthcare, along with the strengths of human perception and judgment, are bringing together the tools needed for scientists and clinicians to unlock medical breakthroughs at a pace like never before.

References

References

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., Beck, A. H., & Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance : A white paper from the Digital Pathology Association. *The Journal of Pathology*, *249*(3), 286-294.

Aeffner, F., Adissu, H. A., Boyle, M. C., Cardiff, R. D., Hagendorn, E., Hoenerhoff, M. J., Klopfleisch, R., Newbigging, S., Schaudien, D., Turner, O., & Wilson, K. (2018). Digital Microscopy, Image Analysis, and Virtual Slide Repository. *ILAR Journal*, *59*(1), 66-79.

Aeffner, F., Wilson, K., Martin, N. T., Black, J. C., Hendriks, C. L. L., Bolon, B., Rudmann, D. G., Gianani, R., Koegler, S. R., Krueger, J., & Young, G. D. (2017). The Gold Standard Paradox in Digital Image Analysis : Manual Versus Automated Scoring as Ground Truth. *Archives of Pathology & Laboratory Medicine*, *141*(9), 1267-1275.

Aeffner, F., Zarella, M. D., Buchbinder, N., Bui, M. M., Goodman, M. R., Hartman, D. J., Lujan, G. M., Molani, M. A., Parwani, A. V., Lillard, K., Turner, O. C., Vemuri, V. N. P., Yuil-Valdes, A. G., & Bowman, D. (2019). Introduction to digital image analysis in whole-slide imaging : A white paper from the digital pathology association. *Journal of Pathology Informatics*, *10*(1), 9.

Al-Janabi, S., Huisman, A., & Van Diest, P. J. (2012). Digital pathology : Current status and future perspectives. *Histopathology*, *61*(1), 1-9.

Al-Kofahi, Y., Lassoued, W., Lee, W., & Roysam, B. (2010). Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images. *IEEE Transactions on Biomedical Engineering*, *57*(4), 841-852.

Allemand, H., Seradour, B., Weill, A., & Ricordeau, P. (2008). [Decline in breast cancer incidence in 2005 and 2006 in France : A paradoxical trend]. *Bulletin Du Cancer*, *95*(1), 11-15.

Ameisen, D., Deroulers, C., Perrier, V., Bouhidel, F., Battistella, M., Legrès, L., Janin, A., Bertheau, P., & Yunès, J.-B. (2014). Towards better digital pathology workflows : Programming libraries for high-speed sharpness assessment of Whole Slide Images. *Diagnostic Pathology*, *9*(Suppl 1), S3.

Ameisen, D., Deroulers, C., Perrier, V., Yunès, J.-B., Bouhidel, F., Battistella, M., Legrès, L., Janin, A., & Bertheau, P. (2013). Stack or trash? Quality assessment of virtual slides. *Diagnostic Pathology*, *8*(Suppl 1), S23.

Ameisen, D., Naour, G. L., & Daniel, C. (2012). Technologie des lames virtuelles—De la numérisation à la mise en ligne. *médecine/sciences*, *28*(11), 977-982.

Andre, F., Berrada, N., & Desmedt, C. (2010). Implication of tumor microenvironment in the resistance to chemotherapy in breast cancer patients. *Current Opinion in Oncology*, *22*(6), 547-551.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878.

Aubreville, M., Stathonikos, N., Bertram, C. A., Klopleisch, R., ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T. A., Maier, A., Breen, J., Ravikumar, N., Chung, Y., Park, J., Nateghi, R., Pourakpour, F., Fick, R. H. J., Ben Hadj, S., Jahanifar, M., … Breininger, K. (2022). Mitosis domain generalization in histopathology images—The MIDOG challenge. In *ArXiv e-prints*. https://ui.adsabs.harvard.edu/abs/2022arXiv220403742A

Autier, P., Boniol, M., LaVecchia, C., Vatten, L., Gavin, A., Héry, C., & Heanue, M. (2010). Disparities in breast cancer mortality trends between 30 European countries : Retrospective trend analysis of WHO mortality database. *BMJ*, *341*.

Balkwill, F., & Mantovani, A. (2001). Inflammation and cancer : Back to Virchow? *The Lancet*, *357*(9255), 539-545.

Bándi, P., Geessink, O., Manson, Q., Dijk, M. V., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F. G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., … Litjens, G. (2019). From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level : The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging*, *38*(2), 550-560.

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M., & Hamilton, P. W. (2017). QuPath : Open source software for digital pathology image analysis. *Scientific Reports*, *7*(1), 16878.

Basavanhally, A. N., Ganesan, S., Agner, S., Monaco, J. P., Feldman, M. D., Tomaszewski, J. E., Bhanot, G., & Madabhushi, A. (2010). Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology. *IEEE Transactions on Biomedical Engineering*, *57*(3), 642-653.

Bataillon, G., Vincent-Salomon, A., Jouvin, N., & Walter, T. (2019, mars 1). La pathologie à l'heure de l'intelligence artificielle : Exemple... *Correspondances en Onco-Théranostic*, 52.

Bayramoglu, N., & Heikkilä, J. (2016). Transfer Learning for Cell Nuclei Classification in Histopathology Images. In G. Hua & H. Jégou (Éds.), *Computer Vision – ECCV 2016 Workshops* (p. 532-539). Springer International Publishing.

Becich, M. J. (2000). The role of the pathologist as tissue refiner and data miner : The impact of functional genomics on the modern pathology laboratory and the critical roles of pathology informatics and bioinformatics. *Molecular Diagnosis: A Journal Devoted to the Understanding of Human Disease Through the Clinical Application of Molecular Biology*, *5*(4), 287-299.

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., Vijver, M. J. van de, West, R. B., Rijn, M. van de, & Koller, D. (2011). Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Science Translational Medicine*, *3*(108), 108ra113-108ra113.

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms : An online database. *Npj Digital Medicine*, *3*(1), 1-8.

Bertheau, P., Chabouis, A., Fabiani, B., Poullier, É., Daniel, C., Cucherousset, J., Bosq, J., Hénin, D., Capron, F., & Guettier, C. (2012). Télépathologie par lames virtuelles ou le diagnostic anatomo-pathologique en réseau numérique. *médecine/sciences*, *28*(11), 983-985.

Bini, S. A. (2018). Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing : What Do These Terms Mean and How Will They Impact Health Care? *The Journal of Arthroplasty*, *33*(8), 2358-2361.

Brawanski, A. (2012). On the myth of the Edwin Smith papyrus : Is it magic or science? *Acta Neurochirurgica*, *154*(12), 2285-2291.

Broeders, M., Moss, S., Nyström, L., Njor, S., Jonsson, H., Paap, E., Massat, N., Duffy, S., Lynge, E., Paci, E., & EUROSCREEN Working Group. (2012). The impact of mammographic screening on breast cancer mortality in Europe : A review of observational studies. *Journal of Medical Screening*, *19 Suppl 1*, 14-25.

Buus, R., Sestak, I., Kronenwett, R., Denkert, C., Dubsky, P., Krappmann, K., Scheer, M., Petry, C., Cuzick, J., & Dowsett, M. (2016). Comparison of EndoPredict and EPclin With Oncotype DX Recurrence Score for Prediction of Risk of Distant Recurrence After Endocrine Therapy. *Journal of the National Cancer Institute*, *108*(11), djw149.

Buyse, M., Consortium, O. behalf of the T., Loi, S., Consortium, O. behalf of the T., van't Veer, L., Consortium, O. behalf of the T., Viale, G., Consortium, O. behalf of the T., Delorenzi, M., Consortium, O. behalf of the T., Glas, A. M., Consortium, O. behalf of the T., Saghatchian d'Assignies, M., Consortium, O. behalf of the T., Bergh, J., Consortium, O. behalf of the T., Lidereau, R., Consortium, O. behalf of the T., Ellis, P., … Consortium, O. behalf of the T. (2006). Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. *JNCI: Journal of the National Cancer Institute*, *98*(17), 1183-1192.

Camp, G. (1996). Problem-Based Learning : A Paradigm Shift or a Passing Fad? *Medical Education Online*, *1*(1), 4282.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, *25*(8), 1301-1309.

Cardoso, F., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rubio, I. T., Zackrisson, S., & Senkus, E. (2019). Early breast cancer : ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Annals of Oncology*, *30*(8), 1194-1220.

Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.-Y., Brain, E., Causeret, S., DeLorenzi, M., Glas, A. M., Golfinopoulos, V., Goulioti, T., Knox, S., Matos, E., Meulemans, B., Neijenhuis, P. A., Nitz, U., Passalacqua, R., … Piccart, M. (2016). 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine*, *375*(8), 717-729.

Chacón, R. D., & Costanzo, M. V. (2010). Triple-negative breast cancer. *Breast Cancer Research*, *12*(2), S3.

Chan, J. Y., & Salto-Tellez, M. (2012). Opinion : Molecular gestalt and modern pathology. *Advances in Anatomic Pathology*, *19*(6), 425-426.

Chang, H. Y., Jung, C. K., Woo, J. I., Lee, S., Cho, J., Kim, S. W., & Kwak, T.-Y. (2019). Artificial Intelligence in Pathology. *Journal of Pathology and Translational Medicine*, *53*(1), 1-12.

Chen, J., & Srinivas, C. (2016). *Automatic Lymphocyte Detection in H&E Images with Deep Neural Networks*.

Chen, J.-M., Qu, A.-P., Wang, L.-W., Yuan, J.-P., Yang, F., Xiang, Q.-M., Maskey, N., Yang, G.-F., Liu, J., & Li, Y. (2015). New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images. *Scientific Reports*, *5*(1), 10690.

Chen, X., & Fan, H. (2020). Improved Baselines with Momentum Contrastive Learning. *ArXiv: Computer Vision and Pattern Recognition*. https://typeset.io/papers/improved-baselines-with-momentum-contrastive-learning-wgoqbbbh9y

Cheng, W.-C., Lam, S., Gong, Q., Lemaillet, P., Food, U., Administration, D., & States, U. (2020). Evaluating whole-slide imaging viewers used in digital pathology. *Image Quality and System Performance*, 6.

Chong, Y., Kim, D. C., Jung, C. K., Kim, D., Song, S. Y., Joo, H. J., & Yi, S.-Y. (2020). Recommendations for pathologic practice using digital pathology : Consensus report of the Korean Society of Pathologists. *Journal of Pathology and Translational Medicine*, *54*(6), 437-452.

Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, *16*(Pt 2), 411-418.

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *The New England Journal of Medicine*, *372*(9), 793-795.

Conde-Sousa, E., Vale, J., Feng, M., Xu, K., Wang, Y., Della Mea, V., La Barbera, D., Montahaei, E., Baghshah, M. S., Turzynski, A., Gildenblat, J., Klaiman, E., Hong, Y., Aresta, G., Araújo, T., Aguiar, P., Eloy, C., & Polónia, A. (2021). HEROHE Challenge : Assessing HER2 status in breast cancer without immunohistochemistry or in situ hybridization. *arXiv:2111.04738 [cs, eess, q-bio]*. http://arxiv.org/abs/2111.04738

Cornish, T. C. (2020). Clinical Application of Image Analysis in Pathology. *Advances in Anatomic Pathology*, *27*(4), 227–235.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, *24*(10), 1559-1567.

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., Girard, N., Elemento, O., Nicholson, A. G., Blay, J.-Y., Galateau-Sallé, F., Wainrib, G., & Clozel, T. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, *25*(10), 1519-1525.

Coussens, L. M., & Werb, Z. (2002). Inflammation and cancer. *Nature*, *420*(6917), 860-867.

Couture, H. D., Williams, L. A., Geradts, J., Nyante, S. J., Butler, E. N., Marron, J. S., Perou, C. M., Troester, M. A., & Niethammer, M. (2018). Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *Npj Breast Cancer*, *4*(1), 1-8.

Cserni, G. (2020). Histological type and typing of breast carcinomas and the WHO classification changes over time. *Pathologica - Journal of the Italian Society of Anatomic Pathology and Diagnostic Cytopathology*, *112*(1), 25-41.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., … Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346-352.

Cuzick, J., Dowsett, M., Pineda, S., Wale, C., Salter, J., Quinn, E., Zabaglo, L., Mallon, E., Green, A. R., Ellis, I. O., Howell, A., Buzdar, A. U., & Forbes, J. F. (2011). Prognostic Value of a Combined Estrogen Receptor, Progesterone Receptor, Ki-67, and Human Epidermal Growth Factor Receptor 2 Immunohistochemical Score and Comparison With the Genomic Health Recurrence Score in Early Breast Cancer. *Journal of Clinical Oncology*, *29*(32), 4273-4278.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*, *5*(10), 2929-2943.

Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks. *Cancer*, *91*(S8), 1615-1635.

Death rates in top four cancer killers fall by a third over 20 years. (2014, août 17). Cancer Research UK - Cancer News. https://news.cancerresearchuk.org/2014/08/18/death-rates-in-top-four-cancer-killers-fall-by-a-third-over-20-years/

Delaune, A., Valmary-Degano, S., Loménie, N., Zryouil, K., Benyahia, N., Trassard, O., Eraville, V., Bergeron, C., Devouassoux-Shisheboran, M., Glaser, C., Bataillon, G., Bacry, E., Combes, S.,

Prevot, S., & Bertheau, P. (2022). [The first data challenge of the french society of pathology : An international competition in 2020, a research tool in A.I. for the future?]. *Annales De Pathologie*, *42*(2), 119-128.

Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P., & Narod, S. A. (2007). Triple-Negative Breast Cancer : Clinical Features and Patterns of Recurrence. *Clinical Cancer Research*, *13*(15), 4429-4434.

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., & Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *14*(16), 5158-5165.

Dietz, R. L., Hartman, D. J., & Pantanowitz, L. (2020). Systematic Review of the Use of Telepathology During Intraoperative Consultation. *American Journal of Clinical Pathology*, *153*(2), 198-209.

Dodington, D. W., Lagree, A., Tabbarah, S., Mohebpour, M., Sadeghi-Naini, A., Tran, W. T., & Lu, F.-I. (2021). Analysis of tumor nuclear features using artificial intelligence to predict response to neoadjuvant chemotherapy in high-risk breast cancer patients. *Breast Cancer Research and Treatment*, *186*(2), 379-389.

Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (s. d.). *CryptoNets : Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*. 10.

Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A. K., Cowens, J. W., Ferree, S., Storhoff, J., Schaper, C., & Cuzick, J. (2013). Comparison of PAM50 Risk of Recurrence Score With Onco *type* DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy. *Journal of Clinical Oncology*, *31*(22), 2783-2790.

Dubsky, P. C. (2020). The EndoPredict score predicts response to neoadjuvant chemotherapy and neoendocrine therapy in hormone receptor-positive, human epidermal growth factor receptor 2-negative breast cancer patients from the ABCSG-34 trial. *European Journal of Cancer*, 8.

Echle, A., Rindtorff, N. T., Brinker, T. J., Luedde, T., Pearson, A. T., & Kather, J. N. (2021). Deep learning in cancer pathology : A new generation of clinical biomarkers. *British Journal of Cancer*, *124*(4), 686-696.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., … Venâncio, R. (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, *318*(22), 2199-2210.

Ektefaie, Y., Yuan, W., Dillon, D. A., Lin, N. U., Golden, J. A., Kohane, I. S., & Yu, K.-H. (2021). Integrative multiomics-histopathology analysis for breast cancer classification. *NPJ Breast Cancer*, *7*(1), 147.

Ellis, I. O., Galea, M., Broughton, N., Locker, A., Blamey, R. W., & Elston, C. W. (1992). Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. *Histopathology*, *20*(6), 479-489.

Elston, C. W., & Ellis, I. O. (1991). pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer : Experience from a large study with long-term follow-up. *Histopathology*, *19*(5), 403-410.

Ertosun, M. G., & Rubin, D. L. (2015). Automated Grading of Gliomas using Deep Learning in Digital Pathology Images : A modular approach with ensemble of convolutional neural networks. *AMIA Annual Symposium Proceedings*, *2015*, 1899-1908.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24-29.

Evans, A. J., Bauer, T. W., Bui, M. M., Cornish, T. C., Duncan, H., Glassy, E. F., Hipp, J., McGee, R. S., Murphy, D., Myers, C., O'Neill, D. G., Parwani, A. V., Rampy, B. A., Salama, M. E., & Pantanowitz, L. (2018). US Food and Drug Administration Approval of Whole Slide Imaging for Primary Diagnosis : A Key Milestone Is Reached and New Questions Are Raised. *Archives of Pathology & Laboratory Medicine*, *142*(11), 1383-1387.

Evans, A. J., Chetty, R., Clarke, B. A., Croul, S., Ghazarian, D. M., Kiehl, T.-R., Perez Ordonez, B., Ilaalagan, S., & Asa, S. L. (2009). Primary frozen section diagnosis by robotic microscopy and virtual slide telepathology : The University Health Network experience. *Human Pathology*, *40*(8), 1070-1081.

Evans, A. J., Kiehl, T.-R., & Croul, S. (2010). Frequently asked questions concerning the use of whole-slide imaging telepathology for neuropathology frozen sections. *Seminars in Diagnostic Pathology*, *27*(3), 160-166.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S. A., Nobel, A. B., van't Veer, L. J., & Perou, C. M. (2006). Concordance among gene-expression-based predictors for breast cancer. *The New England Journal of Medicine*, *355*(6), 560-569.

Farahani, N., & Pantanowitz, L. (2015). Overview of Telepathology. *Surgical Pathology Clinics*, *8*(2), 223-231.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., MacGrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.-L., Brisken, C., Fiche, M., Delorenzi, M., & Iggo, R. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*, *24*(29), 4660-4671.

Fereidouni, F., Harmany, Z. T., Tian, M., Todd, A., Kintner, J. A., McPherson, J. D., Borowsky, A. D., Bishop, J., Lechpammer, M., Demos, S. G., & Levenson, R. (2017). Microscopy with ultraviolet surface excitation for rapid slide-free histology. *Nature Biomedical Engineering*, *1*(12), 957-966.

Ferreira, R., Moon, B., Humphries, J., Sussman, A., Saltz, J., Miller, R., & Demarzo, A. (1997). The Virtual Microscope. *Proceedings of the AMIA Annual Fall Symposium*, 449-453.

Finberg, K. E., Sequist, L. V., Joshi, V. A., Muzikansky, A., Miller, J. M., Han, M., Beheshti, J., Chirieac, L. R., Mark, E. J., & Iafrate, A. J. (2007). Mucinous differentiation correlates with absence of EGFR mutation and presence of KRAS mutation in lung adenocarcinomas with bronchioloalveolar features. *The Journal of Molecular Diagnostics: JMD*, *9*(3), 320-326.

Fitzal, F., Filipits, M., Rudas, M., Greil, R., Dietze, O., Samonigg, H., Lax, S., Herz, W., Dubsky, P., Bartsch, R., Kronenwett, R., & Gnant, M. (2015). The genomic expression test EndoPredict is a prognostic tool for identifying risk of local recurrence in postmenopausal endocrine receptor-positive, her2neu-negative breast cancer patients randomised within the prospective ABCSG 8 trial. *British Journal of Cancer*, *112*(8), 1405-1410.

Fitzgibbons, P. L., Page, D. L., Weaver, D., Thor, A. D., Allred, D. C., Clark, G. M., Ruby, S. G., O'Malley, F., Simpson, J. F., Connolly, J. L., Hayes, D. F., Edge, S. B., Lichter, A., & Schnitt, S. J. (2000). Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. *Archives of Pathology & Laboratory Medicine*, *124*(7), 966-978.

Foulkes, W. D., Smith, I. E., & Reis-Filho, J. S. (2010). Triple-Negative Breast Cancer. *New England Journal of Medicine*, *363*(20), 1938-1948.

Frenois, F. X. (2019). *Pathologie digitale : Fondamentaux technologiques*. 7.

Galea, M. H., Blamey, R. W., Elston, C. E., & Ellis, I. O. (1992). The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment*, *22*(3), 207-219.

Garberis, I., Andre, F., & Lacroix-Triki, M. (2021). L'intelligence artificielle pourrait-elle intervenir dans l'aide au diagnostic des cancers du sein ? – L'exemple de HER2: Could artificial intelligence play a role in breast cancer diagnosis? – The example of HER2. *Bulletin Du Cancer*, *108*(11S), 11S35-11S45.

Gephardt, G. N., & Zarbo, R. J. (1996). Interinstitutional comparison of frozen section consultations. A college of American Pathologists Q-Probes study of 90,538 cases in 461 institutions. *Archives of Pathology & Laboratory Medicine*, *120*(9), 804-809.

Gill, J., Sullivan, R., Taylor, D., & UCL School of Pharmacy. (2015). *Overcoming cancer in the 21st century*. UCL School of Pharmacy.

Glaser, A. K., Reder, N. P., Chen, Y., McCarty, E. F., Yin, C., Wei, L., Wang, Y., True, L. D., & Liu, J. T. C. (2017). Light-sheet microscopy for slide-free non-destructive pathology of large clinical specimens. *Nature Biomedical Engineering*, *1*(7), 0084.

Gnant, M., Filipits, M., Greil, R., Stoeger, H., Rudas, M., Bago-Horvath, Z., Mlineritsch, B., Kwasny, W., Knauer, M., Singer, C., Jakesz, R., Dubsky, P., Fitzal, F., Bartsch, R., Steger, G., Balic, M., Ressler, S., Cowens, J. W., Storhoff, J., … Nielsen, T. O. (2014). Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk : Using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. *Annals of Oncology*, *25*(2), 339-345.

Goldhirsch, A., Winer, E. P., Coates, A. S., Gelber, R. D., Piccart-Gebhart, M., Thürlimann, B., Senn, H.-J., & Panel members. (2013). Personalizing the treatment of women with early breast cancer : Highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *24*(9), 2206-2223.

Goldhirsch, A., Wood, W. C., Coates, A. S., Gelber, R. D., Thürlimann, B., Senn, H.-J., & Panel members. (2011). Strategies for subtypes--dealing with the diversity of breast cancer : Highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *22*(8), 1736-1747.

Goode, A., Gilbert, B., Harkes, J., Jukic, D., & Satyanarayanan, M. (2013). OpenSlide : A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, *4*(1), 27.

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In D. E. Losada & J. M. Fernández-Luna (Éds.), *Advances in Information Retrieval* (p. 345-359). Springer.

Hammond, M. E. H., Hayes, D. F., Dowsett, M., Allred, D. C., Hagerty, K. L., Badve, S., Fitzgibbons, P. L., Francis, G., Goldstein, N. S., Hayes, M., Hicks, D. G., Lester, S., Love, R., Mangu, P. B., McShane, L., Miller, K., Osborne, C. K., Paik, S., Perlmutter, J., … Wolff, A. C. (2010). American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer (Unabridged Version). *Archives of Pathology & Laboratory Medicine*, *134*(7), e48-e72.

Hanna, M. G., Pantanowitz, L., & Evans, A. J. (2015). Overview of contemporary guidelines in digital pathology : What is available in 2015 and what still needs to be addressed? *Journal of Clinical Pathology*, *68*(7), 499-505.

Hartman, D., Laak, J., Gurcan, M., & Pantanowitz, L. (2020). Value of Public Challenges for the Development of Pathology Deep Learning Algorithms. *Journal of Pathology Informatics*, *11*, 7.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Heeke, S., Delingette, H., Fanjat, Y., Long-Mira, E., Lassalle, S., Hofman, V., Benzaquen, J., Marquette, C.-H., Hofman, P., & Ilié, M. (2019). La pathologie cancéreuse pulmonaire à l'heure de l'intelligence artificielle : Entre espoir, désespoir et perspectives. *Annales de Pathologie*, *39*(2), 130-136.

Heindl, A., Sestak, I., Naidoo, K., Cuzick, J., Dowsett, M., & Yuan, Y. (2018). Relevance of Spatial Heterogeneity of Immune Infiltration for Predicting Risk of Recurrence After Endocrine Therapy of ER+ Breast Cancer. *JNCI: Journal of the National Cancer Institute*, *110*(2), 166-175.

Helgason, C. M. (1987). Commentary on the significance for modern neurology of the 17th century B.C. Surgical Papyrus. *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques*, *14*(4), 560-563.

Henriques Abreu, P., Santos, M., Henriques Abreu, M., Aveleira Andrade, B., & Silva, D. (2016). Predicting Breast Cancer Recurrence Using Machine Learning Techniques : A Systematic Review. *ACM Computing Surveys*, *49*, 1-40.

Herrmann, M. D., Clunie, D. A., Fedorov, A., Doyle, S. W., Pieper, S., Klepeis, V., Le, L. P., Mutter, G. L., Milstone, D. S., Schultz, T. J., Kikinis, R., Kotecha, G. K., Hwang, D. H., Andriole, K. P., Iafrate, A. J., Brink, J. A., Boland, G. W., Dreyer, K. J., Michalski, M., … Lennerz, J. K. (2018). Implementing the DICOM Standard for Digital Pathology. *Journal of Pathology Informatics*, *9*, 37.

Hesterberg, T. (2011). Bootstrap. *WIREs Computational Statistics*, *3*(6), 497-526.

Hipp, J. D., Lucas, D. R., Emmert-Buck, M. R., Compton, C. C., & Balis, U. J. (2011). Digital Slide Repositories for Publications : Lessons Learned From the Microarray Community. *The American Journal of Surgical Pathology*, *35*(6), 783–786.

Holzinger, A. (2018). From Machine Learning to Explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 55-66.

Holzinger, A., Plass, M., Holzinger, K., Crisan, G. C., Pintea, C.-M., & Palade, V. (2017). A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. *arXiv:1708.01104 [cs, stat]*. http://arxiv.org/abs/1708.01104

Hossain, Md. S., Hanna, M. G., Uraoka, N., Nakamura, T., Edelweiss, M., Brogi, E., Hameed, M. R., Yamaguchi, M., Ross, D. S., & Yagi, Y. (2019). Automatic quantification of HER2 gene amplification in invasive breast cancer from chromogenic in situ hybridization whole slide images. *Journal of Medical Imaging*, *6*(4), 047501.

Hurk, C. J. G., Eckel, R., Poll-Franse, L. V., Coebergh, J. W. W., Nortier, J. W. R., Hölzel, D., Breed, W. P. M., & Engel, J. (2011). Unfavourable pattern of metastases in M0 breast cancer patients during 1978-2008 : A population-based analysis of the Munich Cancer Registry. *Breast Cancer Research and Treatment*, *128*(3), 795-805.

Ibrahim, A., Gamble, P., Jaroensri, R., Abdelsamea, M. M., Mermel, C. H., Chen, P.-H. C., & Rakha, E. A. (2020). Artificial intelligence in digital breast pathology : Techniques and applications. *The Breast*, *49*, 267-273.

Ignatiadis, M., Azim, H. A., Desmedt, C., Veys, I., Larsimont, D., Salgado, R., Lyng, M. B., Viale, G., Leyland-Jones, B., Giobbie-Hurder, A., Kammler, R., Dell'Orto, P., Rothé, F., Laïos, I., Ditzel, H. J., Regan, M. M., Piccart, M., Michiels, S., & Sotiriou, C. (2016). The Genomic Grade Assay Compared With Ki67 to Determine Risk of Distant Breast Cancer Recurrence. *JAMA Oncology*, *2*(2), 217-224.

Ilse, M., Tomczak, J., & Welling, M. (2018). Attention-based Deep Multiple Instance Learning. *Proceedings of the 35th International Conference on Machine Learning*, 2127-2136. https://proceedings.mlr.press/v80/ilse18a.html

Jaber, M. I., Song, B., Taylor, C., Vaske, C. J., Benz, S. C., Rabizadeh, S., Soon-Shiong, P., & Szeto, C. W. (2020). A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Research*, *22*(1), 12.

Jahn, S. W., Plass, M., & Moinfar, F. (2020). Digital Pathology : Advantages, Limitations and Emerging Perspectives. *Journal of Clinical Medicine*, *9*(11).

Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis : A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, *7*(1), 29.

Jara-Lazaro, A. R., Thamboo, T. P., Teh, M., & Tan, P. H. (2010). Digital pathology : Exploring its applications in diagnostic surgical pathology practice. *Pathology*, *42*(6), 512-518.

Jaroensri, R., Wulczyn, E., Hegde, N., Brown, T., Flament-Auvigne, I., Tan, F., Cai, Y., Nagpal, K., Rakha, E. A., Dabbs, D. J., Olson, N., Wren, J. H., Thompson, E. E., Seetao, E., Robinson, C., Miao, M., Beckers, F., Corrado, G. S., Peng, L. H., … Chen, P.-H. C. (2022). Deep learning models for histologic grading of breast cancer and association with disease prognosis. *Npj Breast Cancer*, *8*(1), 1-12.

Jiang, Y., Yang, M., Wang, S., Li, X., & Sun, Y. (2020). Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Communications*, *40*(4), 154-166.

Joyon, N., Penault-Llorca, F., & Lacroix-Triki, M. (2017). Classification et signatures moléculaires des cancers du sein en 2017. *Oncologie*, *19*(3), 64-70.

Kalinsky, K., Barlow, W. E., Gralow, J. R., Meric-Bernstam, F., Albain, K. S., Hayes, D. F., Lin, N. U., Perez, E. A., Goldstein, L. J., Chia, S. K. L., Dhesy-Thind, S., Rastogi, P., Alba, E., Delaloge, S., Martin, M., Kelly, C. M., Ruiz-Borrego, M., Gil-Gil, M., Arce-Salinas, C. H., … Hortobagyi, G. N. (2021). 21-Gene Assay to Inform Chemotherapy Benefit in Node-Positive Breast Cancer. *New England Journal of Medicine*, *385*(25), 2336-2347.

Kayser, K., GĂśrtler, Jă., Bogovac, M., Bogovac, A., Goldmann, T., Vollmer, E., & Kayser, G. (2009). AI (artificial intelligence) in histopathology—From image analysis to automated diagnosis. *Folia Histochemica et Cytobiologica*, *47*(3), 355-361.

Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y., & Park, S. H. (2019). Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images : Results from Recently Published Papers. *Korean Journal of Radiology*, *20*(3), 405-410.

