From Research to Clinical Diagnostics: Developing and Validating Biomarkers and Artificial Intelligence for Pathology

by

Emma Rewcastle

Thesis submitted in fulfilment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

Faculty of Science and Technology
Department of Chemistry, Bioscience and Environmental Engineering
2024

University of Stavanger
NO-4036 Stavanger
NORWAY
www.uis.no

©2024 Emma Rewcastle

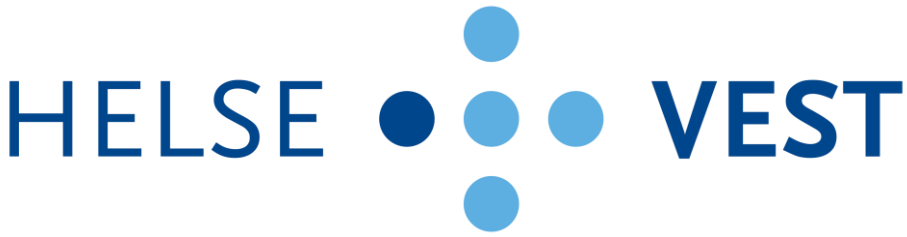
ISBN:978-82-8439-242-4
ISSN:764
PhD: Thesis UiS No. 1890-1387

Scientific Environment

The PhD project was conducted at the Department of Pathology, Stavanger University Hospital, Stavanger, Norway.

In affiliation with the Department of Chemistry, Bioscience and Environmental Engineering, Faculty of Science and Technology, University of Stavanger, Stavanger, Norway.

Supported by the *Pathology services in the Western Norway Health Region – a centre for applied digitization (PiV)*, which received funding from the Helse Vest Strategic Research Fund from 2020-2024.



Acknowledgements

"All we have to decide is what to do with the time that is given us" – JRR Tolkien

I have always been warned that entering a PhD will be one of the most challenging academic journeys to be undertaken: the final qualification. They were not wrong and therefore, I wish to extend my thanks for the exceptional support and opportunities that have been cultivated throughout this journey.

First, my sincerest gratitude to my main supervisor, Professor Emiel Janssen, without whom I would never have received this chance. Thank you for your continued patience in supervising this chronic worrier. For sharing your quick-thinking and expansive knowledge and wisdom, I am eternally grateful. Although this journey may be ending, I hope I may continue to learn from you for many years to come.

To my co-supervisors: to Dr. Einar Gudlaugsson, my resident go-to-guy for all questions medical. Thank you for taking the time to explain all things pathology to a non-pathologist like me. I appreciate your patience and your guidance and the hours you have spent going through the many cases I have placed on your desk these past years. To Dr. Ivar Skaland, your expertise has been invaluable. I have enjoyed our many discussions. Finally, to Professor Sabine Leh, thank you for your warm welcome into the PiV team, for your inclusivity, openness, and willingness to discuss all things digital pathology.

I would like to thank the Heads of the Department of Pathology, Dr. Susanne Buhr-Wildhagen, Dr. Janne Bethuelsen, and once more Prof. Emiel Janssen. To Professor Jan Baak, thank you for sharing your expertise and enthusiasm. You provided invaluable feedback on every manuscript and saved me in my time of need from spending more hours

than I had available counting cells. Additionally, to all my co-workers at the Department, particular the Units for Quantitative and Molecular Pathology, and Immunohistochemistry, thank you for your support and assistance. Also, to the Unit for Histology without whom I would still be sat bent over in front of the microtome sectioning away for hours on end, thank you! To Dr. Ole Gunnar Aasprong and Professor Kjersti Engan, for your time, thank you, your expertise has been greatly appreciated.

Special thanks to Ingrid Lundal and Desmond Abono, you have been the most excellent technical support crew, without whom I would have been buried knee deep in slides and blocks. Also, Silja Fykse, for the hours spent helping me with scanning, programming and all things excel, thank you. To Melinda Lillesand, Eliza Peixoto Albernaz, Live Egeland Eidem, and Dr. Umay Kiraz, I cannot say this enough, thank you for your unconditional support in all things. To my collaborators, PiV team and co-authors, thank you for your time and willingness to contribute and join me on this journey. Finally, to Bianca, my first teacher.

Last, but not least, to my family and extended family. To my unofficial HR representatives, R&R squad, Tea Party and Alfie pack, although not mentioned by name, your support is no less worthy, I am beyond thankful for you all. The days would have been just a little bit less bright without all of you there beside me. Thank you to my dad who has unconsciously instilled in me a thirst for science and the need to understand. The far side comics left around the house were always an added inspiration. Thanks to my brother, Dan, for providing me with an annual supply of coffee beans, which have prevented me from spending more time asleep than spent writing this thesis. Til mamma, uten deg ville jeg aldri kunne ha blitt den jeg er i dag. Du har vært roen i kaoset; Que Sera, Sera. Takk, mamma.

Abbreviations

AI:	Artificial intelligence
ANN:	Artificial neural network
APP:	Application
AUC:	Area under the curve
BEST:	Biomarkers, Endpoints, and other tools
BMI:	Body mass index
CAH:	Complex atypical hyperplasia
CAMELYON	Cancer metastases in lymph nodes challenge
CDK:	Cyclin-dependent kinase
CE-IVD(R):	Conformité Européenne in-vitro diagnostic devices regulation
CH:	Complex