Klimov, S., Miligy, I. M., Gertych, A., Jiang, Y., Toss, M. S., Rida, P., Ellis, I. O., Green, A., Krishnamurti, U., Rakha, E. A., & Aneja, R. (2019). A whole slide image-based machine learning approach to predict ductal carcinoma in situ (DCIS) recurrence risk. *Breast Cancer Research: BCR*, *21*(1), 83.

Koelzer, V. H., Sirinukunwattana, K., Rittscher, J., & Mertz, K. D. (2019). Precision immunoprofiling by image analysis and artificial intelligence. *Virchows Archiv: An International Journal of Pathology*, *474*(4), 511-522.

Koh, J., & Kim, M. J. (2019). Introduction of a New Staging System of Breast Cancer for Radiologists : An Emphasis on the Prognostic Stage. *Korean Journal of Radiology*, *20*(1), 69-82.

Kohavi, R. (2001). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. *14*.

Komura, D., & Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, *16*, 34-42.

Kothari, S., Phan, J. H., Stokes, T. H., & Wang, M. D. (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association: JAMIA*, *20*(6), 1099-1108.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Kumar, V., Abbas, A. K., & Aster, J. C. (2017). *Robbins Basic Pathology*. Elsevier. https://www.us.elsevierhealth.com/robbins-kumar-basic-pathology-9780323790185.html

La Barbera, D., Polónia, A., Roitero, K., Conde-Sousa, E., & Della Mea, V. (2020). Detection of HER2 from Haematoxylin-Eosin Slides Through a Cascade of Deep Learning Classifiers via Multi-Instance Learning. *Journal of Imaging*, *6*(9), 82.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

LeCun, Y., Bottou, L., Bengio, Y., & Ha, P. (1998). *Gradient-Based Learning Applied to Document Recognition*. 46.

Lecuona, I. de, & Villalobos-Quesada, M. (2018). European perspectives on big data applied to health : The case of biobanks and human databases. *Developing World Bioethics*, *18*(3), 291-298.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, *121*(7), 2750-2767.

Li, H., Bera, K., Toro, P., Fu, P., Zhang, Z., Lu, C., Feldman, M., Ganesan, S., Goldstein, L. J., Davidson, N. E., Glasgow, A., Harbhajanka, A., Gilmore, H., & Madabhushi, A. (2021). Collagen fiber orientation disorder from H&E images is prognostic for early stage breast cancer : Clinical trial validation. *Npj Breast Cancer*, *7*(1), 1-10.

Li, W., Germain, R. N., & Gerner, M. Y. (2017). Multiplex, quantitative cellular analysis in large tissue volumes with clearing-enhanced 3D microscopy (Ce3D). *Proceedings of the National Academy of Sciences*, *114*(35), E7321-E7330.

Li, W., Germain, R. N., & Gerner, M. Y. (2019). High-dimensional cell-level analysis of tissues with Ce3D multiplex volume imaging. *Nature Protocols*, *14*(6), 1708-1733.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q. F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., & van der Laak, J. (2018). 1399 H&E-stained sentinel lymph node sections of breast cancer patients : The CAMELYON dataset. *GigaScience*, *7*(6).

Liu, Y., Chen, P.-H. C., Krause, J., & Peng, L. (2019). How to Read Articles That Use Machine Learning : Users' Guides to the Medical Literature. *JAMA*, *322*(18), 1806.

Lu, C., Xu, H., Xu, J., Gilmore, H., Mandal, M., & Madabhushi, A. (2016). Multi-Pass Adaptive Voting for Nuclei Detection in Histopathological Images. *Scientific Reports*, *6*. Scopus.

Lukong, K. E. (2017). Understanding breast cancer—The long and winding road. *BBA Clinical*, *7*, 64-77.

Ma, X.-J., Salunga, R., Dahiya, S., Wang, W., Carney, E., Durbecq, V., Harris, A., Goss, P., Sotiriou, C., Erlander, M., & Sgroi, D. (2008). A Five-Gene Molecular Grade Index and HOXB13:IL17BR Are Complementary Prognostic Factors in Early Stage Breast Cancer. *Clinical Cancer Research*, *14*(9), 2601-2608.

Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology : Challenges and opportunities. *Medical Image Analysis*, *33*, 170-175.

Mallat, S. (2016). Understanding Deep Convolutional Networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150203.

Mantovani, A., Allavena, P., Sica, A., & Balkwill, F. (2008). Cancer-related inflammation. *Nature*, *454*(7203), 436-444.

Marchiò, C., Iravani, M., Natrajan, R., Lambros, M. B., Savage, K., Tamber, N., Fenwick, K., Mackay, A., Senetta, R., Palma, S. D., Schmitt, F. C., Bussolati, G., Ellis, I. O., Ashworth, A., Sapino, A., & Reis-Filho, J. S. (2008). Genomic and immunophenotypical characterization of pure micropapillary carcinomas of the breast. *The Journal of Pathology*, *215*(4), 398-410.

McClelland, C. (2017, décembre 4). The Difference Between Artificial Intelligence, Machine Learning, and Deep Learning. *Medium*. https://medium.com/iotforall/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning-3aa67bff5991

Merino Bonilla, J. A., Torres Tabanera, M., & Ros Mendoza, L. H. (2017). Breast cancer in the 21st century : From early detection to new therapies. *Radiologia*, *59*(5), 368-379.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Vega, J. E. V., Brat, D. J., & Cooper, L. A. D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, *115*(13), E2970-E2979.

Mohaiminul Islam, Md., Huang, S., Ajwad, R., Chi, C., Wang, Y., & Hu, P. (2020). An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Computational and Structural Biotechnology Journal*, *18*, 2185-2199.

Mutasa, S., Chang, P., Nemer, J., Van Sant, E. P., Sun, M., McIlvride, A., Siddique, M., & Ha, R. (2020). Prospective Analysis Using a Novel CNN Algorithm to Distinguish Atypical Ductal Hyperplasia From Ductal Carcinoma in Situ in Breast. *Clinical Breast Cancer*, 757-760.

Naito, Y., & Urasaki, T. (2018). Precision medicine in breast cancer. *Chinese Clinical Oncology*, *7*(3), 8-8.

Nederlof, I., Hajizadeh, S., Sobhani, F., Raza, S. E. A., AbdulJabbar, K., Harkes, R., van de Vijver, M. J., Salgado, R., Desmedt, C., Kok, M., Yuan, Y., & Horlings, H. M. (2022). Spatial interplay of lymphocytes and fibroblasts in estrogen receptor-positive HER2-negative breast cancer. *NPJ Breast Cancer*, *8*(1), 56.

Niazi, M. K. K., Parwani, A. V., & Gurcan, M. N. (2019). Digital pathology and artificial intelligence. *The Lancet Oncology*, *20*(5), e253-e261.

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., Cheang, M. C. U., Mardis, E. R., Perou, C. M., Bernard, P. S., & Ellis, M. J. (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *16*(21), 5222-5232.

Nielsen, T., Wallden, B., Schaper, C., Ferree, S., Liu, S., Gao, D., Barry, G., Dowidar, N., Maysuria, M., & Storhoff, J. (2014). Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC Cancer*, *14*, 177.

Olivotto, I. A., Bajdik, C. D., Ravdin, P. M., Speers, C. H., Coldman, A. J., Norris, B. D., Davis, G. J., Chia, S. K., & Gelmon, K. A. (2005). Population-based validation of the prognostic model ADJUVANT! for early breast cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *23*(12), 2716-2725.

Osareh, A., & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. *2010 5th International Symposium on Health Informatics and Bioinformatics*, 114-120.

Pagès, F., Mlecnik, B., Marliot, F., Bindea, G., Ou, F.-S., Bifulco, C., Lugli, A., Zlobec, I., Rau, T. T., Berger, M. D., Nagtegaal, I. D., Vink-Börger, E., Hartmann, A., Geppert, C., Kolwelter, J., Merkel, S., Grützmann, R., Van den Eynde, M., Jouret-Mourin, A., … Galon, J. (2018). International validation of the consensus Immunoscore for the classification of colon cancer : A prognostic and accuracy study. *Lancet (London, England)*, *391*(10135), 2128-2139.

Paik, S., Kim, C., Baehner, F. L., Park, T., Wickerham, D. L., & Wolmark, N. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *The New England Journal of Medicine*, 10.

Pantanowitz, L., Dickinson, K., Evans, A. J., Hassell, L. A., Henricks, W. H., Lennerz, J. K., Lowe, A., Parwani, A. V., Riben, M., Smith, D., Tuthill, J. M., Weinstein, R. S., Wilbur, D. C., Krupinski, E. A., & Bernard, J. (2014). American Telemedicine Association clinical guidelines for telepathology. *Journal of Pathology Informatics*, *5*(1), 39.

Pantanowitz, L., Hartman, D., Qi, Y., Cho, E. Y., Suh, B., Paeng, K., Dhir, R., Michelow, P., Hazelhurst, S., Song, S. Y., & Cho, S. Y. (2020). Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagnostic Pathology*, *15*(1), 80.

Pantanowitz, L., Valenstein, P. N., Evans, A. J., Kaplan, K. J., Pfeifer, J. D., Wilbur, D. C., Collins, L. C., & Colgan, T. J. (2011). Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, *2*.

Papavramidou, N., Papavramidis, T., & Demetriou, T. (2010). Ancient Greek and Greco–Roman Methods in Modern Surgical Treatment of Cancer. *Annals of Surgical Oncology*, *17*(3), 665-667.

Park, S. H. (2018). Diagnostic Case-Control versus Diagnostic Cohort Studies for Clinical Validation of Artificial Intelligence Algorithm Performance. *Radiology*, *290*(1), 272-273.

Park, S. H., & Han, K. (2018). Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*, *286*(3), 800-809.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., & Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *27*(8), 1160-1167.

Patel, A., Balis, U. G. J., Cheng, J., Li, Z., Lujan, G., McClintock, D. S., Pantanowitz, L., & Parwani, A. (2021). Contemporary Whole Slide Imaging Devices and Their Applications within the Modern Pathology Department : A Selected Hardware Review. *Journal of Pathology Informatics*, *12*(1), 50.

Penault-Llorca, F. (2021). *Désescalade, désescalades, et si on parlait aussi de RxPONDER ?* 3.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747-752.

Petrick, N., Akbar, S., Cha, K. H., Nofech-Mozes, S., Sahiner, B., Gavrielides, M. A., Kalpathy-Cramer, J., Drukker, K., Martel, A. L., & BreastPathQ Challenge Group. (2021). SPIE-AAPM-NCI BreastPathQ challenge : An image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *Journal of Medical Imaging (Bellingham, Wash.)*, *8*(3), 034501.

Pollán, M., Michelena, M. J., Ardanaz, E., Izquierdo, A., Sánchez-Pérez, M. J., Torrella, A., & Breast Cancer Working Group. (2010). Breast cancer incidence in Spain before, during and after the implementation of screening programmes. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *21 Suppl 3*, iii97-102.

Prat, A., & Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*, *5*(1), 5-23.

Qaiser, T., Mukherjee, A., Pb, C. R., Munugoti, S. D., Tallam, V., Pitkäaho, T., Lehtimäki, T., Naughton, T., Berseth, M., Pedraza, A., Mukundan, R., Smith, M., Bhalerao, A., Rodner, E., Simon, M., Denzler, J., Huang, C.-H., Bueno, G., Snead, D., … Rajpoot, N. (2018). HER2 challenge contest : A detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology*, *72*(2), 227-238.

Qaiser, T., & Rajpoot, N. M. (2019). Learning Where to See : A Novel Attention Model for Automated Immunohistochemical Scoring. *IEEE Transactions on Medical Imaging*, *38*(11), 2620-2631.

Rakha, E. A. (2013). Pitfalls in outcome prediction of breast cancer. *Journal of Clinical Pathology*, *66*(6), 458-464.

Rakha, E. A., Alsaleem, M., ElSharawy, K. A., Toss, M. S., Raafat, S., Mihai, R., Minhas, F. A., Green, A. R., Rajpoot, N. M., Dalton, L. W., & Mongan, N. P. (2020). Visual histological assessment of morphological features reflects the underlying molecular profile in invasive breast cancer : A morphomolecular study. *Histopathology*, *77*(4), 631-645.

Rakha, E. A., El-Sayed, M. E., Lee, A. H. S., Elston, C. W., Grainge, M. J., Hodi, Z., Blamey, R. W., & Ellis, I. O. (2008). Prognostic Significance of Nottingham Histologic Grade in Invasive Breast Carcinoma. *Journal of Clinical Oncology*, *26*(19), 3153-3158.

Rayter, Z., & Mansi, J. (Éds.). (2003). *Medical therapy of breast cancer*. Cambridge University Press.

Reis-Filho, J. S., & Lakhani, S. R. (2008). Breast cancer special types : Why bother? *The Journal of Pathology*, *216*(4), 394-398.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer : Classification, prognostication, and prediction. *The Lancet*, *378*(9805), 1812-1823.

Ribback, S., Flessa, S., Gromoll-Bergmann, K., Evert, M., & Dombrowski, F. (2014). Virtual slide telepathology with scanner systems for intraoperative frozen-section consultation. *Pathology - Research and Practice*, *210*(6), 377-382.

Ribelles, N., Perez-Villa, L., Jerez, J. M., Pajares, B., Vicioso, L., Jimenez, B., de Luque, V., Franco, L., Gallego, E., Marquez, A., Alvarez, M., Sanchez-Muñoz, A., Perez-Rivas, L., & Alba, E. (2013). Pattern of recurrence of early breast cancer is different according to intrinsic subtype and proliferation index. *Breast Cancer Research : BCR*, *15*(5), R98.

Riber-Hansen, R., Vainer, B., & Steiniche, T. (2012). Digital image analysis : A review of reproducibility, stability and basic requirements for optimal results. *APMIS: Acta Pathologica, Microbiologica, et Immunologica Scandinavica*, *120*(4), 276-289.

Robertson, S., Azizpour, H., Smith, K., & Hartman, J. (2018). Digital image analysis in breast pathology—From image processing techniques to artificial intelligence. *Translational Research*, *194*, 19-35.

Rojo, M. G., Bueno, G., & Slodkowska, J. (2009). Review of imaging solutions for integrated quantitative immunohistochemistry in the Pathology daily practice. *Folia Histochemica Et Cytobiologica*, *47*(3), 349-354.

Rojo, M. G., García, G. B., Mateos, C. P., García, J. G., & Vicente, M. C. (2006). Critical comparison of 31 commercially available digital slide systems in pathology. *International Journal of Surgical Pathology*, *14*(4), 285-305.

Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G., & Gurcan, M. N. (2013). Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics*, *4*, 8.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Ryad Zemouri, Devalland, C., Valmary-Degano, S., & Zerhouni, N. (2019). Intelligence artificielle : Quel avenir en anatomie pathologique ? *Annales de Pathologie*, *39*(2), 119-129.

Saha, M., Chakraborty, C., Arun, I., Ahmed, R., & Chatterjee, S. (2017). An Advanced Deep Learning Approach for Ki-67 Stained Hotspot Detection and Proliferation Rate Scoring for Prognostic Evaluation of Breast Cancer. *Scientific Reports*, *7*(1), 3213.

Sahiner, B., Tozbikian, G., Lozanski, G., Gurcan, M., & Senaras, C. (2018). *Creating synthetic digital slides using conditional generative adversarial networks : Application to Ki67 staining*.

Salaverry, O. (2013). La etimología del cáncer y su curioso curso histórico. *Rev Peru Med Exp Salud Publica*, *30*(1), 137-141.

Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F. L., Penault-Llorca, F., Perez, E. A., Thompson, E. A., Symmans, W. F., Richardson, A. L., Brock, J., Criscitiello, C., Bailey, H., Ignatiadis, M., Floris, G., … Loi, S. (2015). The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer : Recommendations by an International TILs Working Group 2014. *Annals of Oncology*, *26*(2), 259-271.

Salto-Tellez, M., James, J. A., & Hamilton, P. W. (2014). Molecular pathology – The value of an integrative approach. *Molecular Oncology*, *8*(7), 1163-1168.

Salto-Tellez, M., Maxwell, P., & Hamilton, P. (2019). Artificial intelligence-the third revolution in pathology. *Histopathology*, *74*(3), 372-376.

Samek, W., Binder, A., Montavon, G., Bach, S., & Müller, K.-R. (2015). Evaluating the visualization of what a Deep Neural Network has learned. *arXiv:1509.06321 [cs]*. http://arxiv.org/abs/1509.06321

Schmidhuber, J. (2015). Deep Learning in Neural Networks : An Overview. *Neural Networks*, *61*, 85-117.

Serag, A., Ion-Margineanu, A., Qureshi, H., McMillan, R., Saint Martin, M.-J., Diamond, J., O'Reilly, P., & Hamilton, P. (2019). Translational AI and Deep Learning in Diagnostic Pathology. *Frontiers in Medicine*, *6*.

Sestak, I., Buus, R., Cuzick, J., Dubsky, P., Kronenwett, R., Denkert, C., Ferree, S., Sgroi, D., Schnabel, C., Baehner, F. L., Mallon, E., & Dowsett, M. (2018). Comparison of the Performance of 6 Prognostic Signatures for Estrogen Receptor–Positive Breast Cancer : A Secondary Analysis of a Randomized Clinical Trial. *JAMA Oncology*, *4*(4), 545.

Sgroi, D. C., Sestak, I., Cuzick, J., Zhang, Y., Schnabel, C. A., Schroeder, B., Erlander, M. G., Dunbier, A., Sidhu, K., Lopez-Knowles, E., Goss, P. E., & Dowsett, M. (2013). Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer : A prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the TransATAC study population. *The Lancet Oncology*, *14*(11), 1067-1076.

Shamai, G., Binenbaum, Y., Slossberg, R., Duek, I., Gil, Z., & Kimmel, R. (2019). Artificial Intelligence Algorithms to Assess Hormonal Status From Tissue Microarrays in Patients With Breast Cancer. *JAMA Network Open*, *2*(7).

Shim, H. J., Kim, S. H., Kang, B. J., Choi, B. G., Kim, H. S., Cha, E. S., & Song, B. J. (2014). Breast Cancer Recurrence According to Molecular Subtype. *Asian Pacific Journal of Cancer Prevention*, *15*(14), 5539-5544.

Singh, R., Chubb, L., Pantanowitz, L., & Parwani, A. (2011). Standardization in digital pathology : Supplement 145 of the DICOM standards. *Journal of Pathology Informatics*, *2*, 23.

Sinn, H.-P., & Kreipe, H. (2013). A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. *Breast Care*, *8*(2), 149-154.

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., & McGuire, W. L. (1987). Human breast cancer : Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, *235*(4785), 177-182.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC : A Family of Discriminant Measures for Performance Evaluation. In *AI 2006 : Advances in Artificial Intelligence, Lecture Notes in Computer Science: Vol. Vol. 4304* (p. 1021).

Sørlie, T. (2004). Molecular portraits of breast cancer : Tumour subtypes as distinct disease entities. *European Journal of Cancer (Oxford, England: 1990)*, *40*(18), 2667-2675.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, *100*(14), 8418-8423.

Sotiriou, C., & Pusztai, L. (2009). Gene-expression signatures in breast cancer. *The New England Journal of Medicine*, *360*(8), 790-800.

Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., Geyer, C. E., Dees, E. C., Perez, E. A., Olson, J. A., Zujewski, J., Lively, T., Badve, S. S., Saphner, T. J., Wagner, L. I., Whelan, T. J., Ellis, M. J., Paik, S., Wood, W. C., … Sledge, G. W. (2015). Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *New England Journal of Medicine*, *373*(21), 2005-2014.

Stack, E. C., Wang, C., Roman, K. A., & Hoyt, C. C. (2014). Multiplexed immunohistochemistry, imaging, and quantitation : A review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods*, *70*(1), 46-58.

Stålhammar, G., Fuentes Martinez, N., Lippert, M., Tobin, N. P., Mølholm, I., Kis, L., Rosin, G., Rantalainen, M., Pedersen, L., Bergh, J., Grunkin, M., & Hartman, J. (2016). Digital image analysis outperforms manual biomarker assessment in breast cancer. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *29*(4), 318-329.

Stålhammar, G., Robertson, S., Wedlund, L., Lippert, M., Rantalainen, M., Bergh, J., & Hartman, J. (2018). Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. *Histopathology*, *72*(6), 974-989.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209-249.

Tan, P. H., Ellis, I., Allison, K., Brogi, E., Fox, S. B., Lakhani, S., Lazar, A. J., Morris, E. A., Sahin, A., Salgado, R., Sapino, A., Sasano, H., Schnitt, S., Sotiriou, C., Diest, P. van, White, V. A., Lokuhetty, D., & Cree, I. A. (2020). The 2019 World Health Organization classification of tumours of the breast. *Histopathology*, *77*(2), 181-185.

Tan, W. C. C., Nerurkar, S. N., Cai, H. Y., Ng, H. H. M., Wu, D., Wee, Y. T. F., Lim, J. C. T., Yeong, J., & Lim, T. K. H. (2020). Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications*, *40*(4), 135-153.

Têtu, B., Perron, É., Louahlia, S., Paré, G., Trudel, M.-C., & Meyer, J. (2014). The Eastern Québec Telepathology Network : A three-year experience of clinical diagnostic services. *Diagnostic Pathology*, *9*(Suppl 1), S1.

Tewary, S., & Mukhopadhyay, S. (2021). HER2 Molecular Marker Scoring Using Transfer Learning and Decision Level Fusion. *Journal of Digital Imaging*.

Tian, S., Roepman, P., Van't Veer, L. J., Bernards, R., de Snoo, F., & Glas, A. M. (2010). Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer. *Biomarker Insights*, *5*, 129-138.

Tian, T., Yang, Z., & Li, X. (2021). Tissue clearing technique : Recent progress and biomedical applications. *Journal of Anatomy*, *238*(2), 489-507.

Tian, Y., Chen, X., & Ganguli, S. (2021). Understanding self-supervised learning dynamics without contrastive pairs. *Proceedings of the 38th International Conference on Machine Learning*, 10268-10278. https://proceedings.mlr.press/v139/tian21a.html

Tizhoosh, H. R., & Pantanowitz, L. (2018). Artificial Intelligence and Digital Pathology : Challenges and Opportunities. *Journal of Pathology Informatics*, *9*.

Tőkés, T., Tőkés, A.-M., Szentmártoni, G., Kiszner, G., Madaras, L., Kulka, J., Krenács, T., & Dank, M. (2016). Expression of cell cycle markers is predictive of the response to primary systemic therapy of locally advanced breast cancer. *Virchows Archiv*, *468*(6), 675-686.

Topol, E. J. (2019). High-performance medicine : The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44.

Torti, D., & Trusolino, L. (2011). Oncogene addiction as a foundational rationale for targeted anti-cancer therapy : Promises and perils. *EMBO Molecular Medicine*, *3*(11), 623-636.

Turkki, R., Byckhov, D., Lundin, M., Isola, J., Nordling, S., Kovanen, P. E., Verrill, C., von Smitten, K., Joensuu, H., Lundin, J., & Linder, N. (2019). Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast Cancer Research and Treatment*, *177*(1), 41-52.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., … Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, *347*(25), 1999-2009.

van den Tweel, J. G., & Taylor, C. R. (2010). A brief history of pathology. *Virchows Archiv*, *457*(1), 3-10.

van der Loos, C. M. (2008). Multiple Immunoenzyme Staining : Methods and Visualizations for the Observation With Spectral Imaging. *Journal of Histochemistry and Cytochemistry*, *56*(4), 313-328.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530-536.

Vandenberghe, M. E., Scott, M. L. J., Scorer, P. W., Söderberg, M., Balcerzak, D., & Barker, C. (2017). Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, *7*.

Vaz-Luis, I., Cottu, P., Mesleard, C., Martin, A. L., Dumas, A., Dauchy, S., Tredan, O., Levy, C., Adnet, J., Rousseau Tsangaris, M., Andre, F., & Arveux, P. (2019). UNICANCER : French prospective cohort study of treatment-related chronic toxicity in women with localised breast cancer (CANTO). *ESMO Open*, *4*(5), e000562.

Veta, M., Heng, Y. J., Stathonikos, N., Bejnordi, B. E., Beca, F., Wollmann, T., Rohr, K., Shah, M. A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H. A., … Pluim, J. P. W. (2019). Predicting breast tumor proliferation from whole-slide images : The TUPAC16 challenge. *Medical Image Analysis*, *54*, 111-121.

Veta, M., van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A. B. L., Vestergaard, J. S., Dahl, A. B., Cireşan, D. C., Schmidhuber, J., Giusti, A., Gambardella, L. M., Tek, F. B., Walter, T., Wang, C.-W., Kondo, S., Matuszewski, B. J., … Pluim, J. P. W. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, *20*(1), 237-248.

Vodovnik, A., & Aghdam, M. R. F. (2018). Complete Routine Remote Digital Pathology Services. *Journal of Pathology Informatics*, *9*.

Wagner, R. P. (1999). Anecdotal, historical and critical commentaries on genetics. Rudolph Virchow and the genetic basis of somatic ecology. *Genetics*, *151*(3), 917-920.

Webster, J. D., & Dunstan, R. W. (2014). Whole-slide imaging and automated image analysis : Considerations and opportunities in the practice of pathology. *Veterinary Pathology*, *51*(1), 211-223.

Weigelt, B., Horlings, H. M., Kreike, B., Hayes, M. M., Hauptmann, M., Wessels, L. F. A., Jong, D. de, Vijver, M. V. de, Veer, L. V., & Peterse, J. L. (2008). Refinement of breast cancer classification by molecular characterization of histological special types. *The Journal of Pathology*, *216*(2), 141-150.

Weigelt, B., Kreike, B., & Reis-Filho, J. S. (2009). Metaplastic breast carcinomas are basal-like breast cancers : A genomic profiling analysis. *Breast Cancer Research and Treatment*, *117*(2), 273-280.

Weigelt, B., & Reis-Filho, J. S. (2009). Histological and molecular types of breast cancer : Is there a unifying taxonomy? *Nature Reviews. Clinical Oncology*, *6*(12), 718-730.

Weinstein, R. S., Graham, A. R., Richter, L. C., Barker, G. P., Krupinski, E. A., Lopez, A. M., Erps, K. A., Bhattacharyya, A. K., Yagi, Y., & Gilbertson, J. R. (2009). Overview of telepathology, virtual microscopy, and whole slide imaging : Prospects for the future. *Human Pathology*, *40*(8), 1057-1069.

Whitney, J., Corredor, G., Janowczyk, A., Ganesan, S., Doyle, S., Tomaszewski, J., Feldman, M., Gilmore, H., & Madabhushi, A. (2018). Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. *BMC Cancer*, *18*(1), 610.

Williams, B. J., & Treanor, D. (2020). Practical guide to training and validation for primary diagnosis with digital pathology. *Journal of Clinical Pathology*, *73*(7), 418-422.

Wishart, G. C., Azzato, E. M., Greenberg, D. C., Rashbass, J., Kearins, O., Lawrence, G., Caldas, C., & Pharoah, P. D. P. (2010). PREDICT : A new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research: BCR*, *12*(1), R1.

Wishart, G. C., Bajdik, C. D., Azzato, E. M., Dicks, E., Greenberg, D. C., Rashbass, J., Caldas, C., & Pharoah, P. D. P. (2011). A population-based validation of the prognostic model PREDICT for early breast cancer. *European Journal of Surgical Oncology*, *37*(5), 411-417.

Wishart, G. C., Bajdik, C. D., Dicks, E., Provenzano, E., Schmidt, M. K., Sherman, M., Greenberg, D. C., Green, A. R., Gelmon, K. A., Kosma, V.-M., Olson, J. E., Beckmann, M. W., Winqvist, R., Cross, S. S., Severi, G., Huntsman, D., Pylkäs, K., Ellis, I., Nielsen, T. O., … Pharoah, P. D. P. (2012). PREDICT Plus : Development and validation of a prognostic model for early breast cancer that includes HER2. *British Journal of Cancer*, *107*(5), 800-807.

Wolff, A. C., Hammond, M. E. H., Allison, K. H., Harvey, B. E., Mangu, P. B., Bartlett, J. M. S., Bilous, M., Ellis, I. O., Fitzgibbons, P., Hanna, W., Jenkins, R. B., Press, M. F., Spears, P. A., Vance, G. H., Viale, G., McShane, L. M., & Dowsett, M. (2018). Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer : American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Journal of Clinical Oncology*, *36*(20), 2105-2122.

Wu, D., Hacking, S. M., Chavarria, H., Abdelwahed, M., & Nasim, M. (2022). Computational portraits of the tumoral microenvironment in human breast cancer. *Virchows Archiv: An International Journal of Pathology*.

Xie, P., Bilenko, M., Finley, T., Gilad-Bachrach, R., Lauter, K., & Naehrig, M. (2014). *Crypto-Nets : Neural Networks over Encrypted Data*.

Xie, Q., Faust, K., Ommeren, R. V., Sheikh, A., Djuric, U., & Diamandis, P. (2019). Deep learning for image analysis : Personalizing medicine closer to the point of care. *Critical Reviews in Clinical Laboratory Sciences*, *0*(0), 1-13.

Xu, J., Xue, K., & Zhang, K. (2019). Current status and future trends of clinical diagnoses via image-based deep learning. *Theranostics*, *9*(25), 7556-7565.

Xu, Z., Fernández Moro, C., Bozóky, B., & Zhang, Q. (2019). *GAN-based Virtual Re-Staining : A Promising Solution for Whole Slide Image Analysis*.

Yuan, Y. (2015). Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *Journal of The Royal Society Interface*, *12*(103), 20141153.

Zarella, M. D., Bowman;, D., Aeffner, F., Farahani, N., Xthona;, A., Absar, S. F., Parwani, A., Bui, M., & Hartman, D. J. (2018). A Practical Guide to Whole Slide Imaging : A White Paper From the Digital Pathology Association. *Archives of Pathology & Laboratory Medicine*, *143*(2), 222-234.

Zhu, W., Xie, L., Han, J., & Guo, X. (2020). The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers*, *12*(3).

Annexes

RlapsRisk Report

RAPPORT DE RÉSULTAT RlapsRisk BC

OWKIN

Id Lame [redacted]

Nom du rapport [redacted]_RlapsRiskBC
Création du rapport 02 Nov. 2022 15:07
Version logicielle 1.1.0-rc.10

RÉSULTAT

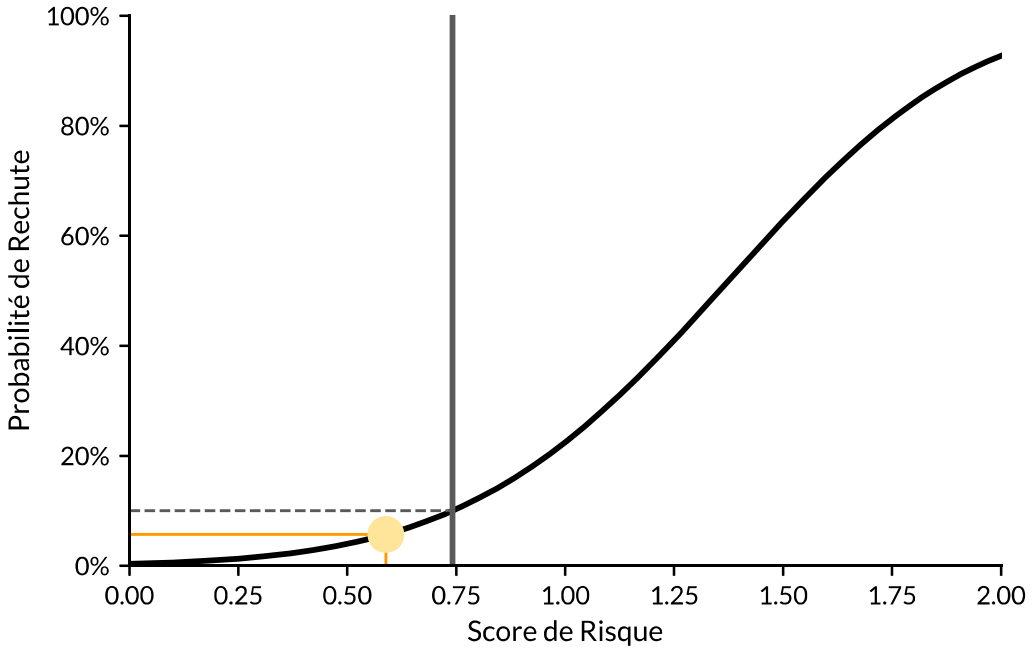
Bas risque

LECTURE DU RÉSULTAT

Le graphique suivant montre la relation entre le score recurisk et l'estimation de la probabilité de récidive à distance à 5 ans pour les patientes traitées pendant 5 ans par hormonothérapie adjuvante seule.

Probabilité de récidive à 5 ans 5.7%

Score de risque Owkin 0.588



Owkin Dx RlapsRisk BC est une analyse de lame HES par intelligence artificielle **pour les patientes atteintes d'un cancer du sein ER+, HER2-, à un stade précoce**. RlapsRisk BC a pour objectif d'informer le clinicien sur le risque de rechute.
Cette analyse d'image détermine un score de risque de rechute à partir duquel est identifiée la probabilité de récidive à distance à 5 ans avec 5 ans d'hormonothérapie adjuvante seule. Le résultat Haut risque ou Bas risque indique la catégorie de risque de récidive à distance avec 5 ans d'hormonothérapie adjuvante seule.

Les informations concernant l'utilisation, les performances cliniques et l'interprétation clinique des résultats sont présentées dans le manuel utilisateur.

RAPPORT DE RÉSULTAT RlapsRisk BC

OWKIN

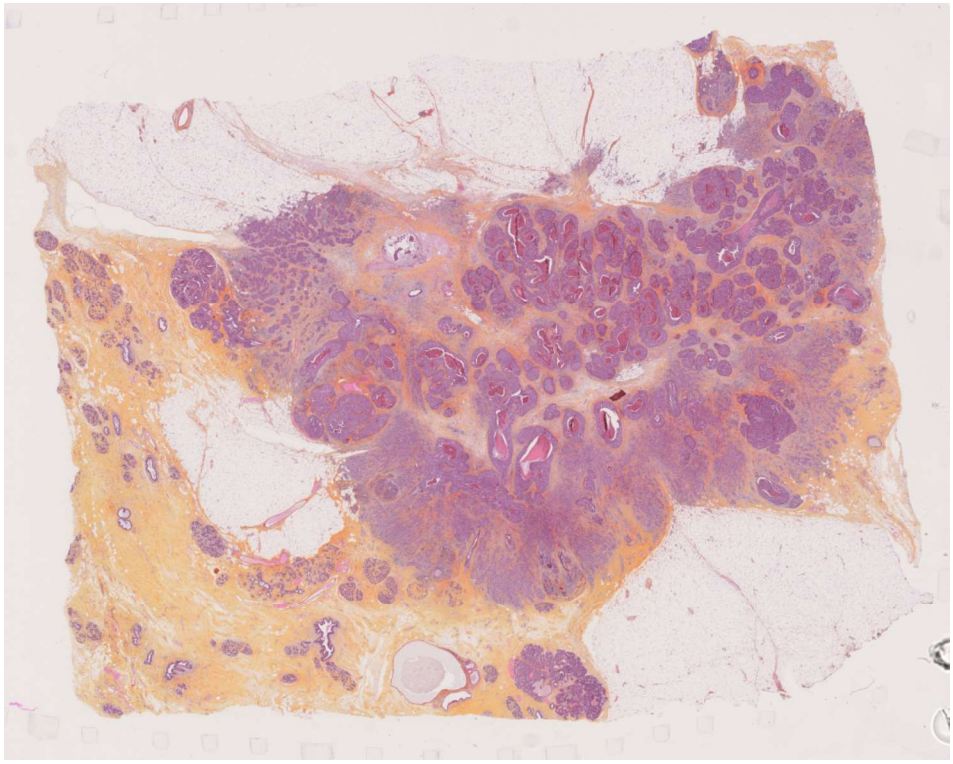
Id Lame	[redacted]	Nom du rapport	[redacted]_RlapsRiskBC
		Création du rapport	02 Nov. 2022 15:07
		Version logicielle	1.1.0-rc.10

INFORMATIONS SUPPLÉMENTAIRES

Surface totale estimée de tissu sur la lame	405.69mm^2
Surface estimée de tissu tumoral sur la lame	184.64mm^2
Taux de surface tumorale	46%

La surface estimée (tissu total ou de tissu tumoral) est obtenue par des algorithmes de segmentation appliqués à chaque image. La surface obtenue est une approximation et ne peut être utilisée en tant que valeur diagnostique.

LAME ANALYSÉE



Les informations concernant l'utilisation, les performances cliniques et l'interprétation clinique des résultats sont présentées dans le manuel utilisateur.

RAPPORT DE RÉSULTAT RlapsRisk BC

OWKIN

Id Lame [redacted]

Nom du rapport [redacted]_RlapsRiskBC
Création du rapport 02 Nov. 2022 17:40
Version logicielle 1.1.0-rc.10

RÉSULTAT

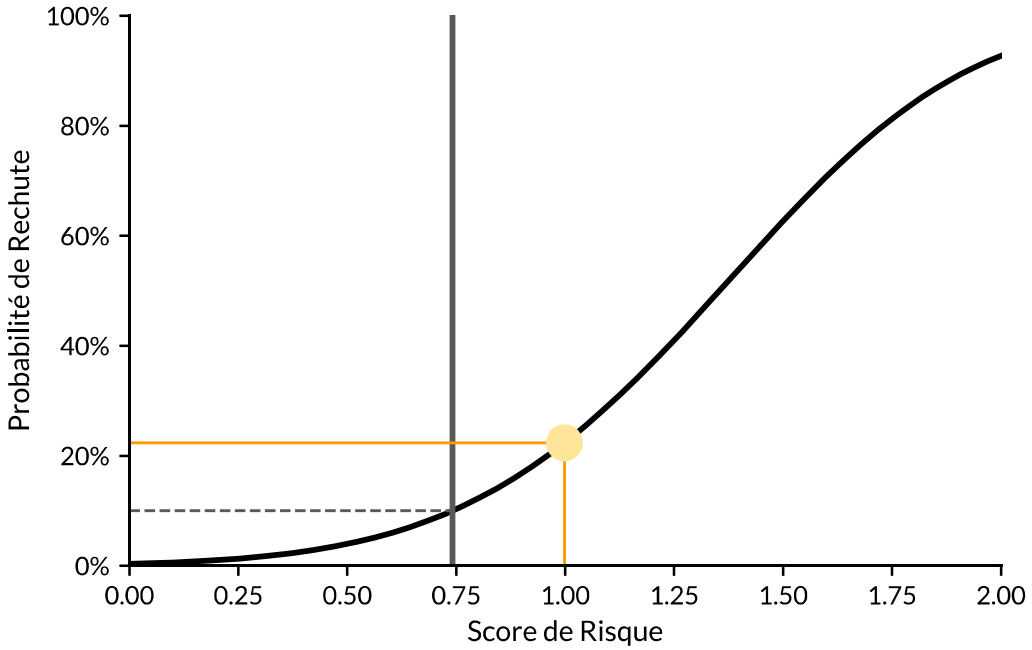
Haut risque

LECTURE DU RÉSULTAT

Le graphique suivant montre la relation entre le score recurisk et l'estimation de la probabilité de récidive à distance à 5 ans pour les patientes traitées pendant 5 ans par hormonothérapie adjuvante seule.

Probabilité de récidive à 5 ans 22.4%

Score de risque Owkin 0.998



Owkin Dx RlapsRisk BC est une analyse de lame HES par intelligence artificielle **pour les patientes atteintes d'un cancer du sein ER+, HER2-, à un stade précoce**. RlapsRisk BC a pour objectif d'informer le clinicien sur le risque de rechute.
Cette analyse d'image détermine un score de risque de rechute à partir duquel est identifiée la probabilité de récidive à distance à 5 ans avec 5 ans d'hormonothérapie adjuvante seule. Le résultat Haut risque ou Bas risque indique la catégorie de risque de récidive à distance avec 5 ans d'hormonothérapie adjuvante seule.

Les informations concernant l'utilisation, les performances cliniques et l'interprétation clinique des résultats sont présentées dans le manuel utilisateur.

RAPPORT DE RÉSULTAT RlapsRisk BC

OWKIN

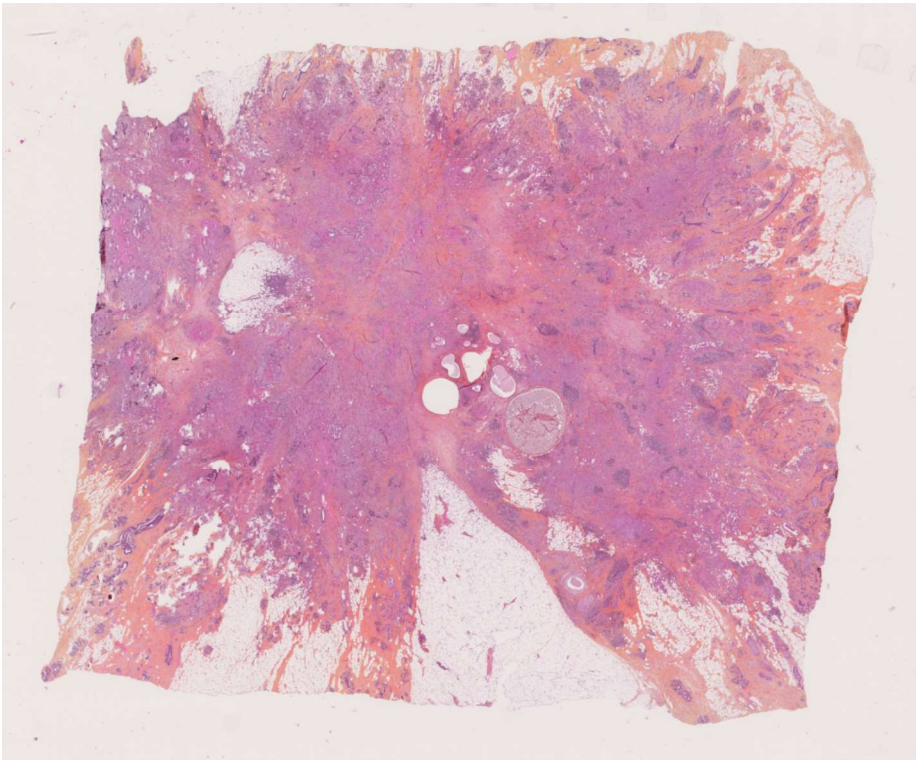
Id Lame	[redacted]	Nom du rapport	[redacted]_RlapsRiskBC
		Création du rapport	02 Nov. 2022 17:40
		Version logicielle	1.1.0-rc.10

INFORMATIONS SUPPLÉMENTAIRES

Surface totale estimée de tissu sur la lame	387.61mm^2
Surface estimée de tissu tumoral sur la lame	284.48mm^2
Taux de surface tumorale	73%

La surface estimée (tissu total ou de tissu tumoral) est obtenue par des algorithmes de segmentation appliqués à chaque image. La surface obtenue est une approximation et ne peut être utilisée en tant que valeur diagnostique.

LAME ANALYSÉE



Les informations concernant l'utilisation, les performances cliniques et l'interprétation clinique des résultats sont présentées dans le manuel utilisateur.

Communications on the RlapsRisk study

A. ESMO Congress 2021 – Paris, France

Proffered Paper oral presentation. "Prediction of distant relapse in patients with invasive breast cancer from deep learning models applied to digital pathology slides". Garberis I.*, Saillard C.*, Drubay D., Schmauch B., Aubert V., Jaeger A., Sapateiro M., de Lavergne A., Kamoun A., Courtiol P., André F., Lacroix-Triki M..

*both authors contributed equally to the study.

B. USCAP 2022 – Los Angeles, California, USA

Poster. "Explainable Deep Learning predicts molecular subtypes and improves risk of relapse assessment from invasive breast cancer histological slides". Saillard C.*, Garberis I.*, Drubay D., Gaury V., Aubert V., Schmauch B., Jaeger A., Herpin L., Elgui K., Linhart J., Kamoun A., André F., Lacroix-Triki M.

*both authors contributed equally to the study.

C. ESMO Congress 2022 – Paris, France

Poster. "Blind validation of an AI-based tool for predicting distant relapse from breast cancer HES stained slides". Garberis I.*, Gaury V.*, Drubay D., Saillard C., Aubert V., Elgui K., Bernigole F., Jacquet A., André F., Lacroix-Triki M.

*both authors contributed equally to the study.

Explainable Deep Learning predicts molecular subtypes and improves risk of relapse assessment from invasive breast cancer histological slides

Abstract #787 - USCAP 2022

Charlie Saillard*[1], Ingrid Garberis*[2], Damien Drubay[3], Valentin Gaury[1], Victor Aubert[1], Benoît Schmauch[1], Alexandre Jaeger[1], Loïc Herpin[1], Kevin Elgui[1], Julia Linhart[1], Aurélie Kamoun[1], Fabrice André[2,4], Magali Lacroix-Triki[5]

[1]Owkin, Paris, France; [2]INSERM UMR981, Gustave Roussy Cancer Campus, Villejuif, France; [3]Oncostat, Gustave Roussy Cancer Campus, Villejuif, France; [4]Department of Oncology, Gustave Roussy Cancer Campus, Villejuif, France; [5]Department of Pathology, Gustave Roussy Cancer Campus, Villejuif, France

Disclosures : No specific disclosures except for Owkin employees who work at Owkin and own Owkin stocks.

Background & Objectives

Breast cancer (BC) is a heterogeneous disease encompassing several subtypes associated to a **wide range of prognosis**.

Tumor molecular biomarkers such as estrogen receptor (ER), progesterone receptor (PR), HER2 and Ki67 are crucial for **treatment decision** and evaluation of the risk of relapse.

We developed 5 *explainable* deep learning (DL) models able to predict :

- Distant relapse from H&E-stained whole slide images (WSI) and routine clinical data
- ER, PR, HER2, Ki67 biomarkers from H&E WSI

→ Creation of a simple companion diagnostic tool, **applicable everywhere**, able to help treatment decision in clinical practice.

Cohort & Models

- Models developed on a cohort of **1802 patients** diagnosed with early BC (1429 ER+/HER2-, 110 ER+/HER2+, 70 ER-/HER2+, 193 ER-/HER2-) followed between 2005 and 2013 at Gustave Roussy.
- Patients underwent surgical resection, with at least 1 H&E digitized tumor slide per case, status of molecular biomarkers and survival follow-up available.
- Models trained to predict: i) the molecular biomarkers, and ii) the 5-year metastasis-free survival (MFS).
- Model performances were evaluated using cross-validation with: i) the area under ROC (AUROC), and ii) the Uno's AUC (UAUC).
- Comparison of multiple approaches for MFS prediction :
  - → From H&E slide alone (AI-H&E)
  - → From standard risk factors only
  - → Combining H&E slide and standard risk factors

Models Design

Our pipeline is composed of 2 steps: a preprocessing and a prediction step.

Preprocessing step

1. Tissue is segmented using UNet
2. Smaller patches ("tiles") are sampled uniformly within detected tissue (10,000 / slide)
3. Features are extracted from each tile using ResNet pretrained with MoCo v2 on millions of histology images

→ Each WSI is represented by a matrix 10,000 x 2048

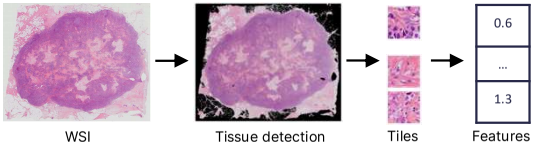


Figure 1. From left to right: overview of a WSI, tissue detection with UNet, tile sampling, feature extraction. Note: scale has not been respected.

Prediction step

For each biomarker and relapse prediction task, attention-based multiple instance learning (MIL) models are trained to :

1. Compute an attention score for each tile
2. Perform attention-weighted average of each slide's tile features
3. Predict the given task using a linear classifier on top of the averaged features

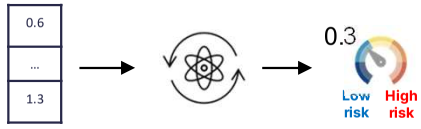


Figure 2. Features are used as input of a MIL model, trained to perform the given task (MFS here), outputting a probability score between 0 and 1.

Performance

Distant relapse

AI-H&E predicted MFS from H&E WSI in ER+HER2- patients with a **UAUC of 0.80**, outperforming standard risk factors (Age, pT, pN, tumor size, number of tumors, type of surgery; **Table 1**).

Combining AI-H&E with these clinical variables, performance further improved to **0.83 UAUC**.

Model	UAUC (std)
Clinical	0.78 (0.08)
AI-H&E	0.80 (0.06)
AI-H&E + Clinical	**0.83 (0.06)**

Table 1. Cross-validation results for MFS prediction from H&E WSI and/or clinical variables.

AI-H&E was still predictive of relapse in the histological grade 2 subgroup (UAUC=0.74), suggesting AI-H&E model goes **beyond the existing histology classification**.

Molecular Biomarkers

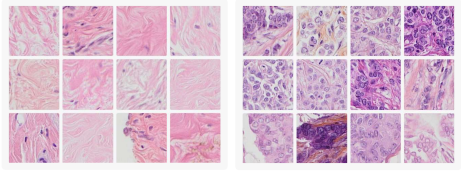
DL models were able to predict biomarkers with AUROCs ranging from 0.76 to 0.94 (**Table 2**).

Biomarker	AUROC (std)
ER	0.91 (0.02)
PR	0.76 (0.03)
Ki67	0.85 (0.02)
HER2	0.94 (0.02)

Table 2. Cross-validation results for biomarker prediction from H&E WSI

Interpretability

Histology regions predictive of high/low risk of relapse were reviewed by 2 pathologists blinded to the predicted outcome.



→ High relapse risk tiles displayed **high tumor cell content, strong nuclear atypia and massive architecture**.

→ Most contributive tiles of low risk corresponded to **fibrotic stroma with a few low-grade tumor cells**.

Conclusion

- DL models **based on a single H&E tumor WSI** show promising performance for biomarker status and relapse prediction that allows to refine the therapeutical strategies.
- AI-based score combined with existing clinic-pathological factors can **improve identification of early BC patients at high risk of relapse**.
- External validation on large independent cohorts and comparison with standard of care are in progress.
- Future steps in our work include :
  - ✓ unraveling the most predictive tiles to discover new biomarkers
  - ✓ developing a novel tool for prediction of relapse, as an alternative to onerous and not available everywhere techniques such as immunohistochemistry or molecular tests.

147P - Blind validation of an AI-based tool for predicting distant relapse from breast cancer HES stained slides

Ingrid Garberis[1], Valentin Gaury[2], Damien Drubay[3], Charlie Saillard[2], Victor Aubert[2], Kevin Elgui[2], Flavie Bernigole[2], Alexandra Jacquet[4], Fabrice André[5], Magali Lacroix-Triki[6]

[1]INSERM UMR981, Gustave Roussy Cancer Campus, Paris-Saclay University, Villejuif, France; [2]Owkin, Paris, France; [3]Department of Biostatistics and Epidemiology, Gustave Roussy Cancer Campus, Paris-Saclay University, Villejuif, France; and Oncostat INSERM UMR1018, Paris-Saclay University, Labeled Ligue Contre le Cancer, Villejuif, France; [4]Unicancer R&D, Unicancer, France; [5]Department of Medical Oncology, Gustave Roussy Cancer Campus, Villejuif, France; [6]Department of Pathology, Gustave Roussy Cancer Campus, Villejuif, France

Disclosures : No specific disclosures except for Owkin employees who work at Owkin and own Owkin stocks.

Background & objectives

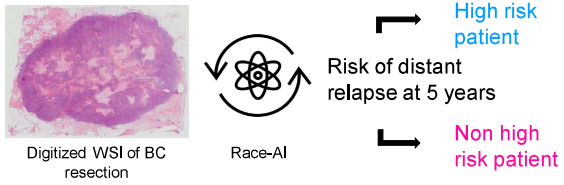
- Breast cancer (BC) is a heterogeneous disease encompassing several subtypes associated to a wide range of prognosis.
- Risk determination is crucial for treatment decision. We developed RlapsRisk™ BC, an AI-based tool that assesses the risk of distant relapse at 5 years of ER+/HER2- early invasive (ei)BC patients from HES (hematoxylin-eosin-safran)-stained whole slide images (WSI) and clinical data. Preliminary results of this tool were presented last year (Abstract 2392/1124O - ESMO2021).
- RlapsRisk™ BC was conceived as a companion diagnostic tool, applicable everywhere, able to help treatment decisions in clinical practice.

170

In the current study, we present a one-shot blind **external validation** of RlapsRisk™ BC.

Model

- RlapsRisk™ BC model was developed on the **GrandTMA** cohort with 1800 HES WSIs.
- It combines Self-Supervised Learning (Moco v2) to extract features from images, and a multiple instance learning model (Deepmil) to predict a risk of distant relapse.



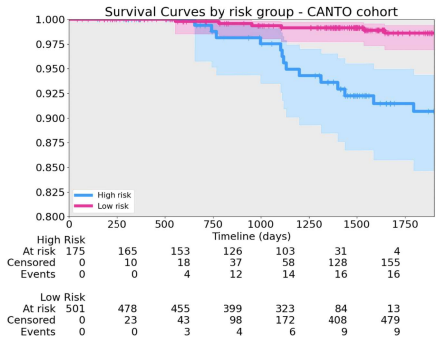
Validation study

- RlapsRisk™ BC validation dataset included 676 HES-stained WSIs from ER+/HER2- eiBC patients diagnosed at Gustave Roussy between 2012 and 2017, included in the CANTO cohort (NCT01993498, comprising 25 patients who relapsed at 5 years).
- We compared RlapsRisk™ BC performance to the two most relevant clinical scores: **Predict Breast and CTS0**.
- Model performances were evaluated through their cumulative sensitivity and dynamic specificity at 5 years to assess the accuracy of the scores to identify distant relapses.
- Each score has been dichotomized into low risk/high risk with respect to a threshold that has been set beforehand (5% for predict Breast, 1.40 for CTS0).

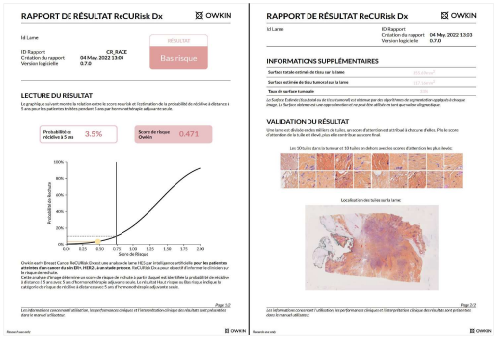
Results

Scores	Cumulative Sensitivity @5Y	Dynamic Specificity @5Y
RlapsRisk™ BC	**0.64 [0.57-0.72]**	**0.78 [0.76-0.80]**
Predict Breast	0.61 [0.52-0.70]	0.77 [0.75-0.79]
CTS0	0.43 [0.34-0.52]	0.80 [0.78-0.82]

- Among the non high risk population treated **without adjuvant chemotherapy**, **the distant relapse rate was 0.3%** (1 out of 324).
- The obtained results showed the ability of RlapsRisk™ BC to generalize on independent data and thus endorses the robustness of the method.



Patient report and interpretability



- High relapse risk tiles displayed **high tumor cell content, strong nuclear atypia and massive architecture**.
- Most contributive tiles of low risk corresponded to **fibrotic stroma with a few low-grade tumor cells**.

Conclusion

We performed the **first fully blind validation of RlapsRisk™ BC**, an AI-based tool to assess the risk of distant relapse.

Additional analyses validate the clinical value of RlapsRisk™ BC and suggest that it could be used for therapeutic de-escalation purposes. **RlapsRisk™ BC has been CE marked in May 2022.**

Future work

- Extension of validation to multi-site and multi-scanner eiBC WSIs from the CANTO cohort (under completion).
- Development of algorithms adapted to different slide conditions (such as HE staining or biopsy specimens) increasing the generalization capacities of our model.
- In-depth tiles analysis centered on the spatiality notion.

Study funded by Owkin, inc.

→ Contact: ingrid-judith.GARBERIS@gustaveroussy.fr

Other publications on the AI domain

A. Review article

- **Garberis I.**, André F., Lacroix-Triki M. "L'intelligence artificielle pourrait-elle intervenir dans l'aide au diagnostic des cancers du sein ? – L'exemple de HER2". *Bulletin du Cancer*, 01 Dec 2021, 108(11S):11S35-11S45.

B. Research articles

- Le Bescond L., Lerousseau M., **Garberis I.**, André F., Christodoulidis S., Vakalopoulou M., Talbot H. "Unsupervised nuclei segmentation using spatial organization priors". *In book: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 (pp.325-335).*

  Poster presented at MICCAI 2022, Singapore.

- Benkirane H., Vakalopoulou M., Christodoulidis S., **Garberis I.**, Michiels S., Cournède P-H. "Hyper-AdaC: Adaptive clustering-based hypergraph whole slide representation for survival analysis".

  In revision for ML4H (Machine Learning for Health) 2022, Paris.

Bull Cancer 2021; 108: 11S1–11S11
en ligne sur / *on line on*
www.em-consulte.com/revue/bulcan
www.sciencedirect.com

L'intelligence artificielle pourrait-elle intervenir dans l'aide au diagnostic des cancers du sein ? – L'exemple de HER2

Ingrid Garberis[1,2,*], Fabrice Andre[1,2,3], Magali Lacroix-Triki[1,4]

1. Inserm UMR 981, Gustave Roussy Cancer Campus, Villejuif, France
2. Université Paris-Saclay, 94270 Le Kremlin-Bicêtre, France
3. Département d'oncologie médicale, Gustave-Roussy, Villejuif, France
4. Département d'anatomie et cytologie pathologiques, Gustave-Roussy, Villejuif, France

Correspondance :
Ingrid Garberis, 114 rue Édouard-Vaillant, 94800 Villejuif, France.
ingrid-judith.garberis@gustaveroussy.fr

Mots clés
Pathologie mammaire
Intelligence artificielle
Apprentissage profond
Diagnostic
HER2

Résumé

Human Epidermal Growth Factor Receptor 2 (HER2) est un important biomarqueur pronostique et prédictif dans le cancer du sein. Sa détection permet de définir quelles patientes pourront bénéficier d'un traitement ciblé. Si la détermination du statut HER2 par immunohistochimie (IHC) en catégories positive *vs* négative est actuellement bien installée et reproductible, l'apparition d'une nouvelle catégorie dite « HER2-faible » pourrait poser quant à elle des problèmes d'interprétation et de reproductibilité.

Nous avons décrit ici les méthodes utilisées actuellement en routine pour préciser le statut HER2 et l'application de techniques innovantes de *Machine Learning* (ML) pour améliorer ces déterminations, ainsi que les principaux défis et opportunités liés à l'exploitation de la pathologie numérique à l'ère de l'intelligence artificielle (IA).

Keywords
Breast pathology
Artificial intelligence
Deep learning
Diagnosis
HER2

Summary

Could artificial intelligence play a role in breast cancer diagnosis? – The example of HER2 Ingrid

HER2 is an important prognostic and predictive biomarker in breast cancer. Its detection makes it possible to define which patients will benefit from a targeted treatment. While assessment of HER2 status by immunohistochemistry in positive vs negative categories is well implemented and reproducible, the introduction of a new "HER2-low" category could raise some concerns about its scoring and reproducibility.

We herein described the current HER2 testing methods and the application of innovative machine learning techniques to improve these determinations, as well as the main challenges and opportunities related to the implementation of digital pathology in the up-and-coming AI era.

Introduction : la révolution des machines

Nous sommes aujourd'hui confrontés à un futur qui aurait pu paraître très improbable pour la plupart des cerveaux d'une époque passée. Il y a presque 60 ans, le grand auteur de science-fiction Isaac Asimov publia dans le *New York Times* un article intitulé « Visite de l'Exposition universelle de New York de 2014 », inspiré par celle de 1964 [1]. Asimov a d'ores et déjà saisi l'importance qu'auront les ordinateurs, qui commençaient à peine à être utilisés par les particuliers. « L'informatisation continuera à progresser inévitablement » et « l'objet mobile informatisé va pénétrer dans la maison », prédit sans se tromper l'écrivain, mais en même temps il méconnaissait jusqu'à quel point ses paroles seraient d'actualité au XXIe siècle.

L'avenir de la médecine sera probablement dicté par deux piliers principaux : la médecine personnalisée et les progrès techniques. Le premier fait référence aux thérapies ciblées, dirigées contre des cibles spécifiques susceptibles d'être détectées par des biomarqueurs. Le deuxième, quant à lui, englobe différents outils et méthodes capables d'améliorer la démarche diagnostique. Ces deux éléments touchent de près certaines spécialités médicales et notamment l'anatomie pathologique. En effet, les pathologistes devront être capables d'intégrer les dimensions morphologiques, cliniques et moléculaires de la maladie, en maîtrisant de nouvelles techniques et technologies pour délivrer des diagnostics de qualité dans les meilleurs délais [2].

Les pathologistes, ainsi que les radiologues, dont le travail est axé sur le repérage des signes de maladie grâce à l'imagerie médicale, sont dans une position stratégique pour tirer le meilleur parti de la révolution déclenchée par l'avènement de l'IA en médecine. Adoptant la neurobiologie computationnelle, la logique mathématique et l'informatique, l'IA pourrait faciliter l'automatisation du travail, éliminer les tâches fastidieuses, et améliorer la précision et l'efficacité de la pratique quotidienne.

Cet article a pour objectif d'effectuer un état des lieux de la caractérisation du statut HER2 dans des échantillons biologiques et des technologies innovantes conduisant à l'amélioration de cette détermination.

Diagnostic à l'heure actuelle dans le cancer du sein

Diagnostics précis pour une maladie très répandue

Le cancer du sein est la première cause de cancer chez les femmes dans le monde, dont l'incidence a été de 2 261 419 cas en 2020 [3]. Dans la pratique diagnostique de routine, le tissu tumoral mammaire est coloré à l'hématoxyline-éosine (HE), puis examiné au microscope optique pour l'évaluation morphologique qui sera détaillée dans le compte-rendu anatomopathologique. Ce rapport sera complété par les résultats de l'analyse immunohistochimique, pour évaluer l'expression des biomarqueurs à des fins pronostiques et prédictives.

En effet, il est bien connu que, en plus de la taille de la tumeur, de l'envahissement des ganglions lymphatiques et de la présence de métastases, la biologie tumorale est d'une importance vitale pour le pronostic et la prédiction de la réponse au traitement. Ainsi, grâce à la caractérisation histopathologique et moléculaire des cancers, la thérapie peut être considérablement améliorée en l'adaptant à chaque cas individuel. Dans une démarche de médecine personnalisée, le panel de routine pour le cancer du sein inclut la détermination des récepteurs hormonaux aux œstrogènes (RE) et à la progestérone (RP), de HER2 et de Ki67 pour évaluer l'index de prolifération tumorale.

HER2 et recommandations internationales d'évaluation

De tous les oncogènes liés au pronostic dans le cancer du sein, *ERBB2* et sa protéine HER2, qui est surexprimée dans environ 15 % des tumeurs de stade précoce, est le plus étudié [4]. Chez les patientes non traitées, la présence d'une amplification d'*ERBB2* a été associée à un pronostic plus péjoratif [5,6]. En outre, la positivité de HER2 est liée à des lésions de haut grade et à un taux de prolifération cellulaire élevé [7].

L'évolution de ces patientes a radicalement changé avec l'introduction des agents ciblant HER2, tels que le trastuzumab. Pour prédire une possible réponse à ce traitement, HER2 est utilisé en tant que biomarqueur compagnon, en combinant les techniques d'IHC et d'hybridation *in situ* (HIS).

Selon les recommandations du College of American Pathologists et de l'American Society of Clinical Oncology (CAP/ASCO), et les recommandations françaises du Groupe d'étude des facteurs pronostiques immunohistochimiques dans le cancer du sein (GEFPICS) (mise à jour sous presse), une tumeur est considérée comme positive pour HER2 (score 3+) si le nombre de cellules tumorales présentant une forte surexpression de HER2 (c'est-à-dire avec un marquage membranaire complet et intense) dépasse 10 % de la population tumorale infiltrante totale ; équivoque (score 2+) si le nombre de cellules tumorales présentant une surexpression modérée de HER2 (c'est-à-dire avec un marquage membranaire complet et d'intensité modérée, ou incomplet basolatéral d'intensité modérée à forte) dépasse 10 % de la population tumorale infiltrante totale ; et négative dans tous les autres cas (scores 0 et 1+) [8]. Le détail du scoring des différentes catégories est accessible en intégralité dans les dernières recommandations du GEFPICS (mise à jour sous presse) [8]. Les patients sont éligibles ou non à la thérapie ciblée selon leur statut HER2 positif ou négatif respectivement, tandis que les cas équivoques (score 2+) sont renvoyés au test HIS pour affiner le statut [9]. Il existe cinq différents groupes d'après la technique d'HIS pour le cancer du sein, selon le rapport HER2/*Chromosome Enumeration Probe 17* (CEP17, sonde d'énumération des centromères pour le chromosome 17) et le nombre moyen de copies de HER2 par noyau. Les normes d'évaluation sont résumées dans la *figure 1*. De façon récente, suite à l'introduction de nouvelles thérapies ciblées (anticorps conjugués à une chimiothérapie, tels que

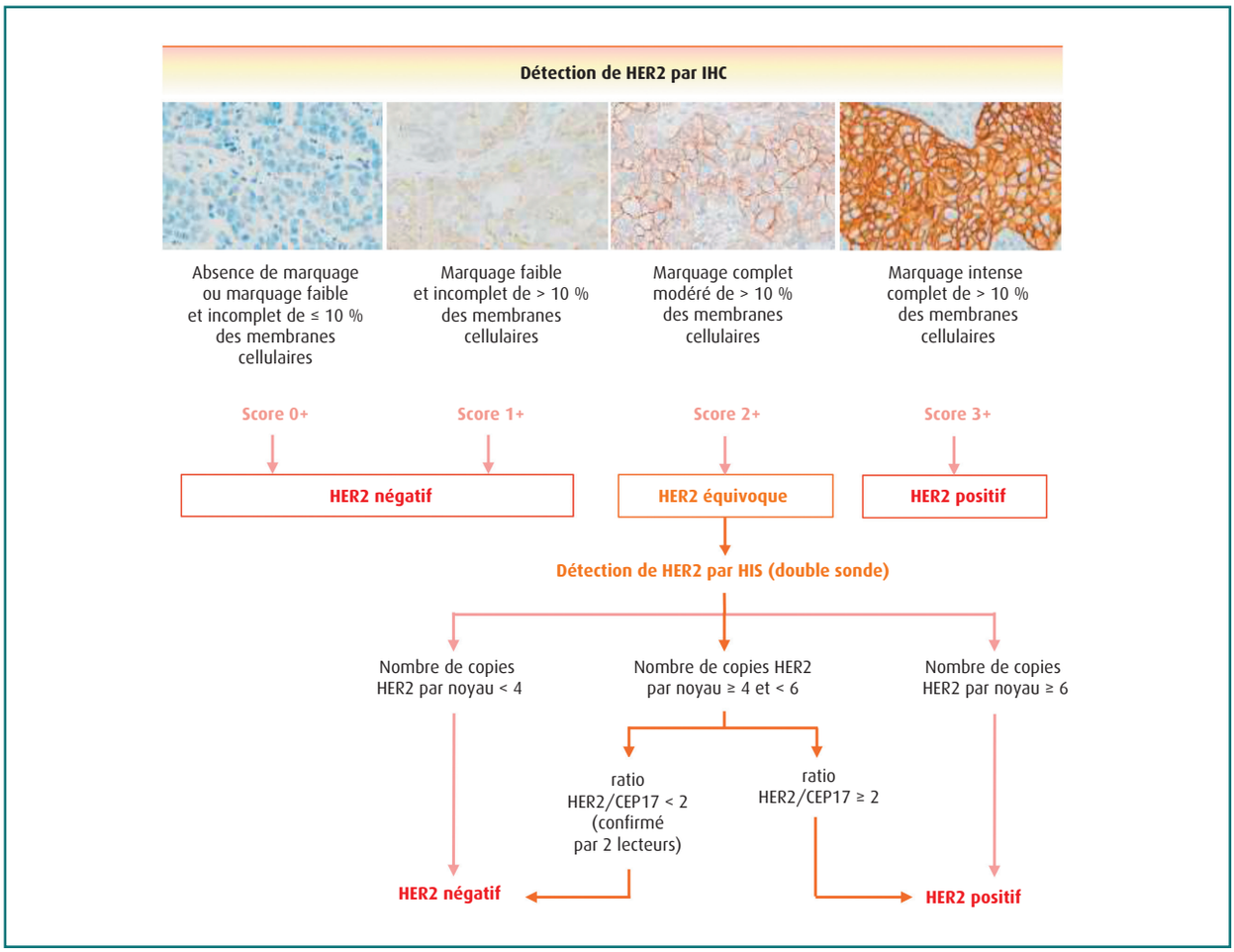


Figure 1
Synthèse des recommandations pour l'évaluation de l'expression de HER2 dans les cancers du sein, par IHC et HIS en utilisant un dosage à double signal (adapté de [9]).

le trastuzumab déruxtécan), une nouvelle catégorie de statut HER2 dite « HER2-faible » est apparue. Cette catégorie inclut les scores IHC 1+ et 2+ non amplifiés en HIS, auparavant classés en négatifs. Les premières données issues des essais cliniques montrent en effet une efficacité de ces molécules (c'est-à-dire trastuzumab déruxtécan) dans les tumeurs « HER2-faibles », mais la reproductibilité interobservateur pour cette catégorie est suboptimale [10,11]. Cette nouvelle catégorie apparaît dans les dernières recommandations nationales (mise à jour sous presse) [8].

Inconvénients de la méthode d'appréciation traditionnelle

Le principal inconvénient des méthodologies actuelles d'évaluation de HER2 est le caractère semi-quantitatif de la détermination, qui est donc sujette à une variabilité interobservateur, particulièrement rapportée dans les scores HER2 1+ et 2+. Par ailleurs, le résultat de la technique de marquage dépend aussi du protocole technique, d'où émanent plusieurs sources de variabilité, et qui conditionne l'intensité du marquage (paramètre crucial des catégories HER2-faibles), comme le temps et la température de la réaction, ou les réactifs employés [12]. Le défi d'atteindre une rigoureuse standardisation de cette nouvelle notion d'HER2-faible est relevé dans la dernière version des recommandations GEFPICS. Un autre facteur à l'origine de la variabilité des résultats, et particulièrement aux discordances de notation entre pathologistes, est représenté par les cas qui arborent une expression hétérogène de HER2 au sein de la tumeur. Différents travaux se sont concentrés sur la résolution de cette difficulté, qui peut potentiellement entraîner des conduites de traitement inadéquates [13,14].
De nouvelles méthodes fiables pour quantifier l'expression protéique pourraient permettre de prédire plus précisément

l'efficacité thérapeutique, notamment vis-à-vis des nouvelles thérapies ciblées. Ce besoin est également fortement soutenu par la demande croissante de la détermination du statut HER2 dans d'autres tumeurs solides qui pourraient tirer un grand bénéfice clinique des médicaments ciblant la voie de signalisation HER2 [15].

La contribution de l'IA à la sénologie

Pathologie numérique : à mi-chemin entre la médecine et l'informatique

La pathologie numérique est un concept apparu d'abord dans le contexte de l'enseignement. Elle réunit toutes les technologies qui permettent d'acquérir des images à partir de coupes tissulaires préparées pour évaluation histopathologique et sa conséquente exploitation *via* des techniques d'analyse assistées par ordinateur (CAD, pour *Computer-Aided Diagnosis*). Le scanner de lames de verre est une partie centrale de ce processus, permettant une acquisition rapide et à haute résolution, de manière automatisée. Souvent, et en raison de son coût élevé, le scanner est aussi le facteur limitant [16]. Donc, la numérisation de lames dans les laboratoires de pathologie n'en est encore qu'à un stade débutant, dépendant de projets d'investissement en cours, mais avec de fortes perspectives de développement dans le court terme.

L'acquisition d'images est basée sur les lames préparées en routine, dont la qualité doit être optimale pour obtenir une image virtuelle de valeur. Pour un échantillon d'une taille habituelle (15 × 15 mm) et au grossissement couramment utilisé (× 20), la vitesse d'acquisition oscillera entre 30 secondes et 5 minutes, dépendant des spécifications du scanner. Ce processus produit des fichiers de très grande taille, ce qui entraîne des contraintes au niveau de leur visualisation et surtout du stockage. Pour minimiser ces obstacles, la numérisation se fait par des petits carrés ou *patches*, qui seront compressés puis assemblés dans un type spécial de fichier nommé « pyramidal » qui simule le fonctionnement du microscope optique, en permettant d'afficher toutes les résolutions possibles d'une image [17]. Ces fichiers s'ouvrent dans des logiciels de visualisation, normalement fournis par le fabricant du scanner, ce qui représente un autre obstacle à franchir : malgré l'existence de *viewers* universels multiformats, l'absence de format unique pour les images générées par des outils provenant de différents constructeurs est préjudiciable pour la généralisation des solutions de pathologie numérique. Les lames virtuelles (WSI, pour *Whole Slide Images*) ainsi obtenues peuvent convenir à différentes finalités. Comme décrit par Al-Janabi et al., les applications principales peuvent se regrouper en diagnostic, recherche, éducation et archivage, avec des avantages appréciés à plusieurs niveaux [18]. Du point de vue du patient, des délais de diagnostic plus rapides, un meilleur accès aux avis d'experts et des rapports anatomopathologiques plus solides en raison de la précision apportée par les nouvelles technologies ne sont que quelques atouts de la pathologie numérique.

En raison du nombre croissant de biomarqueurs évalués quantitativement pour la prise de décision thérapeutique, le CAD des biomarqueurs tissulaires pourrait devenir un aspect crucial de la médecine personnalisée, en assurant le choix de l'option de traitement adéquate pour chaque cas.

Un renforcement réciproque fait avancer l'installation des technologies modernes dans la pratique quotidienne. Tout d'abord, l'utilisation de scanners dans les services de pathologie, qui augmente progressivement, renforcera probablement l'adoption d'une analyse automatisée pour certaines activités de diagnostic. En même temps, l'implémentation de la microscopie numérique dans la routine diagnostique est poussée par l'émergence de l'IA en médecine, qui se nourrit de données massives ou *big data* pour tester leurs algorithmes et méthodes d'analyse automatisée.

L'établissement d'un nouveau flux de travail qui inclut la numérisation et l'annotation de lames est ainsi encouragé (*figure 2*). Avant la mise en œuvre d'un système de pathologie numérique, il est fondamental d'appréhender l'importance et l'impact potentiel des discordances entre la lame de verre et les images digitalisées. Tous les centres devraient disposer de leur propre procédure de validation, nécessaire pour garantir que ce système peut fonctionner de manière adéquate dans la pratique clinique de routine, c'est-à-dire que le pathologiste parviendra au même diagnostic quelle que soit la méthodologie adoptée [19]. Dans le cadre d'une méta-analyse menée par Williams et al., où 8 069 comparaisons ont été évaluées, les discordances mentionnées atteignaient 4 %, dont la moitié correspondaient à des cas de diagnostic difficile et à ceux connus pour être liés à une variabilité interobservateur [20].

Mais… qu'est-ce que l'IA ?

L'IA est un concept au sens large décrivant des systèmes automatisés qui peuvent effectuer des tâches considérées comme nécessitant de l'« intelligence », en imitant ce qu'un humain pourrait faire dans la même situation. Ces systèmes reposent sur la création et l'application d'algorithmes [21].

La différence entre l'IA, l'apprentissage automatique ou *Machine Learning* (ML) et l'apprentissage profond ou *Deep Learning* (DL) n'est pas toujours évidente pour les non-experts. Le ML, une sous-catégorie de l'IA, fait référence au processus par lequel un système est capable d'apprendre et d'analyser à partir de données qui lui sont fournies. Le DL, sous-type de ML, est un modèle informatique inspiré par la biologie, où la profondeur est donnée par les niveaux d'abstraction ou *couches* qui l'intègrent, constituées par des unités de calcul nommées *neurones* [22]. Les réseaux de neurones convolutifs (CNN pour *Convolutional Neural Networks*), qui sont une forme de DL, ont été utilisés avec succès dans une diversité de tâches de reconnaissance visuelle d'objets [23,24].

L'algorithme, entendu comme la suite d'opérations mathématiques qui permettront d'arriver à la résolution d'un problème, est le cœur du système informatique. Extrapolé au domaine de la pathologie et en fonction du rapport que ce système maintient avec l'humain, on peut décrire deux aspects dans l'IA, avec

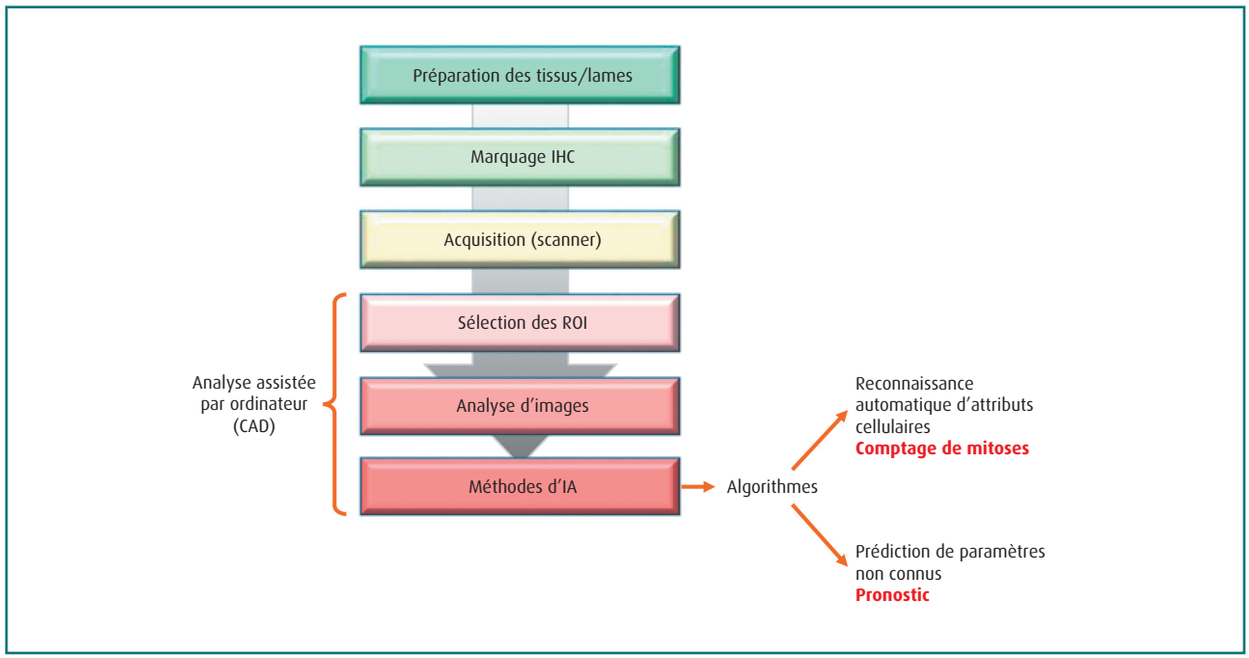


FIGURE 2
Flux de travail pour un laboratoire de pathologie qui incorpore l'analyse assistée par ordinateur (CAD). IHC : immunohistochimie, ROI : régions d'intérêt, IA : intelligence artificielle

des niveaux de complexité croissants. Dans l'approche de ML, l'algorithme essayera de reproduire la démarche du pathologiste : c'est l'humain qui apprend à la machine à résoudre un problème, avec des critères propres de l'anatomie pathologique, comme la prise en compte de la taille et de la forme des cellules. Dans l'approche de DL, la machine n'a pas besoin d'intervention humaine pour trouver une solution à la question demandée : elle va se débrouiller pour créer les algorithmes elle-même, basée sur des critères « invisibles » à l'œil humain, qui ne proviennent pas de la biologie mais des éléments constitutifs de l'image, comme la couleur, le contraste, etc.

Une deuxième considération sur le rôle du pathologiste dans le système porte sur le type d'apprentissage. Dans l'*apprentissage supervisé*, la cohorte d'entraînement présentée au réseau a été préalablement étiquetée : le programme informatique reconnaît les *labels* ou annotations introduites par le pathologiste pour apprendre à distinguer différentes caractéristiques qu'il devra repérer dans les futures cohortes. L'ordinateur pourra ainsi mettre en pratique le système de classification appris pour grader ou scorer de nouvelles images. Dans l'*apprentissage non supervisé*, les données sont présentées au programme sans aucune annotation et, à force d'analyser une grande quantité d'entrées, il devient capable de trouver des associations et des motifs à partir des caractéristiques présentes dans les images. Cela permettra de générer des classifications indépendamment de l'entraîneur. Plus la quantité de données est importante, plus performant deviendra l'algorithme.

Comment l'IA pourrait aider au diagnostic ?

L'utilité de l'IA dans la pathologie diagnostique se traduit par une transformation de l'interprétation subjective des résultats des biomarqueurs en résultats quantitatifs plus précis et reproductibles. À titre d'exemple, on peut mentionner les performances obtenues par Bai et al. en utilisant des modèles mathématiques (*Algorithms for Quantum Applications* [AQUA]) pour mener à bout une détermination quantitative de HER2, par rapport à la méthode standard semi-quantitative [25]. À terme, la pathologie numérique peut fonctionner non seulement comme un moyen pour fournir de meilleurs services, mais aussi en tant qu'un véhicule pour soulager le pathologiste de certaines tâches qui consomment beaucoup de temps, laissant la place à un rôle plutôt centré sur la supervision des résultats obtenus grâce à l'analyse automatisée d'images.

Cependant, donner du sens biologique à une matrice de pixels n'est pas une tâche aisée : le pathologiste doit acquérir des compétences dans le traitement et la segmentation d'images, ainsi que des notions d'informatique. Toutefois, il reste le porteur d'une expertise qui couvre notamment les facteurs les plus importants pour l'étude, tels que l'influence des conditions préanalytiques sur les coupes tissulaires colorées, la qualité des WSI et la performance des algorithmes ; de même que les connaissances physiopathologiques pour interpréter les données générées dans le contexte de la question de recherche et les limites de la conception de l'étude [26].

L'application de l'IA trouve un champ fertile dans la pathologie oncologique sénologique, du fait de la valeur prédictive et

pronostique de l'évaluation histopathologique et moléculaire des tumeurs et des déterminations qui font partie du compte-rendu anatomopathologique [27]. Ces informations obtenues à partir des lames HE et des lames d'IHC pour les biomarqueurs histopronostiques (RE, RP, HER2, Ki67) sont encadrées dans la démarche de la médecine personnalisée, qui cherche à orienter le traitement en fonction des caractéristiques moléculaires de chaque tumeur.

Dans les années à venir, l'IA pourrait même se positionner comme un rival sérieux de l'IHC, au point de remplacer entièrement cette technique, en prédisant les résultats ou le statut des biomarqueurs directement à partir des lames HE. Certains classificateurs basés sur des modèles du DL seraient en mesure de fournir, à partir de la lame HE, des informations sur le sous-type moléculaire intrinsèque des tumeurs, ce qui pourrait concurrencer même d'autres méthodes comme les tests moléculaires [28,29]. Il est probablement plus juste de penser l'IA comme une technique complémentaire de plus à la disposition du pathologiste (au même titre que l'IHC ou la *Fluorescent In Situ Hybridization* [FISH] par exemple) capable de faire dans certaines situations ce que d'autres techniques complémentaires ne font pas (sur le plan diagnostique, pronostique ou théranostique), et capable pour certaines autres situations de remplacer potentiellement (ou restreindre l'utilisation de) ces techniques.

Applications en pathologie mammaire

Bataillon et al. ont décrit deux approches qui pourraient assister le pathologiste à exécuter ses activités quotidiennes. D'un côté, des algorithmes entraînés pour réaliser des tâches de triage, comme la sélection de biopsies prioritaires, la demande de techniques complémentaires ou l'identification de métastases dans les ganglions lymphatiques, tâches qui s'avèrent souvent longues et fastidieuses [30]. Dans ce cadre, qui regroupe les activités tendant à mimer le *modus operandi* du pathologiste, on peut aussi inclure la quantification (de mitoses, de cellules) qui peut être aussi exécutée par un programme informatique, avec des résultats comparables à ceux d'un expert [31-33]. Cet « assistant virtuel », placé au début du flux de travail, pourrait être d'une grande aide pour organiser les activités et mieux répondre aux demandes des cliniciens et patients, tout en déchargeant le pathologiste qui aurait plus de temps à consacrer à des activités d'expertise diagnostique, de recherche et d'enseignement, fréquemment reléguées au second plan derrière le repérage des éléments sur les lames histologiques.

D'un autre côté, on retrouve des algorithmes qui s'éloignent des paramètres connus pour tenter de trouver, de manière autonome, des solutions à des questions complexes. Ces réponses, comme la prédiction de la réaction à un traitement donné, sont basées sur des détails inaperçus par l'œil humain, en raison de leur appartenance aux propriétés intrinsèques de l'image, ou parce que la quantité de données à analyser pour arriver à discerner une trame est trop volumineuse [34]. Malgré le problème d'interprétabilité que cette méthodologie comporte, son étendue produit des résultats remarquables. Dans une étude conduite par Islam et al., un modèle de DL intégratif a été utilisé sur des ensembles de données omiques de cancer du sein pour les classer selon leurs sous-types moléculaires [35]. Un examen plus approfondi du sous-groupe HER2-enrichi a montré qu'en fait il était composé de trois sous-types de cancers avec des valeurs de survie spécifique significativement différentes (*figure 3*).

Stratégies informatiques au service de l'évaluation du statut HER2

Le CAD est un outil très alléchant pour répondre plus facilement aux nouvelles exigences de la médecine personnalisée, comportant les facultés de réduire le temps d'élaboration du diagnostic et d'augmenter la reproductibilité du scoring des biomarqueurs. À l'égard des techniques de DL, les avancées dans la reconnaissance d'images et la détection entièrement automatisée des attributs ont fait leur entrée dans le laboratoire de pathologie pour contribuer amplement à la prise de décision clinique.

Considérant l'évaluation de HER2 dans le cancer du sein, deux approches peuvent être citées : la prédiction du statut HER2 à partir de lames colorées par HE et la prédiction du statut HER2 à partir de lames d'IHC. Deux stratégies seront ensuite brièvement abordées : la prédiction du statut HER2 à partir de lames d'HIS et l'essor des challenges publics pour faire avancer la recherche en IA.

Des applications d'IA pour l'évaluation de HER2 peuvent être consultées dans le *tableau I*.

Algorithmes pour prédire le statut HER2 à partir d'une lame HE

D'après un nombre croissant d'explorations méthodologiques, la capacité du DL pour réaliser des tâches complexes de reconnaissance d'éléments et de motifs pourrait conduire à une nouvelle génération d'outils de CAD.

Il est connu que certaines anomalies moléculaires s'accompagnent de modifications morphologiques identifiables dans des coupes histologiques colorées par HE [36-38]. En général, ces caractéristiques sont trop fines pour être perçues de manière fiable par une inspection manuelle. Rawat et al. appellent ces traits présents sur les lames HE « empreintes tissulaires » et ils s'en servent pour enseigner à la machine les différences biologiques entre les patients. L'algorithme apprend de manière non supervisée (sans annotations) comment déduire le statut des biomarqueurs à partir des WSI [39].

Dans une étude développée par Shamai et al., une méthode de DL a été appliquée à des images de *Tissue Microarrays* (TMA) de cancer du sein, colorées par HE, pour prédire l'expression de 19 biomarqueurs dont RE, RP et HER2. Les résultats ont renforcé l'idée que l'expression des marqueurs moléculaires se reflète phénotypiquement dans la morphologie des tissus, sous forme de trames subtiles, et que ces dernières peuvent être identifiées par des modèles mathématiques. Ainsi, l'IA pourrait être utilisée pour prédire l'expression de biomarqueurs directement à partir des images colorées par HE [40]. Comme le modèle développé

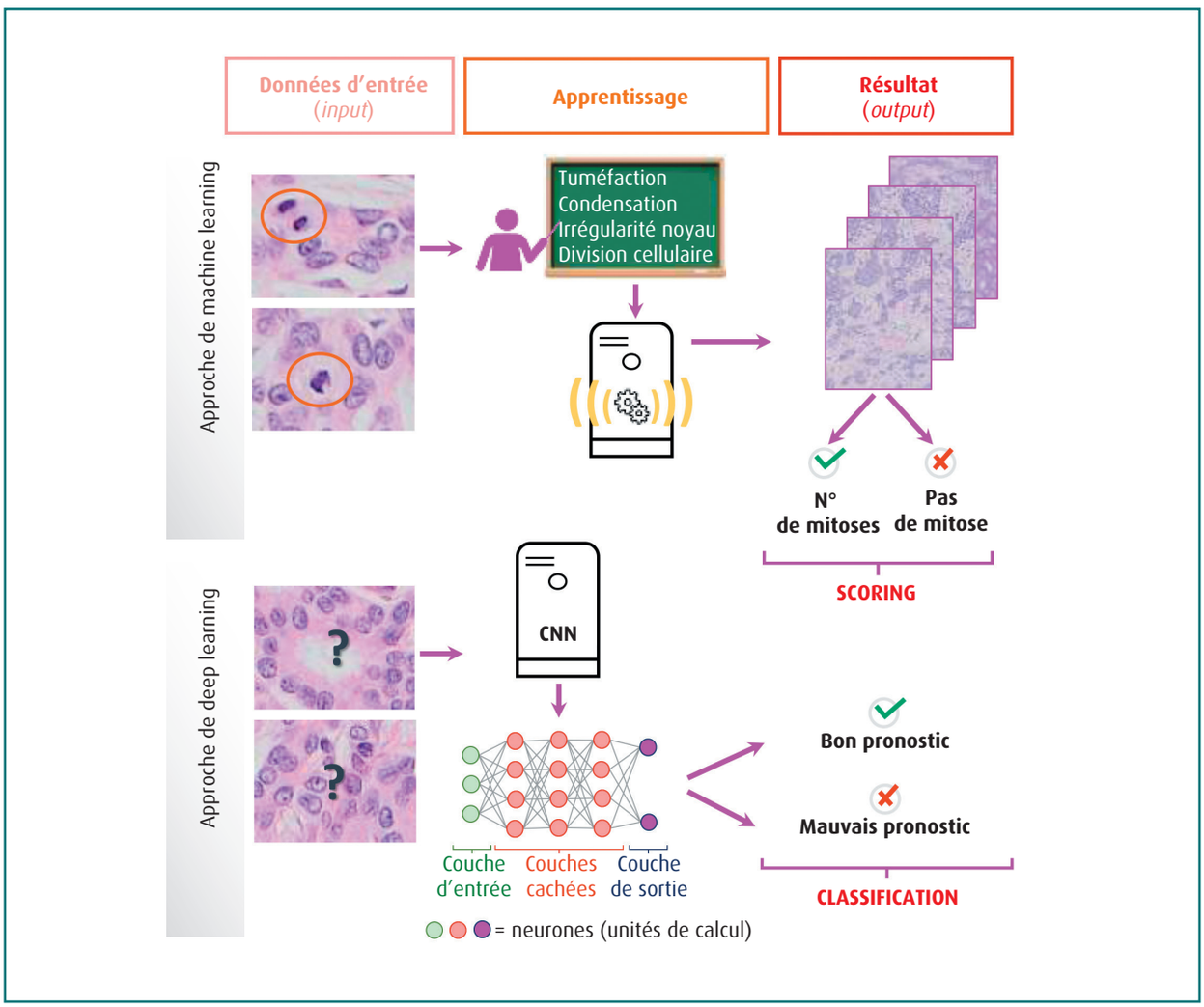


Figure 3

Différentes méthodologies d'IA applicables à la pathologie numérique. Dans l'approche de ML, l'algorithme apprend à reconnaître des traits, qu'il utilise ensuite pour scorer de nouvelles images. Dans l'approche de DL, les caractéristiques des images sont retrouvées directement par l'algorithme, sans intervention humaine, pour répondre à la tâche assignée. CNN : réseau de neurones convolutifs

n'a été validé que pour le RE, et que la taille des échantillons est limitée puisqu'il s'agit des TMA, d'autres études doivent être menées sur les déterminations de HER2, mais les résultats sont déjà prometteurs.

Des résultats semblables ont été obtenus par un modèle de DL développé par Bychkov et al., entraîné avec des données d'HIS et de TMA colorés par HE, capable de prédire le statut HER2 à partir de l'architecture tissulaire [41]. Ce travail a ensuite exploré la corrélation des prédictions à l'efficacité du traitement adjuvant par trastuzumab. Il a été démontré que ce score peut identifier des patients pouvant obtenir un bénéfice plus important avec ce traitement, mais aussi une autre catégorie de patients. Ce groupe inclut les cas avec un résultat HIS négatif (donc non éligibles à une thérapie ciblée anti-HER2), et pourtant possédant une morphologie compatible avec une tumeur positive pour HER2 ; ces patients ont une survie moins favorable. Il reste à déterminer la place que ce type d'information provenant d'un algorithme pourrait occuper à côté des explorations moléculaires existantes. Dans la même ligne de Rawat et al., des WSI ont été utilisées par l'équipe d'Anand pour développer un classificateur qui permet de discerner entre cas positifs et négatifs pour HER2, pouvant s'ériger en une potentielle méthode de dépistage avant le test par IHC [42]. Cette méthode est composée de trois réseaux de neurones séquentiels, entraînés avec des WSI de deux coupes en série de tumeur par patient (IHC pour HER2 et HE). Ce modèle a non seulement donné de bonnes performances sur la cohorte

Tableau I
Comparaison des approches d'IA pour l'évaluation de HER2

Méthode	Année	Dataset		Classificateur	Remarques
		Lames	Coloration		
Masmoudi et al. [56]	2009	WSI	IHC	*Multistage clustering algorithm*	Précision : 90 %
Ficarra et al. [57]	2011	WSI	IHC	Diagrammes de Voronoï, triangulation de Delaunay	CC : 0,98 ; erreur moyenne : 0,14
Brügmann et al. [58]	2012	WSI	IHC	Logiciel HER2-CONNECT™	Sensibilité : 99,2 % ; specificité : 100 %
Tuominen et al. [43]	2012	WSI	IHC	Logiciel ImmunoMembrane™	Valeur Kappa : 0,80
Pitkäaho et al. [59]	2016	WSI	IHC	CNN, architecture AlexNet	Précision : 97,7 %
Vandenberghe et al. [45]	2017	WSI	IHC	LSVM, RF, CNN	Précision : 83 %
Saha et al. [46]	2018	WSI	IHC	*Her2Net-LSTM recurrent network*	Précision : 98 %
Qaiser et al. [49]	2018	WSI	IHC	ResNet, *RNN* pour *deep reinforcement learning*	Précision : 79 %
Khameneh et al. [60]	2019	WSI	IHC	Architecture UNet pour classification tissulaire	Précision : 87 %
Rawat et al. [39]	2020	TMA et WSI	H & E	Architecture *Resnet34, fingerprint-based classifier*	AUC : 0,71 et 0,79 (selon la cohorte)
Anand et al. [42]	2020	WSI	H & E	*CNN*	AUC : 0,82
Tewary et al. [44]	2021	WSI	IHC	*VGG19 architecture (very deep CNN), fully connected dense layers*	Précision : 93 %
Bychkov et al. [41]	2021	TMA et WSI	H & E	*CNN, transfer learning*	Précision : 70 % (sur TMA); 67 % (sur WSI)

AUC : *Area Under Curve* ; CC : coefficient de correlation ; CNN : *Convolutional Neural Network*, réseau de neurones convolutifs ; H & E : hématoxyline-éosine ; IHC : immunohistochimie ; LSTM : *Long Short-Term Memory* ; LSVM : *Linear Support Vector Machine* ; RF : *Random Forest* ; RNN : *Recurrent Neural Network* ; TMA : *Tissue Microarray* ; WSI : *Whole Slide Image*.

d'entraînement, mais a également fonctionné pour un ensemble de données indépendantes et multicentriques, ce qui représente un vrai challenge pour les études de ce genre sur des données médicales.

Algorithmes pour prédire le statut HER2 à partir d'une lame d'IHC (IA pour aider à l'interprétation du scoring)

Dans le cas des lames d'IHC, certains facteurs peuvent augmenter la complexité de la tâche en raison par exemple d'une mauvaise contre-coloration, de noyaux qui se chevauchent ou d'un marquage non spécifique (bruit de fond).

En règle générale, l'évaluation automatisée du marquage HER2 commence par l'extraction de la membrane cellulaire, suivie de la quantification de sa continuité pour construire le score. Cette procédure implique d'abord une segmentation basée sur le signal colorimétrique (délimitation des éléments morphologiques tels que les noyaux, les épithéliums, etc., afin qu'ils puissent être analysés), puis une classification par des méthodes de ML. L'application employée actuellement, ImmunoMembrane [43], qui est disponible en tant que plugin pour ImageJ, et de nouvelles approches comme AutoIHC-Analyzer [44] sont des exemples de cette méthodologie.

Dans l'étude conduite par Vandenberghe et al., l'analyse d'échantillons de tissus mammaires colorés par l'IHC a permis de confronter, en premier lieu, des modèles de ML avec une approche de DL (ConvNet). Les performances du DL ont surpassé celles des modèles de ML. Ensuite, ces résultats ont été comparés à ceux obtenus par un pathologiste. Dans la cohorte étudiée, l'algorithme a fourni des scores HER2 au moins aussi précis que le pathologiste. Ces conclusions suggèrent qu'une fois l'apprentissage accompli, l'approche proposée peut générer des scores HER2 valides de manière entièrement automatisée. En outre, la méthode de DL peut jouer un rôle déterminant dans la mise en relief de cas difficiles qui présentent un risque d'erreur de diagnostic, ce qui s'avère particulièrement utile pour la quantification objective et précise de biomarqueurs dans les cas montrant une grande hétérogénéité. Comme dans un grand nombre d'études, une limite de ce travail est le manque de validation sur des échantillons provenant de divers centres. Cette validation supplémentaire est requise pour généraliser cette approche [45].

Saha et al. ont proposé une méthode qui utilise un réseau plus profond que ConvNet pour déterminer le statut HER2, qu'ils ont nommée *Her2Net*. Her2Net a montré une bonne performance lors de la comparaison avec le scoring fait par des pathologistes, atteignant une précision de 98,33 %. Un des atouts de cette méthode est l'option de l'intégrer dans d'autres structures de calcul pour la segmentation, la classification et le scoring [46].

Algorithmes pour prédire le statut HER2 à partir d'une lame d'HIS

Avec la possibilité de numériser les lames d'HIS, une porte s'ouvre pour l'exploitation de cette technique dans le champ de l'IA. C. Franchet et C. Laurent ont ainsi décrit une méthode d'analyse automatisée pour assister au diagnostic du statut *HER2* sur des lames d'HIS fluorescentes [47]. Dans cette approche, les signaux d'hybridation sont capturés dans la surface de la lame et aussi dans son épaisseur pour obtenir une image assemblée qui contient les pixels les plus intenses. L'image passera ensuite par une étape de segmentation nucléaire, puis par l'annotation des noyaux et signaux d'hybridation dans le but d'être quantifiée. Différentes modalités pourront être utilisées en fonction de l'anomalie recherchée, de l'identification des fusions de gènes à l'énumération des chromosomes, comme c'est le cas pour HER2.

Stratégies pour faciliter le développement des algorithmes d'IA : les challenges publics

Le formidable potentiel des méthodes automatisées pour assister le pathologiste avec des scores d'analyse objectifs pour l'IHC a été révélé dans le contexte des challenges. Il s'agit de « concours » qui mettent à disposition des chercheurs une plateforme incluant un ensemble de données, pour évaluer les performances des algorithmes dessinés en vue de répondre à une question scientifique particulière [48].

Pour le « HER2 Challenge Contest 2016 », où le sujet était la détection automatisée du statut HER2 sur des lames colorées par IHC, 18 soumissions issues de 14 équipes ont été considérées, et 8 des 10 méthodes mieux classées étaient basées sur du DL (CNN). Les résultats étaient cohérents avec les travaux précédents sur le sujet, appuyant l'idée qu'une méthode de scoring automatisée ou semi-automatisée a un fort potentiel de déploiement dans la pratique quotidienne [49]. Le problème de l'hétérogénéité du marquage HER2 a été abordé par une architecture de DL créée pour prédire les régions d'intérêt (ROI) à regarder dans l'échantillon observé [50]. Le récent « HEROHE Grand Challenge 2020 » a relevé une question analogue mais sur des images HE de cancer du sein invasif où le défi était d'obtenir le statut HER2 sans l'image IHC correspondante. Encore une fois, l'idée était d'exploiter les caractéristiques morphologiques en tant que substituts du statut HER2. Une cascade de classificateurs de DL et de l'apprentissage multi-instance (MIL) ont été appliqués à l'ensemble de données, montrant de bons scores d'efficacité pour différentes métriques d'évaluation [51].

Conclusion

L'IA (notamment les modèles de DL) représente une option très attrayante pour augmenter l'efficacité des évaluations réalisées dans le laboratoire de pathologie. En effet, les méthodes d'usage courant ne sont pas exemptées d'erreur, en plus d'être fréquemment coûteuses et chronophages. Les nouveaux outils informatiques pourraient améliorer le flux de travail par la réduction des coûts, par exemple en évitant l'utilisation de l'IHC pour le dépistage, ou en ayant la possibilité d'analyser plusieurs biomarqueurs ou d'évaluer l'hétérogénéité intratumorale avec seulement une lame HE [52].

Pourtant, il faut garder un œil sur l'importance d'un rigoureux contrôle humain tout au long du processus et surtout dans l'étape de validation. D'une part, une attention spéciale doit être accordée à la normalisation des données en ce qui concerne la phase préanalytique (fixation, coloration HE, protocole d'IHC, utilisation de témoins, participation à une évaluation externe de la qualité, suivi des recommandations en vigueur), du type de scanner et des protocoles de numérisation dans les cohortes provenant de divers centres. D'autre part, il ne faut pas oublier que les images sont des représentations imparfaites de la réalité, et qu'elles doivent être analysées de manière intégrée en fonction du contexte anatomoclinique, tâche qui peut s'avérer plus complexe pour une machine, et à la lumière des particularités de l'expression de HER2 en IHC. En d'autres mots, il ne faut pas prendre chaque pixel littéralement.

Afin de faire des approches de DL une pratique de routine, il est essentiel de compter avec des cohortes de qualité, exhaustives et d'accès simplifié, comportant des images annotées, assemblées à des données cliniques, morphologiques, radiologiques et moléculaires structurées. Cela permettra de fluidifier le test de nouvelles méthodologies en prospectif et rétrospectif, et de pouvoir ainsi les mettre en pratique plus rapidement.

Puisqu'une forte critique faite à ces approches est l'ignorance des opérations qui ont lieu dans l'ordinateur pour élaborer une réponse (la fameuse « boîte noire »), l'épreuve ultime à surmonter sera de doter les méthodes de l'interprétabilité requise pour qu'elles soient acceptées par la communauté médicale [53]. Des efforts visant à limiter cet effet incluent des solutions comme les *attention maps*, qui permettent de sélectionner des sous-ensembles d'images dans les coupes histologiques, contenant des ROI qui seront ensuite traitées d'une façon analogue à la méthode du pathologiste [54], ou des systèmes capables d'interpréter chaque ROI en décrivant les caractéristiques microscopiques et d'expliquer ce que le réseau voit lors de la description des observations sur les lames [55].

Finalement, outre le manque d'interprétabilité de l'IA, certains pathologistes craignent le changement de *workflow*. En effet, en plus de devoir s'appuyer sur les résultats des algorithmes, souvent inexplicables, et de s'éloigner de la microscopie optique telle qu'on la connaît, qui a servi fidèlement la spécialité depuis plus de 100 ans, les pathologistes doivent apprendre non seulement à utiliser de nouvelles plateformes, mais aussi à les adapter à

la pratique quotidienne pour standardiser les comptes-rendus anatomopathologiques. Si la résistance au changement est effectivement présente parmi les pathologistes comme dans d'autres spécialités, l'anatomie-pathologie est une discipline dynamique qui implémente régulièrement et avec succès de grandes avancées scientifiques à ses pratiques usuelles. Il est donc impératif de former pleinement les futures générations de pathologistes aux nouvelles technologies qui font partie de la pathologie numérique et de l'IA en santé, sans oublier néanmoins que le pathologiste reste le garant du diagnostic, responsable de la validation et la mise en pratique de ces avancées. Le rôle du pathologiste se déroulera à l'interface entre le diagnostic et la recherche, et sera primordial pour vérifier la qualité du diagnostic et des analyses réalisées sur les coupes tissulaires.

Il semble bien que, tel que prédit par Asimov, des appareils continueront à soulager l'humanité des travaux fastidieux, non seulement pour accomplir de lourdes activités comme la cueillette de fruits par des robots, mais aussi pour alléger la charge que représentent certaines tâches de laboratoire, telles que la quantification des éléments en microscopie ! Et, paraphrasant son article de 1964 paru dans le *New York Times*, « si les machines sont si intelligentes aujourd'hui, qui sait ce qu'elles feront dans 50 ans… ? ».

Déclaration de liens d'intérêts :
Les auteurs déclarent ne pas avoir de liens d'intérêts.

Cet article fait partie d'un numéro supplément *Cancer du sein et HER2, doit-on se limiter à la surexpression ?* réalisé avec le soutien institutionnel du laboratoire Daiichi-Sankyo/AstraZeneca.

Références

[1] Asimov I. Visit to the World's Fair of 2014 1964. https://archive.nytimes.com/www.nytimes.com/books/97/03/23/lifetimes/asi-v-fair.html?wptouch_preview_theme=enabled (accessed June 9, 2021).
[2] Salto-Tellez M, James JA, Hamilton PW. Molecular pathology – The value of an integrative approach. Mol Oncol 2014;8:1163-8.
[3] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209-49.
[4] Yarden Y. Biology of HER2 and Its Importance in Breast Cancer. Oncology 2001;61:1-13.
[5] Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987;235:177-82.
[6] Ross JS, Slodkowska EA, Symmans WF, Pusztai L, Ravdin PM, Hortobagyi GN. The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine. Oncologist 2009;14:320-68.
[7] Sørlie T. Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. Eur J Cancer Oxf Engl 1990/2004;40:2667-75.
[8] Penault-Llorca F, Vincent-Salomon A, MacGrogan G, Roger P, Treilleux I, Valent A, et al. Mise à jour 2014 des recommandations du GEFPICS pour l'évaluation du statut HER2 dans les cancers du sein en France. Ann Pathol 2014;34:352-65.
[9] Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. J Clin Oncol 2018;36:2105-22.
[10] Modi S, Park H, Murthy RK, Iwata H, Tamura K, Tsurutani J, et al. Antitumor Activity and Safety of Trastuzumab Deruxtecan in Patients With HER2-Low-Expressing Advanced Breast Cancer: Results From a Phase Ib Study. J Clin Oncol 2020;38:1887-96.
[11] Schettini F, Chic N, Brasó-Maristany F, Paré L, Pascual T, Conte B, et al. Clinical, pathological, and PAM50 gene expression features of HER2-low breast cancer. NPJ Breast Cancer 2021;7:1.
[12] Gown AM. Current issues in ER and HER2 testing by IHC in breast cancer. Mod Pathol 2008;21:S8-15.
[13] Potts SJ, Krueger JS, Landis ND, Eberhard DA, Young GD, Schmechel SC, et al. Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. Lab Invest 2012;92:1342-57.
[14] Buckley NE, Forde C, McArt DG, Boyle DP, Mullan PB, James JA, et al. Quantification of HER2 heterogeneity in breast cancer-implications for identification of sub-dominant clones for personalised treatment. Sci Rep 2016;6:23383.
[15] Oh D-Y, Bang Y-J. HER2-targeted therapies – a role beyond breast cancer. Nat Rev Clin Oncol 2020;17:33-48.
[16] Jara-Lazaro AR, Thamboo TP, Teh M, Tan PH. Digital pathology: exploring its applications in diagnostic surgical pathology practice. Pathology 2010;42:512-8.
[17] Ameisen D, Naour GL, Daniel C. Technologie des lames virtuelles – de la numérisation à la mise en ligne. Med Sci 2012;28:977-82.
[18] Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. Histopathology 2012;61:1-9.
[19] Hanna MG, Pantanowitz L, Evans AJ. Overview of contemporary guidelines in digital pathology: what is available in 2015 and what still needs to be addressed? J Clin Pathol 2015;68:499-505.
[20] Williams BJ, DaCosta P, Goacher E, Treanor D. A Systematic Analysis of Discordant Diagnoses in Digital Pathology Compared With Light Microscopy. Arch Pathol Lab Med 2017;141:1712-8.
[21] Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. JAMA 2019;322:1806-16.
[22] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
[23] LeCun Y, Bottou L, Bengio Y, Ha P. Gradient-Based Learning Applied to Document Recognition: Proc. of the IEEE 1998:46.
[24] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84-90.
[25] Bai Y, Cheng H, Bordeaux J, Neumeister V, Kumar S, Rimm DL, et al. Comparison of HER2 and Phospho-HER2 Expression between Biopsy and Resected Breast Cancer Specimens Using a Quantitative Assessment Method. PLoS One 2013;8:e79901.
[26] Aeffner F, Adissu HA, Boyle MC, Cardiff RD, Hagendorn E, Hoenerhoff MJ, et al. Digital Microscopy, Image Analysis, and Virtual Slide Repository. ILAR J 2018;59:66-79.
[27] Tizhoosh HR, Pantanowitz L. Artificial Intelligence and Digital Pathology: Challenges and Opportunities. J Pathol Inform 2018;9.
[28] Jaber MI, Song B, Taylor C, Vaske CJ, Benz SC, Rabizadeh S, et al. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. Breast Cancer Res 2020;22:12.
[29] Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, et al. Image analysis with deep learning to predict breast cancer grade, ER status,

histologic subtype, and intrinsic subtype. NPJ Breast Cancer 2018;4:1-8.
[30] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017;318:2199-210.
[31] Pantanowitz L, Hartman D, Qi Y, Cho EY, Suh B, Paeng K, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. Diagn Pathol 2020;15:80.
[32] Koopman T, Buikema HJ, Hollema H, de Bock GH, van der Vegt B. Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement. Breast Cancer Res Treat 2018;169:33-42.
[33] Del Rosario Taco Sanchez M, Soler-Monsó T, Petit A, Azcarate J, Lasheras A, Artal C, et al. Digital quantification of KI-67 in breast cancer. Virchows Arch 2019;474:169-76.
[34] Bataillon G, Vincent-Salomon A, Jouvin N, Walter T. La pathologie à l'heure de l'intelligence artificielle : exemple... Corresp en onco-théranostic 2019:52.
[35] Mohaiminul Islam Md, Huang S, Ajwad R, Chi C, Wang Y, Hu P. An integrative deep learning framework for classifying molecular subtypes of breast cancer. Comput Struct Biotechnol J 2020;18:2185-99.
[36] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-67.
[37] Finberg KE, Sequist LV, Joshi VA, Muzikansky A, Miller JM, Han M, et al. Mucinous differentiation correlates with absence of EGFR mutation and presence of KRAS mutation in lung adenocarcinomas with bronchioloalveolar features. J Mol Diagn 2007;9:320-6.
[38] Rakha EA, Alsaleem M, ElSharawy KA, Toss MS, Raafat S, Mihai R, et al. Visual histological assessment of morphological features reflects the underlying molecular profile in invasive breast cancer: a morphomolecular study. Histopathology 2020;77:631-45.
[39] Rawat RR, Ortega I, Roy P, Sha F, Shibata D, Ruderman D, et al. Deep learned tissue « fingerprints » classify breast cancers by ER/PR/Her2 status from H&E images. Sci Rep 2020;10:7275.
[40] Shamai G, Binenbaum Y, Slossberg R, Duek I, Gil Z, Kimmel R. Artificial Intelligence Algorithms to Assess Hormonal Status From Tissue Microarrays in Patients With Breast Cancer. JAMA Netw Open 2019;2:e197700.
[41] Bychkov D, Linder N, Tiulpin A, Kücükel H, Lundin M, Nordling S, et al. Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. Sci Rep 2021;11:4037.
[42] Anand D, Kurian NC, Dhage S, Kumar N, Rane S, Gann PH, et al. Deep Learning to Estimate Human Epidermal Growth Factor Receptor 2 Status from Hematoxylin and Eosin-Stained Breast Tissue Images. J Pathol Inform 2020;11:19.
[43] Tuominen VJ, Tolonen TT, Isola J. ImmunoMembrane: a publicly available web application for digital image analysis of HER2 immunohistochemistry. Histopathology 2012;60:758-67.
[44] Tewary S, Arun I, Ahmed R, Chatterjee S, Mukhopadhyay S. AutoIHC-Analyzer: computer-assisted microscopy for automated membrane extraction/scoring in HER2 molecular markers. J Microsc 2021;281:87-96.
[45] Vandenberghe ME, Scott MLJ, Scorer PW, Söderberg M, Balcerzak D, Barker C. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Sci Rep 2017;7:45938.
[46] Saha M, Chakraborty C. Her2Net: A Deep Framework for Semantic Segmentation and Classification of Cell Membranes and Nuclei in Breast Cancer Evaluation. IEEE Trans Image Process 2018;27:2189-200.
[47] Franchet C, Laurent C. Analyse automatisée d'images d'hybridation in situ en fluorescence... Corresp en onco-théranostic 2019;8:52.
[48] Hartman D, Laak J, Gurcan M, Pantanowitz L. Value of Public Challenges for the Development of Pathology Deep Learning Algorithms. J Pathol Inform 2020;11:7.
[49] Qaiser T, Mukherjee A, Pb CR, Munugoti SD, Tallam V, Pitkäaho T, et al. HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology 2018;72:227-38.
[50] Qaiser T, Rajpoot NM. Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring. IEEE Trans Med Imaging 2019;38:2620-31.
[51] La Barbera D, Polónia A, Roitero K, Conde-Sousa E, Della Mea V. Detection of HER2 from Haematoxylin-Eosin Slides Through a Cascade of Deep Learning Classifiers via Multi-Instance Learning. J Imaging 2020;6:82.
[52] Laurinavicius A, Rasmusson A, Plancoulaine B, Shribak M, Levenson R. Machine-Learning-Based Evaluation of Intratumoral Heterogeneity and Tumor-Stroma Interface for Clinical Guidance. Am J Pathol 2021;191:1724-31.
[53] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206-15.
[54] Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S. Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides. JAMA Netw Open 2019;2:e1914645.
[55] Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell 2019;1:236-45.
[56] Masmoudi H, Hewitt SM, Petrick N, Myers KJ, Gavrielides MA. Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. IEEE Trans Med Imaging 2009;28:916-25.
[57] Ficarra E, Di Cataldo S, Acquaviva A, Macii E. Automated segmentation of cells with IHC membrane staining. IEEE Trans Biomed Eng 2011;58:1421-9.
[58] Brügmann A, Eld M, Lelkaitis G, Nielsen S, Grunkin M, Hansen JD, et al. Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. Breast Cancer Res Treat 2012;132:41-9.
[59] Pitkäaho T, Lehtimäki TM, McDonald J, Naughton TJ. Classifying HER2 breast cancer cell samples using deep learning. Proc Ir Mach Vis Image Process Conf 2016;1-104.
[60] Khameneh FD, Razavi S, Kamasak M. Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. Comput Biol Med 2019;110:164-74.

Unsupervised Nuclei Segmentation using Spatial Organization Priors

Loïc Le Bescond[1,2], Marvin Lerousseau[1], Ingrid Garberis[2], Fabrice André[2], Stergios Christodoulidis[3], Hugues Talbot[1], and Maria Vakalopoulou[3]

[1] Université Paris-Saclay, CentraleSupélec, Centre de Vision Numérique, 91190, Gif-sur-Yvette, France.
[2] Gustave Roussy Cancer Campus, 94800 Villejuif, France.
[3] Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, 91190, Gif-sur-Yvette, France.

Abstract. In digital pathology, various biomarkers (e.g., KI67, HER2, CD3/CD8) are routinely analysed by pathologists through immunohistochemistry-stained slides. Identifying these biomarkers on patient biopsies allows for a more informed design of their treatment regimen. The diversity and specificity of these types of images make the availability of annotated databases sparse. Consequently, robust and efficient learning-based diagnostic systems are difficult to develop and apply in a clinical setting. Our study builds on the observation that the overall organization and structure of the observed tissues is similar across different staining protocols. In this paper, we propose to leverage both the wide availability of hematoxylin-eosin stained databases and the invariance of tissue organization and structure in order to perform unsupervised nuclei segmentation on immunohistochemistry images. We implement and evaluate a generative adversarial method that relies on high-level nuclei distribution priors through comparison with largely available hematoxylin-eosin stained cell nuclei masks. Our approach shows promising results compared to classic unsupervised and supervised methods, as we demonstrate on two publicly available datasets.

Keywords: Precision medicine · Biomedical imaging · Digital pathology · Generative Adversarial Networks.

1 Introduction

Learning nuclei segmentation models is a challenging problem for immunohistochemistry (IHC) stained histological images. In routine pathology, IHC images are used to provide a distinct readout for proteins at the surface of nuclei or cell membranes that would otherwise be invisible to the human eye, using immunostains [3]. IHC is widely used for diagnostic and for treatment selection, notably in cancer pathology, since it bypasses the need to perform expensive and time-consuming genetic testing. There are over 100 immunostains routinely used by pathologists, highlighting different proteins such as Ki67 and HER2, which can provide clues to tumor proliferation. The segmentation of nuclei stained as such

provides essential information for distinguishing benign cells from malignant cells or those which express a specific protein from those which do not. The ability to automatically identify and segment nuclei in IHC images is crucial since It could *(i)* accelerate the diagnosis time of cancers, *(ii)* reduce misdiagnosis in routine pathology, and *(iii)* improve the performance of cell-based learning system for therapy response prediction.

The most popular nuclei segmentation approaches currently rely on manually obtained, careful pixel-based annotations of nuclei [16,17,26,30]. However, producing such annotations is time-consuming, cumbersome, tedious and error-prone, which hampers the development of segmentation models for a wide range of immunostains. Some semi-supervised methods such as [11] have been proposed to alleviate this need, requiring however manual interactions making their use on whole slide level time consuming. On the other hand, current unsupervised segmentation approaches, such as those based on color clustering, perform inadequately, preventing their application in clinical settings.

This study introduces an approach that revolves around a simple idea: we exploit the fact that the spatial organization and shape characteristics of cells in histological tissue do not change significantly with the type of stain used to color tissue slides. Specifically, we design and evaluate a powerful and highly versatile adversarial-based approach that leverages already publicly available nuclei annotations for haematoxylin-eosin (H&E) stainings to learn segmentation models for potentially many types of immunostains.We show in our experiments that our approach is effective for two of the most prevalent types of nuclear-based and membranous-based immunostains. On these examples, our approach obtains results which are close to fully supervised approaches evaluated on two publicly available datasets, without requiring any annotation.

2 Related Work

Nuclei segmentation is attracting a lot of attention lately with different challenges focusing on methods that can provide accurate segmentations for the many and diverse nuclei present on histology slides [5,32]. These challenges however focus on fully supervised methods, mostly in the domain of H&E stains. Similar approaches relying on manually obtained pixel-based annotations on H&E sometime generalize to some IHC stains e.g., for HER2-stained segmentation [27] and StarDist [29]. However, these methods essentially use color augmentation strategies [18,24], which would be specific to each new staining. In practice, there is a trade-off between the available amount of annotated tiles and the expressive power of the annotations: a higher number of annotated tiles can improve the generalization performance of segmentation systems due to the higher variability of the training data.

Conversely, a variety of thresholding-based approaches have been investigated for unsupervised nuclei segmentation, either based on Otsu thresholding [12,20] or constrained local thresholding [2,19,31]. Self-supervised learning has also been investigated for nuclei segmentation. In [28], authors train a network to accu-

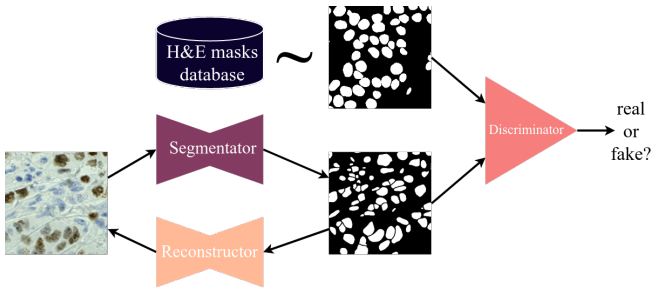


Fig. 1. An overview of our proposed framework.

rately classify the magnification of an input tile using an attention module, and show that the attention maps can be used to produce detection maps of nuclei in H&E staining which can be further converted into nuclei segmentation maps.

Cross-domain learning is a paradigm that consists of the adaptation of a model from one domain to another; for instance from the H&E domain to the IHC domain. In this vein, authors of [13] tackle H&E-IHC cross-domain learning by matching the distribution of high-level features obtained from both domains, for tissue segmentation. Other recent approaches have leveraged the use of generative adversarial networks (GANs) to train segmentation network with various approaches. GANs can be used to generate images via style transfer and use annotations provided in a domain into another, which can then be used to train a supervised network like a U-Net of a Mask-RCNN [13,14]. Moreover, in [33] an auto-encoder like approach for image-to-image translation for style transfer is proposed, learning the segmentation and transfer simultaneously.

Contrary to these approaches, our method exploits the available information at the segmentation level, by encoding and identifying the histological tissue characteristics that are independent of the explored staining. To the best of our knowledge this is the first time that such a scheme is explored, and shown to provide close to fully supervised performance.

3 Methodology

In this study, we propose an unsupervised method for nuclei segmentation incorporating priors from public available datasets. The intuition for our work is that the underlying spatial organization of cells within tissues is the same irrespective of staining. For a given immunostain, rather than relying on specific pixel-based annotations, our approach exploits generic pixel-based segmentation annotations from classical H&E-stained histological images.

Our architecture is composed of three different components trained jointly, as illustrated in Figure 2. The first is a generator (S) which generate segmentation maps from the IHC inputs. The output of S is then processed by a discriminator (D_S) which predicts if the produced segmentation is plausible or not. Moreover, real, unpaired segmentation from public datasets are given to the discriminator,

in order to guide and encode real tissue characteristics. The last component of our method relies in a reconstruction generator (R) which trained to reconstructs IHC-looking nuclei from a segmentation. This way our framework enforces consistency priors between the generated and the real tiles.

Formally, given an input IHC tile t from a training of T, the segmentator S produces a predicted segmentation map $S(t)$. Given a database $\mathcal{DB}$ of segmentation maps from any type of staining (e.g., H&E), a ground-truth segmentation map S_{GT} is sampled from $\mathcal{DB}$ for each IHC tile t. The discriminator D is then asked to correctly predict that each S_{GT} is real (label 1) and that each predicted IHC segmentation map S_{GT} is fake (label 0); this is done by minimizing $\mathcal{L}_D$:

$$\mathcal{L}_D = \big(D(S_{GT}) - 1\big)^2 + \big(D(S(t))\big)^2 \tag{1}$$

Conversely, the segmentator S is optimized by maximizing its ability to fool, i.e. by minimizing the loss function $\mathcal{MSE}_G$ that is:

$$\mathcal{L}_G = \big(D(S(t)) - 1\big)^2 \tag{2}$$

The two examined losses $\mathcal{L}_D$ and $\mathcal{L}_G$ should train S to produce segmentation maps that contain nuclei object that resembles true nuclei segmentation (shape prior), and that display nuclei distribution similar to the nuclei distribution of $\mathcal{DB}$ (organization prior). In fact, with both losses $\mathcal{L}_D$ and $\mathcal{L}_G$ combined, the system is optimized when the distributions of $\mathcal{DB}$ and $\{S(t), t \in T\}$ are matched.

However, we found that the segmentation model S tended to produce false negatives by failing to segment some nuclei. The reconstructor R is intended to circumvent this, by reconstructing the input IHC tile t from its predicted segmentation $S(t)$; nuclei that would be missed by S would then induce errors in the reconstruction $R(S(t))$, therefore inducing S to minimize the number of false negatives. R is trained by minimizing the reconstruction $\mathcal{L}_R$, where an ℓ_1 norm is used for sparsity:

$$\mathcal{L}_R = \big\|R(S(t)) - t\big\|_1 \tag{3}$$

Following CycleGAN [34], we add another discriminator on the IHC and reconstructed IHC domains, in order to train R (and therefore S through backpropagation) with an additional loss beyond pixel-based. For simplicity, we merge this discriminator loss within $\mathcal{L}_D$, and the corresponding adversarial loss of R within $\mathcal{L}_R$ and $\mathcal{L}_G$.

Furthermore, we introduce an additional consistency loss for robustness and to ensure that the segmentator does not solely focus on color for decision making. For an IHC tile t, we consider two color augmentations c_1 and c_2 (e.g. color jitter) and two augmented views $c_1(t)$ and $c_2(t)$ of t. The consistency loss is defined as the ℓ_1 norm between the predicted segmentation maps of both augmented views:

$$\mathcal{L}_C = \big\|S(c_1(t)) - S(c_2(t))\big\|_1 \tag{4}$$

Finally, we sharpen the predicted segmentation maps $S(t)$ by multiplying the predicted logits of the segmentator S using a sharpening factor r=60, similarly

to [7]. This result in saturated value of 0 or 1 rather than float values in $[0, 1]$ which can be used by the discriminator D to easily identify fake segmentations.

The system is trained end-to-end, by optimizing both modules through minimization of the following loss function:

$$\mathcal{L}_{\text{system}} = \mathcal{L}_D + \mathcal{L}_G + \mathcal{L}_R + \mathcal{L}_C \tag{5}$$

4 Experimental Configuration

4.1 Databases

We performed extensive experiments on 3 immunohistochemistry datasets to measure the performance of all benchmarked approaches. In detail, we utilize the **DeepLIIF dataset** [6] with 1667 Ki67-stained fields of view of size 512 pixels at 40x magnification. We used the same data splitting than publicly available, i.e. 709 images for training, 303 for validation and 598 for testing. Immunofluorescence correspondences in the dataset were discarded for the current study. Each image is supplied with ground-truth annotations, which were used for testing purposes but never for our training except for the fully supervised benchmark comparisons. Ki767 IHC images are actually colored with haematoxylin, which marks all nuclei, and Ki67, which marks pKi67 at the surface of some nuclei; nuclei of Ki67-stained images are thus either brown or blue. We also employ he **BCDataset** [8] which consists of 1338 Ki67-stained 640 pixel-width 40x fields of view. Each nucleus is annotated with a single point highlighting its center. These were never used for our training except for testing purposes. Lastly, we use also the **Warwick HER2 dataset** [22,23] which contains 84 HER2-stained whole slide images (WSI) split in 50 training and 34 testing images. We extracted 256x256 patches from each tiles after performing contours detection and filtering based on texture and lightness criteria [15]. To get a good representation of each tissue, we performed K-Means clustering on the Resnet features of each patch and selected for each one the closest to centroids [10]. As KMeans is sensitive to outliers, we applied an isolation forest algorithm to remove the few artifacts that may remain after our pre-processing steps. For the testing set, we divided the patch sets into 2 folds leading to 68 patches. Similarly, for the training set, we divided the patch sets into 14 folds leading to 700 patches. The testing tiles were finally annotated by a expert anatomopathologist.Compared to Ki67, HER2-stained images are more challenging since HER2 marks the membranes of cells (and not their nuclei).

4.2 Baselines

We compare the performance of our proposed method with five competing methods, including two fully supervised approaches. Specifically: a fully supervised model based on **Unet** [9] was utilized. Moreover, **NuClick** [11] was also employed, a weakly supervised approach specifically designed to compute nuclei masks from point annotations at the center of each cell. To train this model,

a senior pathologist manually annotated all nuclei centers in HER2, and such centers were obtained by computing the centroid of each nuclei ground-truth mask for both DeepLIIF and BCDataset datasets. Furthermore, **StarDist** [29] is a supervised method originally trained on H&E images. For our problem, this approach can be considered unsupervised since it does not rely on extra annotations. StarDist was used as a plugin within QuPath [1]. **Thresholding** was performed by applying Otsu thresholding on the Gaussian filtered luminance image. We also applied the same protocol to images obtained through color deconvoluton [25] but this method was found to perform worse. The **proposed** approach was implemented with Unet-styles segmentator S and reconstructor R, and PatchGAN-based discrimators D_R and D_S [9]. At each iteration, a segmentation map $S(t)$ is produced by G for an input IHC tile t. $S(t)$ is forwarded into the discriminator D_S, along with a randomly sampled segmentation map from the Pannuke dataset [4,5] which contains nuclei instance masks of H&E tiles extracted at either 20X or 40X magnification. Similarly, the reconstructor R outputs from these masks simulated IHC images that are compared to the real ones through discriminator D_R. As we found that the reconstruction represents a key factor in the training of our method, we leveraged HER2 membranous nature to train our approach reconstructing only the deconvolved hematoxylin images as nucleus are only highlighted by this marker in this setting [25].

4.3 Implementation Details

We trained the generator using 64 filters in the last convolutional layer and a dropout of 0.5 and Adam optimizer with a learning rate of 0.0002 and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For the discriminator, we used 64 filters and 3 layers in total, with the same parameters for the optimizer. We exploited nucleus invariance to rotation and flipping to perform data augmentation. Moreover, as our datasets are all extracted from slides scanned at 40X magnifications, we performed random resizing to simulate 20X magnifications images and reproduce Pannuke distribution.

The fully supervised Unet based architecture was tuned on the number of filters $ngf \in [64, 128]$, the dropout value $p \in [0.3 : 0.5]$, the learning rate $lr \in [10^{-5} : 10^{-2}]$, the decay rate $dr \in [10^{-10} : 10^{-3}]$ and the batch size $bs \in [10, 30, 60, 120, 140]$ using a gaussian process algorithm during 50 iterations maximising the F1 score on the validation set after 10 training epochs.

Both fully supervised Unet and the unsupervised proposed approach were then trained on a single A100 GPU for up to 600 epochs with PyTorch v1.10 [21]. For Unet, the model with the lowest validation score was inferred on the (shared) testing set of the DeepLIIF dataset (and was not trained on both other datasets because of missing ground-truth training data). For proposed, the final model was selected by finding the minimum of the system loss $\mathcal{L}_{\text{system}}$ after 250 epochs to discard early training instabilities. To extract nucleus on our method, we first applied a median filter with a window size of 5 to remove the noise that may remain on our final predictions. For HER2 images, we applied in addition an

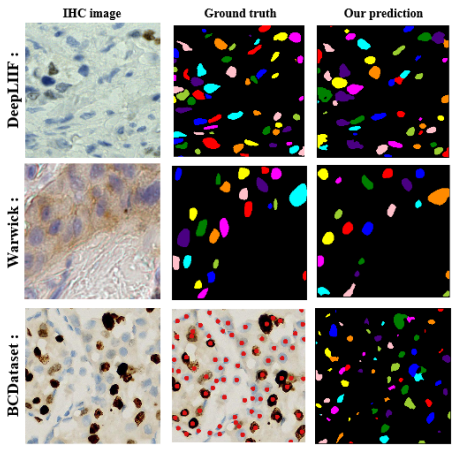


Fig. 2. Examples of predictions of our approach on the different datasets

erosion operation with a radius of 5 to remove remaining artifacts. We lastly applied a watershed algorithm to retrieve final instances [28].

5 Results & Discussion

		Semantic				Object		
Method	Unsupervised	Dice	Accuracy	Precision	Recall	AJI	Dice	Hausdorff
Unet	✗	**77.56**	**84.49**	81.59	**73.95**	40.02	64.30	5.10
Nuclick	✗	76.19	82.57	**86.12**	68.36	**56.70**	**73.18**	**4.76**
Threshold	✓	64.74	75.59	76.68	56.24	29.88	51.66	5.80
StarDist	✓	62.06	73.03	**88.16**	47.95	40.80	52.06	**5.21**
Proposed	✓	**70.27**	**79.86**	74.55	**66.60**	**41.91**	**54.43**	5.77

Table 1. Results on the DeepLIIF dataset [6]. Accuracy is balanced. **Bold** indicates the top performing method for each metric, for both supervised and unsupervised groups.

Table 1 reports semantic and object-level results on the DeepLIIF dataset for Ki67-stained images. The proposed approach obtained the highest Dice score of 70.27 and the highest balanced accuracy of 79.86 among all unsupervised approaches, i.e. approaches that do not necessitate additional annotations. While StarDist [29] obtained a higher precision, the proposed obtained the best recall of unsupervised approaches with 66.60; trading recall for precision is better for clinical considerations as false negative could aggravate the course of the patient care, while false positive can be more easily corrected. The proposed approach obtained competitive results with the fully supervised Unet and the

weakly supervised NuClick which obtained no more than 6% of improvement on the balanced accuracy without requiring any further annotations.

		Semantic				Object		
Method	Unsupervised	Dice	Accuracy	Precision	Recall	AJI	Dice	Hausdorff
Unet	✗	NA	NA	NA	NA	NA	NA	NA
Nuclick	✗	44.55	71.81	40.10	50.46	38.95	56.40	5.21
Threshold	✓	39.85	63.57	64.40	29.05	17.71	39.85	5.59
StarDist	✓	56.85	71.83	**77.86**	44.94	34.81	57.73	**4.29**
Proposed	✓	**62.59**	**77.05**	70.67	**56.51**	**39.29**	**60.01**	4.68

Table 2. Results on the Warwick dataset [22,23]. Accuracy is balanced. **Bold** indicates the top performing method for each metric, for both supervised and unsupervised groups. Unet results are not available (NA) since ground-truth segmentation maps were unavailable.

The Table 2 outlines the results on the testing set extracted from Warwick dataset. Our approach outperforms the other methods on almost all metrics, showing great improvements in semantic metrics with a Dice score of 62.59 and a recall of 56.51 while the classic methods top at 56.85 and 50.46 respectively. Once again, our approach proved to be better tailored to a clinical use with a higher recall and better object metrics. It is also advocating for a great adaptability of our method to the diversity of IHC staining. Indeed, providing minor changes in the training and post-processing, we successfully applied our method to two different staining conditions, thus underlying that our method can better leverage H&E information than directly applying pre-trained state-of-the-art algorithms.

Finally, we assessed the generalisation of the proposed approach trained on DeepLIIF to BCDataset dataset. As highlighted in Fig.2, our approach managed to provide a segmentation matching many ground truth annotations without adding any additional knowledge.

We performed an ablation study that can be found in the supplementary material by successively removing some key components of our method and computing the performances on both DeepLIIF (Ki67 staining) and Warwick (HER2 staining). On both dataset, removing the cycle loss decreased the performances significantly on all the metrics, and produced masks uncorrelated to the input, thus underlying the key role of the proposed cycling architecture. For the consistency loss and the sharpening factor, we noticed that these two elements balanced each other, with a stronger precision but a lower recall when decreasing the sharpening factor, and inversely when removing the consistency loss.

6 Conclusion

In this paper, we introduced a simple yet effective and unsupervised framework for nuclei segmentation integrating spatial organization priors. Extensive exper-

iments on 3 highly heterogeneous datasets highlight the potential of this approach. In particular, we found that our approach outperformed all other benchmarked unsupervised methods as well as some weakly supervised approaches.

There are several axes of improvements over this work. First, besides the nuclei segmentation and detection information, the type of nuclei is also an important information in routine pathology. The current formulation could integrate such information by outputting one segmentation mask per stain and counterstain of IHC images (e.g. HER2 and haematoxylin). Another very interesting direction include the integration of additional datasets or segmentation masks, which would unravel further shape and organization priors for the nuclei.

References

1. Bankhead, P., Loughrey, M.B., Fernández, J.A., et al.: Qupath: Open source software for digital pathology image analysis. Scientific reports **7**(1), 1–7 (2017)
2. Di Cataldo, S., Ficarra, E., Acquaviva, A., Macii, E.: Automated segmentation of tissue images for computerized ihc analysis. Computer methods and programs in biomedicine **100**(1), 1–15 (2010)
3. Duraiyan, J., Govindarajan, R., Kaliyappan, K., Palanisamy, M.: Applications of immunohistochemistry. Journal of pharmacy & bioallied sciences **4**(Suppl 2), S307 (2012)
4. Gamper, J., Alemi Koohbanani, N., Benet, K., et al.: Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: European Congress on Digital Pathology. pp. 11–19. Springer (2019)
5. Gamper, J., Koohbanani, N.A., Benes, K., et al.: Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778 (2020)
6. Ghahremani, P., Li, Y., Kaufman, A., et al.: Deep learning-inferred multiplex immunofluorescence for ihc image quantification. bioRxiv (2021)
7. Hou, L., Nguyen, V., Kanevsky, A.B., et al.: Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. Pattern recognition **86**, 188–200 (2019)
8. Huang, Z., Ding, Y., Song, G., et al.: Bcdata: A large-scale dataset and benchmark for cell detection and counting. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 289–298. Springer International Publishing, Cham (2020)
9. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017)
10. Kalra, S., Tizhoosh, H., Choi, C., et al.: Yottixel – an image search engine for large archives of histopathology whole slide images. Medical Image Analysis **65**, 101757 (2020)
11. Koohbanani, N., Jahanifar, M., Tajadin, N.Z., Rajpoot, N.: Nuclick: a deep learning framework for interactive segmentation of microscopic images. Medical Image Analysis **65**, 101771 (2020)
12. Kuok, C.P., Wu, P.T., Jou, I.M., et al.: Automatic segmentation and classification of tendon nuclei from ihc stained images. In: Seventh International Conference on Graphic and Image Processing (ICGIP 2015). vol. 9817, p. 98170J. International Society for Optics and Photonics (2015)
13. Lin, Z., Li, J., Yao, Q., et al.: Adversarial learning with data selection for cross-domain histopathological breast cancer segmentation. Multimedia Tools and Applications pp. 1–20 (2022)
14. Liu, D., Zhang, D., Song, Y., et al.: Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4243–4252 (2020)
15. Lu, M.Y., Williamson, D.F., Chen, T.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
16. Mahanta, L.B., Hussain, E., Das, N., et al.: Ihc-net: A fully convolutional neural network for automated nuclear segmentation and ensemble classification for allred scoring in breast pathology. Applied Soft Computing **103**, 107136 (2021)

17. Mao, K.Z., Zhao, P., Tan, P.H.: Supervised learning-based cell image segmentation for p53 immunohistochemistry. IEEE Transactions on Biomedical Engineering **53**(6), 1153–1163 (2006)
18. Mi, H., Bivalacqua, T.J., Kates, M., et al.: Predictive models of response to neoadjuvant chemotherapy in muscle-invasive bladder cancer using nuclear morphology and tissue architecture. Cell Reports Medicine **2**(9), 100382 (2021)
19. Mouelhi, A., Rmili, H., Ali, J.B., et al.: Fast unsupervised nuclear segmentation and classification scheme for automatic allred cancer scoring in immunohistochemical breast tissue images. Computer methods and programs in biomedicine **165**, 37–51 (2018)
20. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
21. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
22. Qaiser, T., Mukherjee, A., Reddy PB, C., et al.: Her2 challenge contest: a detailed assessment of automated her2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology **72**(2), 227–238 (2018)
23. Qaiser, T., Rajpoot, N.M.: Learning where to see: A novel attention model for automated immunohistochemical scoring. IEEE Transactions on Medical Imaging **38**(11), 2620–2631 (2019)
24. Rassamegevanon, T., Feindt, L., Müller, J., et al.: Molecular response to combined molecular-and external radiotherapy in head and neck squamous cell carcinoma (hnscc). Cancers **13**(22), 5595 (2021)
25. Ruifrok, A.C., Johnston, D.A.: Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology **23**(4), 291–299 (2001)
26. Saha, M., Arun, I., Ahmed, R., et al.: Hscorenet: A deep network for estrogen and progesterone scoring using breast ihc images. Pattern Recognition **102**, 107200 (2020)
27. Saha, M., Chakraborty, C.: Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. IEEE Transactions on Image Processing **27**(5), 2189–2200 (2018)
28. Sahasrabudhe, M., Christodoulidis, S., Salgado, R., et al.: Self-supervised nuclei segmentation in histopathological images using attention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 393–402. Springer (2020)
29. Schmidt, U., Weigert, M., Broaddus, C., Myers, G.: Cell detection with star-convex polygons. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 265–273. Springer (2018)
30. Sheikhzadeh, F., Ward, R.K., van Niekerk, D., Guillaud, M.: Automatic labeling of molecular biomarkers of immunohistochemistry images using fully convolutional networks. PLoS One **13**(1), e0190783 (2018)
31. Shu, J., Liu, J., Zhang, Y., et al.: Marker controlled superpixel nuclei segmentation and automatic counting on immunohistochemistry staining images. Bioinformatics **36**(10), 3225–3233 (2020)
32. Verma, R., Kumar, N., Patil, A., et al.: Monusac2020: A multi-organ nuclei segmentation and classification challenge. IEEE Transactions on Medical Imaging **40**(12), 3413–3423 (2021)

33. Yao, K., Huang, K., Sun, J., Jude, C.: AD-GAN: End-to-end unsupervised nuclei segmentation with aligned disentangling training. arXiv preprint arXiv:2107.11022 (2021)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

UNSUPERVISED NUCLEI SEGMENTATION USING SPATIAL ORGANIZATION PRIORS

LEVERAGING TISSUE INVARIANCE ACROSS STAINING TO PERFORM UNSUPERVISED NUCLEI SEGMENTATION.

Loïc Le Bescond, Marvin Lerousseau, Ingrid Garberis, Fabrice André, Stergios Christodoulidis, Maria Vakalopoulou & Hugues Talbot

INTRODUCTION

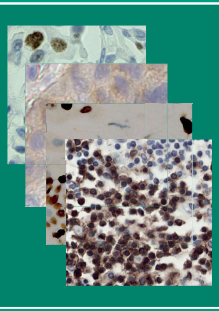
Due to the variety and specificity of staining conditions, IHC images analysis remains challenging.

Current approaches relies on manually obtained pixel-based annotations or thresholding methods requiring specific tuning.

Our approach revolves around the observation that the overall organization and structure of the observed tissues are similar across different staining protocols.

We leverage the wide availability of H&E stained databases to infer shape and distribution prior through a GAN model.

ASSUMPTIONS

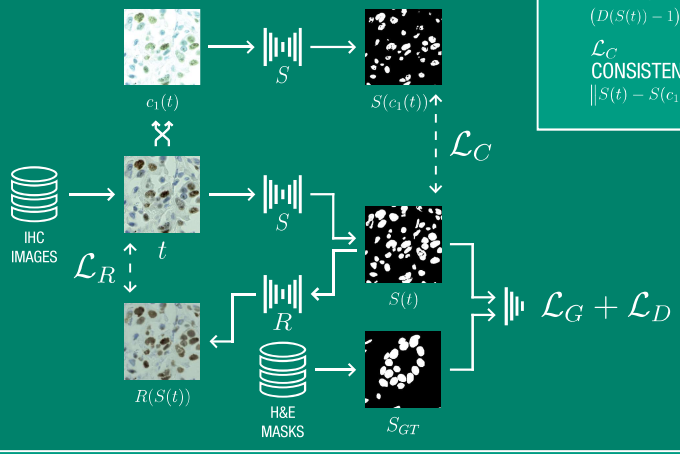


SIMILAR TISSUE ORGANIZATION AND STRUCTURE

PUBLIC H&E DATASETS

PANNUKE
MONUSEG
MONUSAC
CONIC
NUCLS

METHODOLOGY



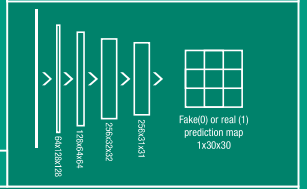
DATASET GENERATOR DISCRIMINATOR

$\mathcal{L}_G + \mathcal{L}_D$
ADVERSARIAL LOSS:
$(D(S(t)) - 1)^2 + (D(S_{GT}) - 1)^2 + (D(S(t)))^2$

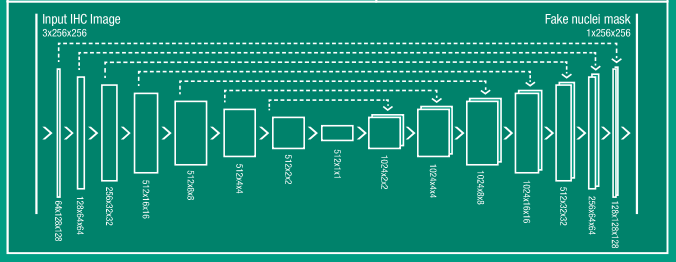
$\mathcal{L}_C$
CONSISTENCY LOSS:
$\|S(t) - S(c_1(t))\|_1$

$\mathcal{L}_R$
RECONSTRUCTION LOSS:
$\|R(S(t)) - t\|_1$

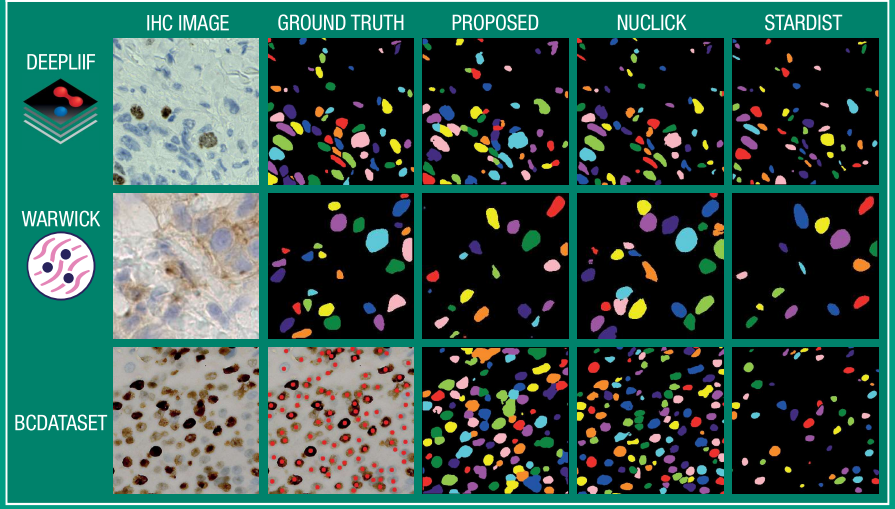
DISCRIMINATOR



GENERATOR



RESULTS



TABLES

DEEPLIIF

		SEMANTIC				OBJECT		
Method	Unsupervised	Dice	Accuracy	Precision	Recall	AJI	Dice	Hausdorff
Unet	✗	**77.28**	**84.30**	81.42	**73.67**	51.01	72.52	12.81
Nuclick	✗	76.14	82.65	**85.85**	68.62	**61.09**	**75.81**	**9.46**
Threshold	✓	64.21	75.65	76.65	56.42	37.08	58.85	17.30
StarDist	✓	61.92	73.04	87.89	48.00	40.80	60.38	**14.79**
Proposed	✓	**69.81**	**79.65**	**74.54**	**66.30**	**41.91**	**63.47**	16.18

WARWICK

		SEMANTIC				OBJECT		
Method	Unsupervised	Dice	Accuracy	Precision	Recall	AJI	Dice	Hausdorff
Unet	✗	NA	NA	NA	NA	NA	NA	NA
Nuclick	✗	70.63	89.08	64.10	82.79	56.61	70.88	7.39
Threshold	✓	43.03	67.46	**75.71**	34.65	25.49	44.56	9.87
StarDist	✓	51.95	71.32	72.41	42.71	35.44	54.95	**6.12**
Proposed	✓	**58.46**	**81.64**	64.01	**64.34**	**39.07**	**58.65**	8.71

CONCLUSION

Simple yet effective and unsupervised framework for nuclei segmentation integrating spatial organization priors.
Extensive experiments on 3 highly heterogeneous datasets highlight the potential of this approach.

Several axis of improvement:
- Integrate nuclei classification by outputting one segmentation mask per stain and counterstain.
- Include additional datasets or segmentation masks

FEEL FREE TO CONTACT US:
loic.le-bescond@centralesupelec.fr

Hyper-AdaC: Adaptive clustering-based hypergraph representation of whole slide images for survival analysis

Hakim Benkirane HAKIM.BENKIRANE@CENTRALESUPELEC.FR
CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, 91190
Oncostat U1018, Inserm, Université Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, CESP, Villejuif, 94805

Maria Vakalopoulou MARIA.VAKALOPOULOU@CENTRALESUPELEC.FR
Stergios Christodoulidis STERGIOS.CHRISTODOULIDIS@CENTRALESUPELEC.FR
CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, 91190

Ingrid-Judith Garberis INGRID-JUDITH.GARBERIS@GUSTAVEROUSSY.FR
Inserm UMR981, Gustave Roussy Cancer Campus, Villejuif, France
Université Paris-Saclay, 94270 Le Kremlin-Bicêtre, France

Stefan Michiels STEFAN.MICHIELS@GUSTAVEROUSSY.FR
Oncostat U1018, Inserm, Université Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, CESP, Villejuif, 94805
Bureau de Biostatistique et d'Épidémiologie, Gustave Roussy, Université Paris-Saclay, Villejuif, 94805

Paul-Henry Cournède PAUL-HENRY.COURNEDE@CENTRALESUPELEC.FR
CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, 91190

Abstract

The emergence of deep learning in the medical field has popularized the development of models to predict survival outcomes from histopathology images in precision oncology. Due to their large sizes, learning proper representations from whole slide images (WSI) is a crucial problem for computational pathology. Graph-based formalism has opened interesting perspectives for this challenging task, as they can be context-aware and model local and global topological structures in the tumor's microenvironment. However, the critical issue in using graph representations lies in their generalizability. They can suffer from overfitting due to their large sizes or high discrepancies between nodes due to random sampling from WSI. In addition, graphs are limited to pairwise interactions, which can sometimes fail to represent the reality observed in histopathology and hinder the interpretability of those interactions. In this work, we present Hyper-AdaC, an adaptive clustering-based hypergraph representation to model high-order correlations among different regions of the WSIs while being compact enough to help Graph Neural Networks generalize in the case of survival prediction. We evaluate our approach on 5 different cancer datasets. We outperform most state-of-the-art graph-based methods for survival prediction with WSIs, creating a more efficient and robust alternative to other graph representations. The code is available at : [Github Link Here]

Keywords: Histopathology, Hypergraphs, Survival Analysis, Representation Learning, Interpretability.

1. Introduction

Computational Pathology has rapidly developed over the past decade due to the development of whole-slide image (WSI) scanners that digitize histopathology, immunohistochemistry, or cytology slides into high-resolution images (Zhang, 2019; Lin, 2019). It has increased their exploitation for cancer diagnosis and prognosis, relying on massive progress in gigapixel image analysis with statistical learning. In this perspective, WSIs have been used for numerous prediction tasks, one of the most challenging being survival prediction (Zhu et al., 2016, 2017). It consists in modeling the survival function until the occurrence of a particular event (e.g., death, relapse). For this purpose, multiple approaches were adopted in the literature to deal with the challenge of processing large images to train a survival network.

One of the most popular methods, Multiple Instance Learning (MIL), performs weakly-supervised learning on WSIs by extracting small image patches as independent instances and aggregating them in bags of unordered instances (Sudharshan et al., 2019). However, even if this approach has performed well for some tasks like cancer grading (Zhou et al., 2019) and subtyping (Anand et al., 2020), its adaptation to survival prediction is not straightforward as it should rely on local instances as well as global-level features. Standard MIL approaches only consider bags of instances as independent and thus do not incorporate context information, failing to learn general associations in the tumor or its environment to assess patient mortality risk (Saltz et al., 2018). To alleviate this issue, graph representations have a known growing interest as they can embed global interactions between patches in a network that allows communication between instances (Adnan et al., 2020; Li et al., 2018; Chen et al., 2021). However, existing works on this subject either consider huge graphs that can hinder Graph Neural Networks' (GNNs) generalizability (Yehudai et al., 2021) or involve sampling, which covers only part of the information field and neglects lots of pathological tissues. Moreover, graph representations being limited to pairwise associations can sometimes fail to model local structures when there are significant discrepancies between instances (Garg et al., 2020).

In this work, we propose a novel hypergraph representation (Hyper-AdaC) based on adaptive clustering (Müllner, 2011) for accurate survival prediction. The contribution of this model to the representation of gigapixel WSIs is twofold. First, we deal with the limitations of the graph size by using hierarchical clustering based on both morphological similarity and spatial proximity to summarize WSIs information efficiently. This method is easy to adapt and does not rely on constraining hypotheses, like the number of clusters to consider, since it can change with the tissue morphological characteristics. This method can also be seen as a way to efficiently bypass the limitations of random patch sampling as it filters the most relevant patches from the WSI, resulting in less loss of information. Secondly, we overcome the constraints induced by the local structures thanks to our hypergraph representations of those clustered instances depending on morphological and spatial features. To validate our method, we quantitatively evaluate it on 5 different cancer datasets from The Cancer Genome Atlas (TCGA) and compare it to several other state-of-the-art methods for survival outcome prediction, proving better performance.

2. Related Work

Several methods have been developed for survival analysis in computational pathology, mainly using MIL approaches (Mobadersany et al., 2018; Lu et al., 2021; Yao et al., 2020). Those methods rely on sampling a limited number of patches to deal with the large size of WSIs. They can suffer from coverage and generalization limitations, as shown in multiple studies (Ciga et al., 2021; Di et al., 2022). To overcome those limitations, many approaches have been proposed, in which patches are grouped using clustering algorithms such as K-Means algorithm before sampling (Zhu et al., 2017; Yao et al., 2020) to identify morphological phenotypes in WSIs and reduce the dimensionality. Recent more advanced methods (Chen et al., 2021; Shao et al., 2021) started taking an interest in correlations between small instances of gigapixel images, which is neglected by the initial hypothesis of the MIL approach (Carbonneau et al., 2018). Following this idea, graph-based representations have become an excellent alternative for robust context-aware representations (Li et al., 2018; Zheng et al., 2021). To alleviate the issue of limited sampling, Chen et al. (2021) proposed a way to model interactions between features of adjacent patches using a k-nearest neighbors (k-nn) graph. As classical graph representations can only model pairwise interactions between image patches, new methods are considering broader representations by trying to lift the i.i.d. hypothesis from standard MIL (Shao et al., 2021), or by switching to hypergraph representations (Di et al., 2020, 2022). Contrary to these methods, Hyper-AdaC relies on hypergraphs to capture interesting spatial and morphological features from WSIs, harvesting informative global and local WSI dependencies for survival models.

3. Method

Within the scope of this study, we design, implement and evaluate a hypergraph based survival network for overall survival outcome prediction. For $1 \leq i \leq N$, let us denote by W_i, the WSI of a patient, T_i its event time, and C_i its censoring status. The goal of this study is to build and train a survival neural network $\mathcal{S}$ and to determine a function ϕ that maps the WSI into a hypergraph representation, such that $\mathcal{S}(\phi(W_i), \mathbf{\Theta}) = r_i$, with $\mathbf{\Theta}$ a set of trainable parameters and r_i the hazard rate of the time-to-event outcome of interest.

3.1. Hypergraph Construction

We denote by G_i a hypergraph representation of W_i such that $\phi(W_i) = G_i$. Prior to the construction of the hypergraph, we first performed automatic tissue and background separation using Lu et al. (2021). We then extract non-overlapping 256×256 patches at $20\times$ magnification that are fed to a ResNet-18 trained using the same contrastive learning strategy as in Ciga et al. (2022) that represents with a 1024-dimensional feature vector $h \in \mathbb{R}^{1024}$ each patch. The set of $(h_j)_{1 \leq j \leq n_p}$ associated to a W_i with n_p patches will be stacked into a feature matrix $\mathbf{X}_i \in \mathbb{R}^{n_p \times 1024}$. Each patch x_j is characterized by its ResNet-18 feature representation h_j that embeds the morphological properties of the patch and a set of coordinates $g_j = (g_{x,j}, g_{y,j})$ that represents the spatial position of the center of the patch. Since the hypergraph should not be too large for the generalizability of the GNN (Yehudai et al., 2021), we perform a first step of Adaptive Agglomerative Clustering on the different patches. For that, we compute two similarity matrices $K_h \in \mathbb{R}^{n_p \times n_p}$ and $K_g \in \mathbb{R}^{n_p \times n_p}$ such that $K_h = (\kappa_h(x_i, x_j))_{1 \leq i,j \leq n_p}$ and $K_g = (\kappa_g(x_i, x_j))_{1 \leq i,j \leq n_p}$ where $\kappa_h(x_i, x_j) = e^{-\lambda_h ||h_i - h_j||^2}$ is a morphological similarity

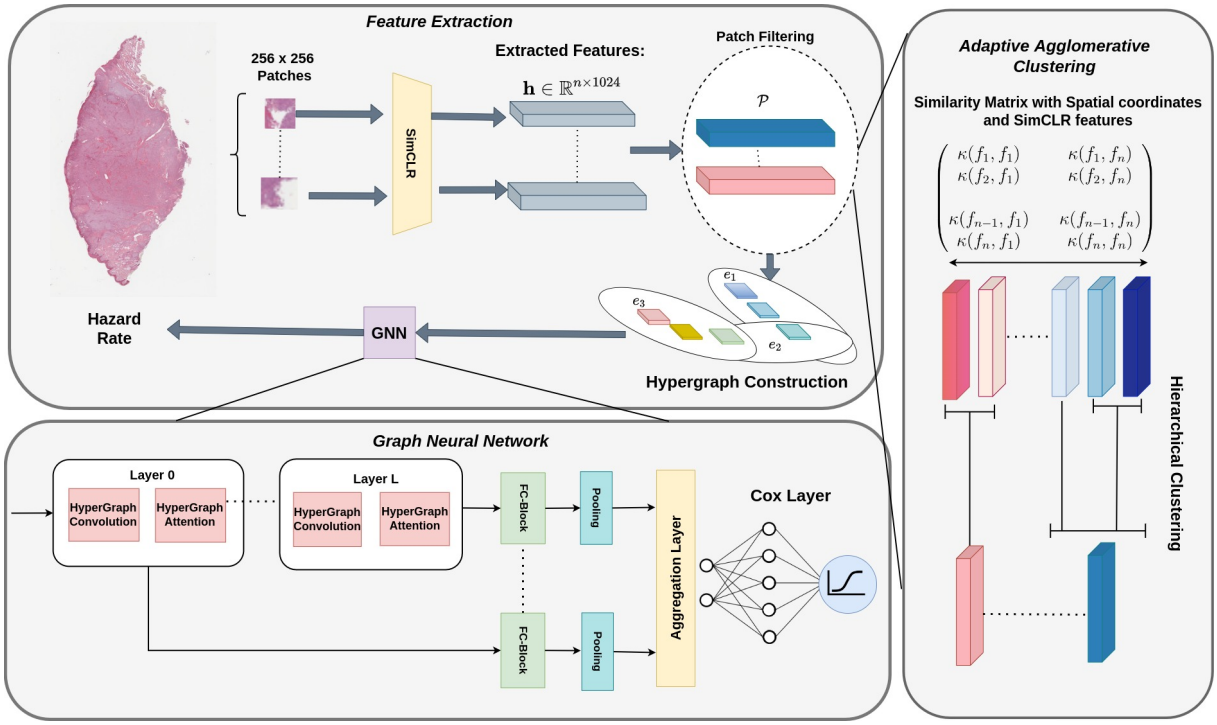


Figure 1: Overview of the Hyper-AdaC pipeline. We first perform a feature extraction step using a SimCLR framework trained on TCGA images. Then the features are processed into a clustering step that performs agglomerative clustering based on a similarity metric κ. The clustered features serve as nodes to construct a hypergraph that is then fed to a Graph Neural Network (GNN) that predict the hazard ratio. The GNN is composed of multiple Hypergraph convolutions and attention modules, followed by an FC-block (Fully-Connected block) and a global pooling layer.

metric and $\kappa_g(x_i, x_j) = e^{-\lambda_g ||g_i - g_j||^2}$ is a spatial proximity metric. Following the ideas presented in Müllner (2011), we use the kernel $\kappa(x_i, x_j) = \kappa_h(h_i, h_j)\kappa_g(g_i, g_j)$ as a similarity kernel for agglomerative clustering, as done in Lu et al. (2022). This kernel will be computed for each pair of patches from the same WSI and all patches for which similarity will be greater than a threshold δ will be considered belonging to the same cluster C_k and merged hierarchically into a single patch representation $p_k = (\tilde{h}_k, \tilde{g}_k)$ where $\tilde{h}_k = \frac{1}{|C_k|}\sum_{j\in C_k} h_j$ and $\tilde{g}_k = \frac{1}{|C_k|}\sum_{j\in C_k} g_j$. Now that we have a reduced set of points $\mathcal{P}_i$, a hypergraph denoted by $G_i =< V_i, E_i, \mathbf{X}_i >$ is constructed. For a single WSI, we consider each clustered patch as a vertex of the hypergraph such that $V_i = [p_j]_{j\in\mathcal{P}_i}$. Each hyperedge is associated to the neighbourhood of each node V_i. This neighborhood is defined as $\gamma(p_j) = \{p_k \in \mathcal{P}_i; \kappa_h(p_k, p_j) \geq \delta_h\}$, where δ_h is a threshold value to fine-tune. Those hyperedges are indicated by an incidence matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{P}_i|\times|E_i|}$ such that,

$$h(k,j) = \begin{cases} 1 & \text{if } p_j \in \gamma(p_k) \\ 0 & \text{else} \end{cases} \quad (1)$$

The interesting aspect about hypergraph compared to regular graph is that the neighborhood of each node is depicted as a single hyperedge. This allows us to train our model with fewer parameters and thus decrease the time complexity of the convolution. In addition to this, it creates a community effect that gives more importance to bigger hyperedges, which will represent denser regions of our WSI.

3.2. Construction of the Graph Neural Network

Network's Architecture: The GNN we propose (Fig. 1) consists of a series of hypergraph convolutions and attention as defined in Bai et al. (2021), with each layer using a multi-layer perceptron to generate embedding of nodes based on the features of the node itself and its neighbors. Each layer consists of batch normalization and dropout layers to avoid instability during training. We also use the idea introduced in Lu et al. (2022) of accumulating the feature representations of the convolution layers in the GNN. Those node-level representations are then pooled to generate a graph-level representation. This representation is then fed to a survival network composed of a series of multi-layer perceptrons that predict the hazard rate used for survival outcome prediction.

Network's Loss Function: The entire network is trained using the Cox-proportional hazard loss introduced in Ching et al. (2018), it uses the partial log-likelihood as the cost function, defined as follows:

$$pl(\mathbf{\Theta}) = \frac{1}{|\{i : C_i = 1\}|} \sum_{i:C_i=1} [\mathcal{S}(\phi(W_i), \mathbf{\Theta})) - log \sum_{T_i \geq T_j} exp(\mathcal{S}(\phi(W_j), \mathbf{\Theta}))] \quad (2)$$

where ψ is a neural network modelling the hazard ratio and $\mathbf{\Theta}$ are the network's parameters. The cost function to train the model is therefore defined by:

$$\mathcal{L}(\mathbf{\Theta}) = pl(\mathbf{\Theta}) + \lambda ||\mathbf{\Theta}||_2^2 \quad (3)$$

4. Experimental Setup

4.1. Dataset

For this study, we performed extensive experiments using five different cohorts from The Cancer Genome Atlas (TCGA) detailed in Table 3. We chose those 5 datasets based on size and distribution of uncensored-to-censored patients. On average, each WSI contains approximately 12691 patches at 20× magnification that are then reduced by hierarchical clustering to around 3147 points.

4.2. Implementation Details

The architecture of the GNN is constructed using three hypergraph convolution layers of 256 neurons each followed by a three layers survival network of respectively 256, 128, and 64 neurons that outputs the hazard ratio using a Sigmoid activation function in the output layer. The entire architecture is built using fully-connected blocks. For each layer, we use a batch normalization technique to address the problem of internal covariate shift. Also, to avoid overfitting problems, we use dropout with a rate of 0.2. For the graph construction, we select a similarity threshold of 80% with $\lambda_h = 3\lambda_g$ to give more importance to morphological features during the clustering. This choice of hyperparameters has been validated with the experiments presented in Appendix A. To train Hyper-AdaC, we used Adam optimization with a learning rate of 10^{-3} with an exponential scheduler, a weight decay of 10^{-5} and 20 epochs. All models were trained using an Nvidia Tesla V100S with 32 GB of memory.

Table 1: A detailed description of the cohorts used for the study. The table includes the different cancer types, as well as the number of patients and WSIs per type.

Cancer Type	**# of Patients**	**# of WSIs**
Bladder Urothelial Carcinoma (BLCA)	437	457
Breast Invasive Carcinoma (BRCA)	1022	1133
Glioblastoma & Lower Grade Glioma (GBMLGG)	389	860
Lung Adenocarcinoma (LUAD)	515	541
Ulterine Corpus Endometrial Carcinoma (UCEC)	538	566

To evaluate Hyper-AdaC, we perform 5-fold cross-validation for each cancer type, and we compute the concordance index (C-index) Uno et al. (2011) across all the validation folds to measure the predictive performance of the method. We also compare our proposed method to multiple other representations of WSIs from the literature to evaluate its contribution when faced with different state-of-the-art approaches. For all our experiments, we used the same survival loss function, the exact SimCLR feature embeddings, and training hyperparameters for all methods for a fair comparison. The basis of comparison we consider is the following:

- **DeepAttnMISL** (Yao et al., 2020): Performs standard Multiple-Instance Learning by first applying K-Means algorithm to cluster instance-level features and then process each cluster using Siamese networks.
- **DeepGraphSurv** (Li et al., 2018): A graph-based representation over sampled patches, which uses spectral GCN to consider the topological relationships between patches. We also integrate K-Means algorithm before sampling from clusters in another setup we will call C.DeepGraphSurv.
- **Patch-GCN** (Chen et al., 2021): Current state-of-the-art for GNN for survival. Performs Graph Multiple instance learning by considering the WSI as a 2D-point cloud, building a k-nearest neighbors graph.
- **knn-hypergraph** (Di et al., 2020): k-nearest neighbors hypergraph construction using sampling of patches. We use the same pipeline as Hyper-AdaC.

When comparing our approach to other methods, we see that Hyper-AdaC outperforms most of the primary methods in terms of C-index (Table 2 and Figure 2). In general, our approach is at least 1.6% better overall than every other method and in most of the separate datasets (except for BLCA and GBMLGG). When comparing with the results of DeepGraphSurv, we can immediately see the limitations of sampling patches from WSIs as this method is the weakest in these comparisons. It only covers around 20% of the WSI and fails to train GNNs due to significant discrepancies between sampled patches. We also witness a clear improvement by adding context information, as almost all the graph representations outperform the multiple-instance learning method DeepAttnMISL. Another exciting aspect is our proposed method's standard deviation between the C-index values across the 5-fold. One can observe that Hyper-AdaC reports the lowest variability of the validation C-index, suggesting a more robust model due to the compact form of its representation.

Table 2: Survival Performances of state-of-the-art methods using concordance index on 5 TCGA cohorts: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Glioblastoma & Lower Grade Glioma (GBMLGG), Lung Adenocarcinoma (LUAD) and Ulterine Corpus Endometrial Carcinoma (UCEC).

Model	BLCA	BRCA	GBMLGG	LUAD	UCEC
DeepAttnMISL (Yao et al., 2020)	0.514 ± 0.052	0.564 ± 0.050	0.781 ± 0.037	0.558 ± 0.060	0.595 ± 0.067
DeepGraphSurv (Li et al., 2018)	0.495 ± 0.045	0.551 ± 0.077	0.816 ± 0.031	0.563 ± 0.050	0.614 ± 0.052
C.DeepGraphSurv (Li et al., 2018)	0.504 ± 0.042	0.564 ± 0.043	0.787 ± 0.028	0.559 ± 0.036	0.625 ± 0.057
Patch-GCN (Chen et al., 2021)	0.561 ± 0.042	0.587 ± 0.043	$\mathbf{0.834 \pm 0.029}$	0.570 ± 0.050	0.632 ± 0.059
k-nn Hypergraph (Di et al., 2020)	$\mathbf{0.611 \pm 0.049}$	0.545 ± 0.071	0.805 ± 0.044	0.584 ± 0.061	0.615 ± 0.020
Hyper-AdaC (ours)	0.564 ± 0.034	$\mathbf{0.592 \pm 0.025}$	0.778 ± 0.024	$\mathbf{0.595 \pm 0.012}$	$\mathbf{0.667 \pm 0.022}$

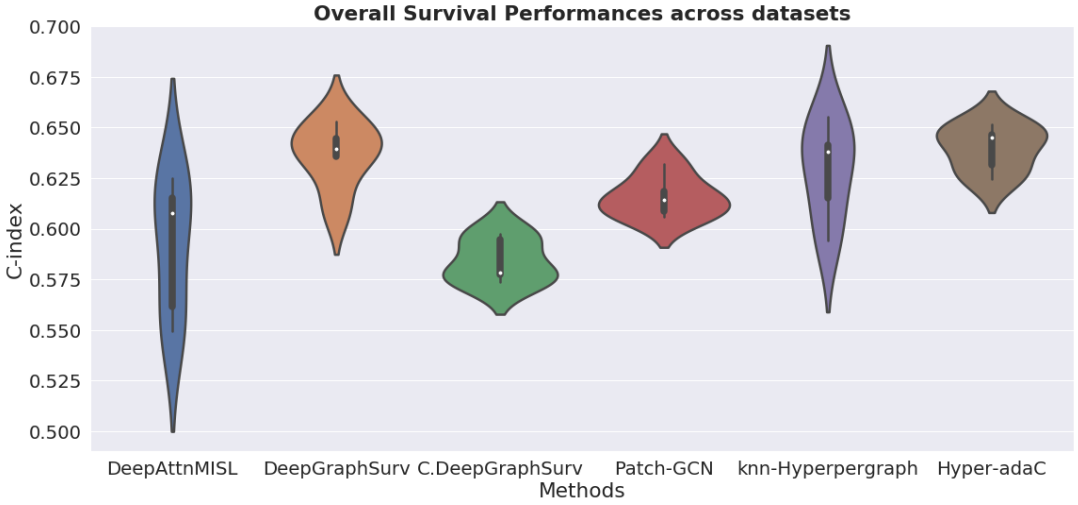


Figure 2: Overall survival performances across all datasets. They are computed by taking the C-indices on all folds of the evaluation, for all datasets.

Moreover, as the representation is smaller on Hyper-AdaC, the computing time is lower than considering a whole WSI graph because graph convolution has a worst-case complexity of $O(n^3)$ where n is the number of nodes. However, this reduction comes with a trade-off since the graph construction part is heavier due to the hierarchical clustering step that comes with the additional complexity of $\mathcal{O}(kn^2)$, where k is the final number of clusters and n is the initial number of patches. In practice, our method is about 30% slower than graphs constructed using the whole WSI like Patch-GCN or random sampling like DeepGraphSurv. On the other hand, we are almost 20% faster during training due to more compact representations, better summarized WSI, and fewer parameters. Finally, when we compare the adaptive clustering to k-means through C.DeepGraphSurv, we see that the adaptive property of the hierarchical clustering compared to K-means provides us with more information as it sums up quite well the discrepancies in the tissue without

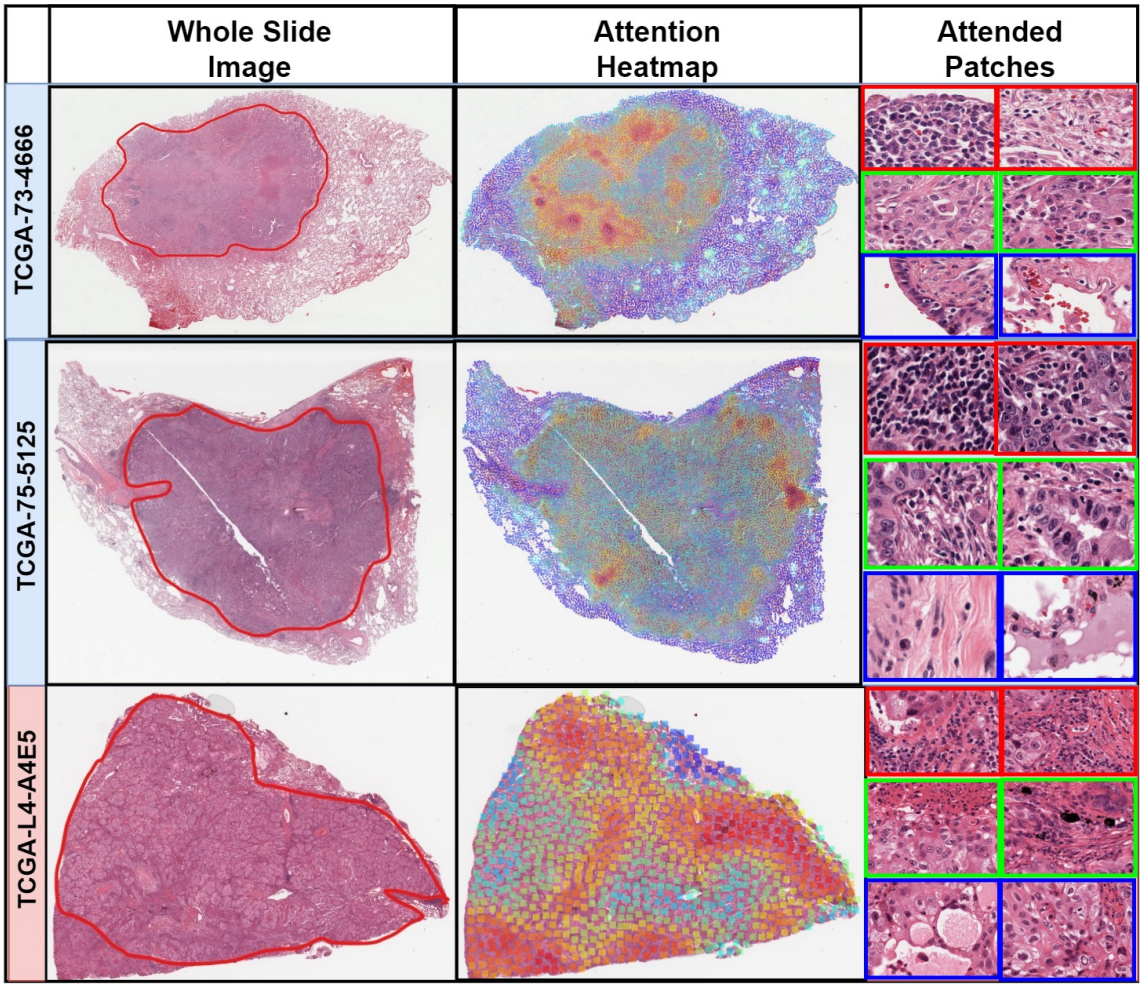


Figure 3: Comparison between the model attention heatmaps and manual annotations of tumor regions for three different patients from the TCGA-LUAD dataset (Blue for low risk patients and red for high risks patients). First column: Annotations of tumor regions (in red) in the WSI by a pathologist. Second column: Attention Heatmaps. Third column: sampled patches from 3 different attention regions; High attention (Red border), Medium attention (Green border) and low attention (Blue Border).

having to include the number of clusters as it can change from one slide to another.

Our experiments indicate lower performance on BLCA and GBMLGG datasets. To analyze more this point, we performed some additional experiments detailed in Appendix B. In fact, for BLCA we see that the number of elements conserved after the agglomerative clustering is still too high, resulting in a bigger graph and, therefore weaker performances. This reasoning can be inverted for GBMLGG for which agglomera-

tive clustering conserves only very little information, meaning that the morphological structure of this particular cancer is more homogeneous than others, and we lose a lot of information as this clustering disregards local variability. To alleviate these issues, dataset-specific hyperparameter tuning can be performed (while we originally preferred common hyperparameters for all datasets to enhance the generalizability of our model). In practice, we add more constraints on the graph construction for BLCA dataset by setting the similarity threshold δ to 85% and relax them on the GBMLGG dataset where we set $\delta = 70\%$. We also set $\lambda_h = 2\lambda_g$ for the GBMLGG dataset to focus less on morphological properties for similarity as the tissue is generally highly homogeneous, so the clustering will be more uniform across the WSI. By doing that, we can witness a spike in performance as the C-index for our method in the BLCA dataset gets to 0.619 ± 0.037 and to 0.812 ± 0.025 for the GBMLGG dataset, similar to the state-of-the-art results.

Examples of WSIs annotated by a pathologist and the corresponding model attention heatmaps are presented in Figure 3. We can observe that our model succeeds in discriminating zones based on their morphological and spatial features. Moreover, for the tumoral zone depicted by the pathologist in the first column of Figure 3 coincides with the regions where attention is at its highest. An additional exciting point is that the model focused on dense inflammatory cell regions for patients with low predicted risk, which are signs of good immunity response. The multiple purple dots highlight those inflammatory cell regions in high attention regions for the two low-risk patients (third column of Figure 3 showing a zoom of the attended patches). We witness increasing importance given to tumor cells for patients with high risk because of their density. This is where the hypergraph construction proves its advantage: it creates a community behavior with hyperedges. It can assess the density of small regions through their weights, and thanks to message passing between hyperedges, areas with more significant communities have a more decisive influence on survival prediction.

5. Conclusion

Computational Pathology has made tremendous progress when dealing with WSI global representations. However, many approaches still suffer from generalizability problems and do not properly model the whole tumor's microenvironment. In this work, we have introduced a compact hypergraph representation, Hyper-AdaC, that solves the size issue of graphs for GNNs without losing important and patient-specific information from whole-slide images. We showed through our experimentation that Hyper-AdaC creates an efficient and robust representation for training GNNs and allows broader associations between patches. An interesting perspective is to explore the efficiency of this representation in the promising context of multimodal learning for survival outcome prediction, combining WSIs with multi-omics and clinical data.

Acknowledgments

The project is supported by the Prism project, funded by the Agence Nationale de la Recherche under grant number ANR-18-IBHU-0002 and by the Public Health graduate school of Paris-Saclay University. The Data has been made available by the TCGA research network.

References

Mohammed Adnan, Shivam Kalra, and Hamid R Tizhoosh. Representation learning of histopathology images using graph

neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 988–989, 2020.

Deepak Anand, Shrey Gadiya, and Amit Sethi. Histographs: graphs in histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, pages 150–155. SPIE, 2020.

Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110: 107637, 2021.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

Ozan Ciga, Tony Xu, Sharon Nofech-Mozes, Shawna Noy, Fang-I Lu, and Anne L Martel. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, 11(1):1–10, 2021.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

Donglin Di, Shengrui Li, Jun Zhang, and Yue Gao. Ranking-based survival prediction on histopathological whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 428–438. Springer, 2020.

Donglin Di, Jun Zhang, Fuqiang Lei, Qi Tian, and Yue Gao. Big-hypergraph factorization neural network for survival prediction from whole slide image. *IEEE Transactions on Image Processing*, 31: 1149–1160, 2022.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pages 3419–3430. PMLR, 2020.

Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2018.

et al. Lin, Huangjing. Fast scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection. *IEEE Transactions on Medical Imaging*, 38:1948–58, August 2019.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. Slidegraph+: Whole slide image level graphs to predict her2 status

in breast cancer. *Medical Image Analysis*, page 102486, 2022.

Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, 2019.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, pages 11975–11986. PMLR, 2021.

et al. Zhang, Zizhao. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1:236–45, May 2019.

Yi Zheng, Rushin Gindra, Margrit Betke, Jennifer Beane, and Vijaya B Kolachalama. A deep learning based graph-transformer for whole slide image classification. *medRxiv*, 2021.

Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE, 2016.

Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7234–7242, 2017.

Table 3: Average number of nodes remaining after the hierarchical clustering step for each dataset. Due to our selection criteria, the GBMLGG dataset had a significant lower number of nodes (as the ratio is also lower, it may indicate higher homogeneity among tissues), which may explain the lower performance with respect to the other cancer types. In this study, we selected the same hyperparameters for all the cancer types to prove the generalizability of our method, outperforming the other state-of-the-art methods. Some specific hyperparameters tuning for the GBMLGG and BLCA may resolve this issue.

Cancer Type	**# of patches**	**# of nodes**	$\frac{\text{\# of nodes}}{\text{\# of patches}}$
Bladder Urothelial Carcinoma (BLCA)	58586	9187	0.16
Breast Invasive Carcinoma (BRCA)	38107	5304	0.14
Glioblastoma & Lower Grade Glioma (GBMLGG)	15855	961	0.06
Lung Adenocarcinoma (LUAD)	43445	6003	0.14
Ulterine Corpus Endometrial Carcinoma (UCEC)	56162	7748	0.14

Appendix A. Patch Clustering

We compute the average number of elements remaining after the hierarchical clustering step for each dataset separately, the results along with the ratio between initial and filtered patches are represented in Table 3. We observe that, in general, this step leaves about 14% of the WSI, and as shown in Figure 3, those elements are well spread across the WSI. However, we can see that both BLCA and GBMLGG datasets behave differently from the others. For BLCA, the ratio of remaining elements over the total number of patches is higher than all the other datasets, whereas for GBMLGG it is the opposite. Our method does not perform well for those particular test cases.

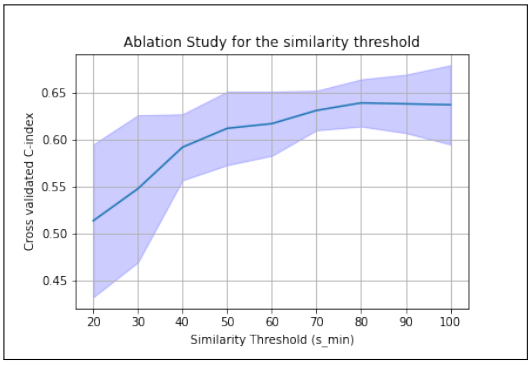


Figure 4: Ablation Study for the similarity threshold δ used in the hierarchical clustering step. We evaluate for each hyperparameter the 5-fold cross-validated C-index on the overall 5 TCGA datasets used in this study.

Appendix B. Ablation Studies

We perform an ablation study on the different graph hyperparameters to justify our construction choices. In Figure 4, we can see the effect of the similarity threshold δ_h on the survival performances. The stricter the constraint, the better the performance, indicating that larger graphs fail at learning generalizable properties. This idea is also supported by the standard deviation across the 5-folds that decreases, proof that the model is less robust with larger graphs. A similarity threshold of 80% achieves the peak performance; past that point, the performances start to decrease again because we tend to oversimplify the WSI and start neglecting information.

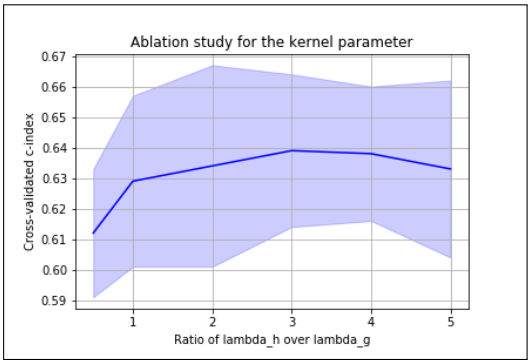


Figure 5: Ablation study for $\frac{\lambda_h}{\lambda_g}$ used in the hierarchical clustering step. We evaluate for each hyperparameter the 5-fold cross-validated C-index on the overall 5 TCGA datasets.

Figure 5 highlights the relationship between morphological features and geographical properties with respect to the overall survival performance. We see that, in general, focusing on morphological properties is more beneficial to the performances than spatial properties as they hold more information about the structure of the tissue (including, to a certain extent, spatial information because similar patches tend to be close). However, focusing too much on morphological features can hinder the accuracy of our survival predictions, as sometimes the homogeneity of specific tissues can make the filtering biased and overlook chunks of WSIs that may hold vital information.

Appendix C. About Hypergraphs

A Hypergraph is a generalization of the graph structure that extends the interaction between instances to a higher-level. To describe this complex relationship where an edge can connect to more than two nodes, we define a hypergraph $\mathcal{G} = (V, E)$ as a hypergraph with M vertices and N hyperedges. The hypergraph can then be generated using an incidence matrix $\mathbf{H} \in \mathbb{R}^{N \times M}$. For each vertex i, the vertex degree is defined as $D_{ii} = \sum_{e \in E} H_{ie}$ and the hyperedge degree will be $B_{ee} = \sum_{i \in V} H_{ie}$.

C.1. Hypergraph Convolution

This hypergraph can be associated to a feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$ where F is the feature dimension of one node. In the context of our study, this node feature will represent the aggregated Resnet-18 features of one cluster. A step of this convolution is defined in Bai et al. (2021) as follows:

$$\mathbf{X}^{(l+1)} = \sigma(\mathbf{D}^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{B}^{-1}\mathbf{H}^T\mathbf{D}^{-\frac{1}{2}}\mathbf{X}^{(l)}\mathbf{P}) \tag{4}$$

where $\mathbf{W}$ is the weight matrix, σ a non-linear transformation and $\mathbf{P}$ is the weight matrix between layer l and l+1.

C.2. Hypergraph Attention

To build the attention visualization, we used an attention mechanism for Hypergraphs described in Bai et al. (2021) as:

$$\alpha_{ij} = \frac{exp(\sigma(sim(x_i\mathbf{P}, x_j\mathbf{P})))}{\sum_{k \in \mathcal{N}_i} exp(\sigma(sim(x_i\mathbf{P}, x_k\mathbf{P})))} \tag{5}$$

where the similarity function computes similarity between two vertices as follows:

$$sim(x_i, x_j) = \mathbf{a}^T[x_i||x_j] \tag{6}$$

where $\mathbf{a}$ is a weight vector and [.||.] denotes concatenation.

Additional work performed during this thesis

- Rodrigues-Ferreira S., Nehlig A., Monchecourt C., Nasr S. Fuhrmann L, Lacroix-Triki M., **Garberis I.**, Scott V., Delaloge S., Pistilli B., Vielh P., Dubois T, Vincent-Salomon A., André F., Nahmias C. "Combinatorial expression of microtubule-associated EB1 and ATIP3 biomarkers improves breast cancer prognosis". *Breast Cancer Research and Treatment 2019, 173(3), pp.573-583*.

- Mosele F., Stefanovska B., Lusque A., Tran Dien A., **Garberis I.**, Droin N., Le Tourneau C., Sablin M-P., Lacroix L., Enrico D., Miran I., Jovelet C., Bièche I., Soria J-C., Bertucci F., Bonnefoi H., Campone M., Dalenc F., Bachelot T., Jacquet A., Jimenez M., André F. « Outcome and molecular landscape of patients with PIK3CA-mutated metastatic breast cancer". *Annals of Oncology 2020 Mar;31(3):377-386*.

- Bachelot B, Filleron T., Bieche I., Arnedos, Campone M., Dalenc F., Coussy F., Sablin M-P., Debled M., Lefeuvre-Plesse C., Goncalves A., Mouret Reynier M-A., Jacot W., You B., Barthelemy P., Verret B., Isambert N., Tchiknavorian X., Levy C., Thery J-C., L'Haridon T., Ferrero J-C., Mege A., Del Piano F., Rouleau E., Tran-Dien A., Adam J., Lusque A., Jimenez M, Jacquet A., **Garberis I.**, Andre F. "Durvalumab compared to maintenance chemotherapy in metastatic breast cancer: the randomized phase II SAFIR02-BREAST IMMUNO trial". *Nature Medicine 2021 Feb, 27(2):1-6*.

- Mosele F., Deluche E., Lusque A., Le Bescond L., Filleron T., Pradat Y., Ducoulombier A., Pistilli B., Bachelot T., Viret F., Levy C., Signolle N., Alfaro A., Tran D. T. N., **Garberis I.**, Talbot H., Christodoulidis S., Vakalopoulou M., Droin N., Stourm A., Kobayashi M., Kakegawa T., Lacroix L., Saulnier P., Job B., Deloger M., Jimenez M., Baris V., Laplante P., Kannouche P., Marty V., Lacroix-Triki M., Diéras V., André F. "Mechanism of action and resistance to Trastuzumab Deruxtecan in patients with metastatic breast cancer: the DAISY trial".

  Manuscript submitted to Nature Medicine, 2022.

- Rassy E., **Garberis I.**, Tran-Dien A., Chung-Scott V., Bouakka I., Bassil J., Ferkh R.,

Lacroix-Triki M., Zanconati F., Giudici F., Generali D., Rouleau E., Lacroix L., André F., Pistilli F. "Distinguishing new primary breast cancers from true recurrences: a comparative genomic profiling of primary and locally recurrent hormone receptor-positive breast cancers".

Manuscript submitted to Clinical Cancer Research, 2022.

Synthèse en français

C. Introduction

Le cancer du sein, malgré la stabilisation ou même la diminution des taux d'incidence dans plusieurs pays, reste la tumeur la plus fréquente chez les femmes. Cette maladie est caractérisée par une vaste hétérogénéité tant au niveau clinique comme moléculaire, incluant un large éventail d'entités avec des approches particulières. Une classification précise et exhaustive est cruciale pour identifier les tumeurs qui auront une évolution indolente ou celles qui se développeront rapidement, ainsi que les patients qui nécessiteront de traitements plus agressifs.

A l'heure actuelle, différents sous-groupes sont reconnus par les pathologistes selon des variables clinico-pathologiques, notamment le grade et le type histologiques, et d'autres caractéristiques comme la taille tumorale, la présence d'invasion lympho-vasculaire, l'envahissement des ganglions lymphatiques, et l'expression de biomarqueurs détectée par l'immunohistochimie, comprenant les récepteurs hormonaux aux œstrogènes et à la progestérone (RE, RP), HER2 et Ki67 pour évaluer la prolifération. Ces sous-groupes ont des implications pronostiques et prédictives différentes les uns des autres. La classification courante peut-être simplifiée, à l'égard de l'expression ou non des biomarqueurs mentionnés, en trois sous-types moléculaires : les tumeurs luminales (A et B), le groupe HER2-positif et les tumeurs triple-négatives.

Les avancées dans le génotypage des tumeurs sont venues à l'aide de la prise de décision thérapeutique avec les signatures moléculaires, utilisées pour prédire quels patients bénéficieront le plus probablement d'une chimiothérapie adjuvante pour réduire le risque de rechute et, tout aussi important, quels patients pourraient ne pas en avoir besoin, en évitant des traitements lourds avec des effets indésirables associés. D'autres outils pour la prédiction du risque de rechute fonctionnent avec des algorithmes appliqués à des données clinico-pathologiques, tel que Predict Breast, consultable online, et le CTS (Clinical Treatment Score).

Avec l'augmentation du volume et de la complexité des déterminations requises pour une évaluation complète de chaque cas, les laboratoires de pathologie sont confrontés au défi d'une charge de travail croissante incluant de nouvelles techniques et outils. Des modalités de

travail plus adaptées pour absorber cette charge supplémentaire deviennent urgentes. La pathologie numérique, qui commence à se déployer dans la pratique quotidienne, permet la mise en œuvre de solutions informatiques pour la gestion d'échantillons et pour l'automatisation de certaines tâches, ainsi que de l'intelligence artificielle (IA) pour, par exemple, la classification des tumeurs, qui permettront à terme d'accélérer l'administration de thérapies appropriées et de fournir un pronostic précis aux cliniciens et patients. Ce flux de travail implique la numérisation de lames histologiques dans des fichiers d'images numériques ou *whole slide images* (WSI) qui seront utilisées pour l'interprétation, l'analyse automatisée et l'archivage dans des serveurs.

L'IA est en train de gagner terrain en médecine et en particulier en pathologie en tant que nouvel outil pour améliorer la précision des stratégies diagnostiques, basé sur la création et l'application d'algorithmes, et notamment des réseaux de neurones artificielles, formés par une séquence de couches (apprentissage profond ou *deep learning*, DL), qui constituent une approche idéale pour la reconnaissance visuelle, comme le traitement des WSI. En général, ces modèles mathématiques sont dessinés en réponse à une question scientifique (par exemple, reconnaître des cellules cancéreuses dans les images histologiques) puis entraînés pour « apprendre » les traits qui pourraient résoudre la question demandée (dans l'exemple, prédire la configuration dans des cas futurs). En identifiant les trames de reconnaissance visuelle, l'IA permet le traitement de grands ensembles de données qui, en raison de leur échelle et de leur diversité, pourraient être très difficiles à gérer s'ils étaient analysés manuellement. La première étape ou entraînement consiste à fournir d'énormes quantités de données au système qui permettront à l'algorithme de s'ajuster et de s'améliorer. Dans la deuxième étape ou prédiction, comme son nom l'indique, le réseau de neurones mouline les données d'entrée pour produire des prédictions. La taille et la variété des jeux de données, ainsi que leur caractère prospectif, sont des facteurs importants à considérer afin d'éviter des résultats erronés. Quand ces données d'entrée sont annotées, on parle d'apprentissage supervisé.

L'apprentissage supervisé a toujours été l'approche la plus courante en pathologie numérique mais, en raison du coût des annotations précises, l'apprentissage faiblement supervisé devient de plus en plus populaire. L'apprentissage à instances multiples (*multiple instance learning*, MIL), un type d'algorithme d'apprentissage faiblement supervisé utilisé dans notre étude, en est un exemple.

Un point capital à considérer quand on applique des approches d'IA est l'interprétabilité des résultats. Les systèmes émulent des boîtes noires où les entrées et les sorties sont perceptibles mais le processus de décision reste "caché". Cette condition peut, d'un part, empêcher l'acceptation des solutions employant l'IA par les médecins et autres experts et, d'autre part, engendrer des difficultés pour la correction des algorithmes dont le raisonnement est inconnu. Une manière de contourner cette contrainte est l'IA « explicable », où l'effort est dirigé pour dévoiler et trouver le sens biologique aux traits qui ont un poids plus important pour la prédiction élaborée par l'algorithme.

Comme pour toute autre technologie médicale, les algorithmes d'IA doivent passer par une validation clinique avant leur application généralisée. Deux types successifs de validation ont lieu : la validation interne, qui fait référence à l'évaluation de l'algorithme avec les données qui ont été utilisées pour développer le modèle, et la validation externe, qui peut être accomplie grâce à l'étude d'un nombre suffisant d'échantillons qui n'ont pas été utilisés pour l'entraînement et qui présentent une variété de caractéristiques représentative de l'ensemble du spectre du problème étudié.

Dans le domaine de la pathologie mammaire, l'application de l'IA, en plus d'améliorer la précision du diagnostic et l'évaluation des biomarqueurs, peut fournir des résultats au-delà de ce qui peut être obtenu par une évaluation oculaire des caractères histologiques. Ces méthodes pourraient ainsi constituer une alternative moins chère et plus rapide à certains tests multigéniques comme les signatures moléculaires, pour prédire l'évolution du cancer du sein, pouvant les remplacer complètement ou du moins comme une étape préalable à la mise en œuvre de ces tests onéreux et de disponibilité encore limitée.

D. Objectives

L'objectif principal de ce travail était de développer un outil de pathologie numérique basé sur l'IA couplée aux données cliniques et supports déjà accessibles au laboratoire de pathologie comme les WSI, pour évaluer le risque de rechute à distance à 5 ans chez les patients avec un cancer du sein invasif en phase précoce (eiBC). Le but ultime en cas de succès était d'améliorer la stratification de patientes qui pourraient être observées en toute sécurité plutôt que de nécessiter d'autres thérapies complémentaires, avec les effets secondaires que cela entraîne.

Deuxièmement, nous souhaitions décrypter les éléments morphologiques pris en compte pour prédire le risque de rechute, par l'interprétation des résultats produits par les algorithmes développés. Pour ce faire, nous avons prévu d'évaluer la corrélation entre les sous-groupes pathologiques générés par l'IA et le risque de récidive, en accordant une attention particulière aux caractéristiques morphologiques extraites par le modèle qui pourraient être liées à différents résultats.

Ainsi, nos principaux objectifs étaient de :

1) évaluer si l'IA appliquée au WSI pouvait prédire la rechute métastatique à cinq ans,

2) évaluer si les éléments pronostiques fournis par l'IA ajoutent des informations supplémentaires aux paramètres pronostiques clinico-pathologiques utilisés en routine,

3) décrypter les caractéristiques utilisées par l'IA pour estimer le risque (interprétabilité).

E. Matériels et Méthodes

1. **Patients**

Nous avons utilisé deux cohortes pour notre projet.

La première, employée pour construire nos modèles, correspond aux données collectées rétrospectivement, pour la base du projet GrandTMA, auprès de patients diagnostiqués d'un cancer du sein et traités à Gustave Roussy, France. Cela a conduit à l'inclusion de 1429 patients diagnostiqués avec un eiBC ER+/HER2- avec un suivi complet et au moins 1 lame de tumeur colorée par l'hématoxyline-éosine-safran (HES) et numérisée. Les informations clinico-pathologiques supplémentaires comprenaient l'âge, la taille tumorale, le grade et le sous-type histologiques, le nombre de ganglions lymphatiques envahis, le sous-type moléculaire et le statut des biomarqueurs (ER, PR, HER2 et Ki67).

La deuxième cohorte, sélectionnée pour valider notre modèle, est issue d'une étude observationnelle et prospective française, CANTO, comprenant 915 HES WSI de patients diagnostiqués d'un eiBC ER+/HER2-.

2. **Endpoint**

L'*endpoint* choisi est l'intervalle libre de métastase (MFI, *metastasis free interval*) à 5 ans.

3. **Modèle**

Pour prédire un score de risque de rechute à distance à partir d'une lame histologique, nous avons procédé en trois étapes : le découpage des tissus dans les WSI, l'extraction des caractéristiques et la prédiction du risque. Les WSI ont été d'abord divisées en petits carrés de 76 × 76 micromètres (224 × 224 pixels) appelés « tuiles », à partir desquels des caractéristiques ont été extraites par un réseau de neurones convolutionnel pré-entraîné. Au cours du développement du modèle, les caractéristiques des tuiles ont été introduites dans le réseau avec les données de survie, et un modèle d'attention a appris à attribuer un poids à chaque tuile en fonction de sa pertinence pour prédire la rechute à distance. Enfin, le réseau a agrégé ces caractéristiques à l'aide d'une moyenne pondérée dont les poids étaient les scores d'attention, et a ainsi créé une représentation unique de chaque WSI utilisée pour la prédiction.

4. **Validation**

L'évaluation du modèle de DL et des scores cliniques standards a été réalisée sur la cohorte indépendante CANTO. Nous avons comparé les performances en termes de capacité de stratification, de sensibilité cumulée et de spécificité dynamique à 5 ans (adaptations de la sensibilité et de la spécificité classiques pour inclure non seulement l'occurrence ou pas de l'évènement d'intérêt mais aussi le temps écoulé jusqu'à cet évènement).

Nous avons aussi confronté nos performances à des scores cliniques pertinents utilisés dans la pratique quotidienne, Predict Breast (PB) et CTS (basés sur l'âge, la taille et le grade histologique de la tumeur, le nombre de ganglions lymphatiques envahis, le statut des biomarqueurs ER, HER2 et KI67. Enfin, et concernant la question de l'interprétabilité des algorithmes, nous avons évalué les caractéristiques présentes dans les tuiles sélectionnées par les modèles mathématiques afin d'identifier les éléments responsables de la prédiction du risque.

Pour faire la distinction entre les patients à haut risque et à faible risque de rechute à distance à 5 ans, nous avons défini un seuil pour le score de risque continu fourni par notre algorithme, fixé à 10 %, choisi en accord avec l'état de l'art acceptable pour la pratique clinique.

F. Résultats

1. **Sur l'ensemble des données d'entraînement**

En ce qui concerne le pouvoir discriminant du score pour prédire le MFI, notre score a

montré un c-index moyen de 0,77 (95%CI: [0.71, 0.83]). Les scores cliniques n'ont pas surpassé le score de notre modèle, avec des performances de 0,76 (95%CI: [0.70, 0.82]) tant pour PB que pour le CTS.

Notre modèle a permis de stratifier selon le MFI avec un hazard ratio (HR) de 4,19 (95%CI 2,88–6,1 ; p <0,0001) pour le groupe à haut risque par rapport au groupe à faible risque.

2. **Sur l'ensemble des données de validation**

Notre modèle a prédit le MFI en obtenant un c-index de 0.76 (95%CI: [0.72,0.77]). En comparaison, le CTS0 a atteint un c-index de 0.81 (95%CI: [0.80,0.83]) et PB atteint un c-index de 0.73 (95%CI: [0.70, 0.76]).

Un HR de 4,97 (95%CI 2.73–9.04; p<0.0001) a été observé dans la série de validation, comparable aux estimateurs de risque bien établis en routine, qui combinent des données d'entrée pathologiques, cliniques et moléculaires, atteignant des valeurs équivalentes ou supérieures en termes de sensibilité et de spécificité. De plus, la combinaison de notre score de risque avec d'autres scores de risque a donné un pouvoir discriminant encore plus élevé pour le groupe à haut risque versus le groupe à faible risque), montrant que notre modèle fournit de nouvelles informations et qu'il pourrait fonctionner comme une détermination additionnelle pour la stratification de patients à pronostic incertain.

3. **Interprétabilité**

L'analyse d'interprétabilité a mis en évidence la capacité du modèle à s'appuyer sur des caractéristiques histologiques connues ainsi que sur des interactions cellulaires complexes (comme le rapprochement spatial de fibroblastes et de lymphocytes dans des tuiles à faible risque) pour la prédiction du risque de rechute, ce qui a validé biologiquement notre approche. Les caractéristiques les plus prédictives d'un risque élevé de rechute ont été la haute densité en cellules tumorales, un degré élevé de pléomorphisme nucléaire, une architecture massive et la faible formation de tubules, des structures trabéculaires ainsi qu'une activité mitotique élevée.

G. Conclusion

Dans cette étude, nous avons conçu un outil basé sur l'IA pour prédire si un patient diagnostiqué d'un eiBC ER+/HER2- rechutera dans les cinq ans suivant la chirurgie initiale. Pour fournir un score de risque, la méthode utilise simplement une lame représentative de la tumeur,

colorée et numérisée de manière standard, déjà disponible à des fins de diagnostic au laboratoire de pathologie. Nous avons validé notre modèle sur une cohorte externe indépendante obtenant des performances au moins équivalentes à celles des scores cliniques actuellement utilisés dans la pratique quotidienne.

Contrairement à d'autres approches de pathologie numérique qui prédisent des caractéristiques morphologiques (le grade histologique, le scoring KI67) et sont donc indirectement liés au pronostic, nous avons entraîné notre modèle directement pour prédire le MFI, répondant directement à notre question clinique.

Nous avons conclu que notre étude met en évidence les avantages de l'IA appliquée à la pathologie numérique pour améliorer la stratification du risque des patientes atteintes d'un eiBC et pour élargir l'accès à des stratégies thérapeutiques personnalisées, y compris la désescalade thérapeutique.

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie : biologie - médecine - santé (CBMS)

Titre : L'intelligence artificielle appliquée à la pathologie numérique pour découvrir de nouveaux prédicteurs de l'évolution des patientes atteintes d'un cancer du sein

Mots clés : cancer du sein ; pathologie mammaire ; intelligence artificielle ; prédiction de la rechute ; pathologie numérique ; apprentissage profond

Résumé : L'intelligence artificielle (IA) émerge en médecine comme un nouvel outil pour améliorer la précision des stratégies diagnostiques. En identifiant les trames de reconnaissance visuelle, l'IA permet le traitement de grands ensembles de données qui, en raison de leur échelle et de leur diversité, pourraient être très difficiles à gérer s'ils étaient analysés manuellement. L'apprentissage profond (AP), un domaine de l'apprentissage automatique ou machine learning où la profondeur est générée par une séquence de couches, est une approche idéale pour la reconnaissance visuelle, comme le traitement de lames histologiques numérisées. Le mécanisme est basé sur l'entraînement d'algorithmes pour concevoir des modèles mathématiques capables de prédire la configuration dans des cas futurs.

L'« entraînement » consiste à fournir d'énormes quantités de données au système et à permettre à l'algorithme de s'ajuster et de s'améliorer.

Le cancer du sein (CS) est la première cause de cancer chez les femmes dans le monde. Les CS invasifs sont classés par les pathologistes en différents sous-types en fonction du grade et du type histologique et d'autres caractéristiques telles que la taille tumorale, la présence d'une invasion lympho-vasculaire, l'atteinte des ganglions lymphatiques, l'expression des récepteurs hormonaux et de HER2, ou l'index de prolifération évalué par le Ki67, avec différentes implications pronostiques et prédictives. La vaste hétérogénéité du CS a un impact particulier dans le pronostic et la réponse au traitement. De plus, il a été démontré que le risque et le moment de la récidive de la maladie dépendent également des caractéristiques tumorales mentionnées. Ce fait souligne le besoin de meilleurs outils pour la stratification des patients concernant la prise de décision thérapeutique, à la fois pour la mise en œuvre de traitements adaptés et aussi pour éviter des traitements lourds chez les patients identifiés comme ayant un risque faible de récidive. L'applicabilité de l'IA sur la pathologie numérique pourrait gérer la grande quantité de données générées afin d'optimiser son interprétation, améliorant ainsi la reconnaissance des différents sous-groupes du CS précoce qui pourraient s'associer à différentes évolutions.

L'objectif de ce travail était de développer un outil de pathologie numérique basé sur l'IA pour évaluer le risque de rechute à distance à 5 ans chez les patients avec un CS invasif en phase précoce, applicable partout et à un coût abordable. Nous avons utilisé des lames tumorales colorées par hématoxyline-éosine-safran (HES) et scannées, provenant de résections chirurgicales de CS, ainsi que des données cliniques et histopathologiques, comme données d'entrée pour entraîner des réseaux de neurones afin de prédire un risque de rechute. Dans un deuxième temps, nous avons validé les résultats obtenus sur un jeu de données externe et nous avons comparé nos performances à des scores cliniques pertinents utilisés dans la pratique quotidienne, atteignant des valeurs équivalentes ou supérieures en termes de sensibilité et de spécificité. Enfin, et concernant la question de l'explainable AI, nous avons évalué les caractéristiques présentes dans les tuiles sélectionnées par les modèles mathématiques afin d'identifier les éléments responsables de la prédiction du risque, et de valider biologiquement notre approche.

Nous avons conclu que notre étude met en évidence les avantages de l'IA appliquée à la pathologie numérique pour améliorer la stratification du risque des patientes atteintes d'un CS et pour élargir l'accès à des stratégies thérapeutiques personnalisées, y compris la désescalade thérapeutique.

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie : biologie - médecine - santé (CBMS)

Title : Artificial intelligence applied to digital pathology to discover new predictors of breast cancer patient outcome

Key words : breast cancer; breast pathology; artificial intelligence; outcome prediction; digital pathology; deep learning

Abstract : Artificial intelligence (AI) is emerging in Medicine as a novel tool for improving the precision of diagnostic strategies. By identifying patterns of recognition, AI allows the processing of large datasets, which, due to its scale and diversity, could be very difficult to handle if manually analyzed. Deep learning (DL), an area of machine learning where depth is generated by a sequence of multiple layers, is an ideal approach for visual recognition, such as the processing of scanned histological slides. The mechanism is based on training algorithms to design mathematical models that predict the configuration in future cases. "Training" involves providing huge amounts of data to the system and permitting the algorithm to adjust itself and improve.

Breast cancer (BC) is the first cause of cancer in women worldwide. Invasive BCs are classified by pathologists into different subtypes using the histological grade, type, and other features such as the tumor size, the presence of lymph-vascular invasion, lymph-node involvement and expression of hormone receptors, HER2 and Ki67, with different prognostic and predictive implication. BC vast heterogeneity has a particular impact on prognosis and response to treatment. Moreover, it has been demonstrated that the risk and timing of disease recurrence also depends on tumor characteristics aforementioned. This fact highlights the need for better tools for patients' stratification regarding treatment decision, both for the implementation of tailored therapeutic schemes and for avoiding heavy treatments in the so-called low-risk patients. The applicability of AI on digital pathology could manage the data generated in order to optimize its interpretation, improving the recognition of different pathological subgroups of early BC that could associate with different outcomes.

The aim of this work was to develop an AI-based digital pathology tool to assess the risk of distant relapse at 5 years in early invasive BC patients, applicable everywhere and at an affordable cost. We used scanned hematoxylin-eosin-safran (HES)-stained tumor slides from BC resection specimens, along with clinico-pathological data, as inputs to train neural networks in order to predict a risk of relapse. Subsequently, we validated the obtained results on an external dataset and we compared our performances to relevant clinical scores used in daily practice, attaining equivalent or superior values in terms of sensitivity and specificity. Finally, and regarding explainable AI question, we assessed the features present in the tiles selected by the mathematical models in order to identify the elements that supported the risk prediction, in order to find biologically coherent results that validated our approach.

We concluded that our study highlights the benefit of AI-based digital pathology to improve the risk stratification of breast cancer patients and expands access to personalized therapeutic strategies, including treatment de-escalation purposes.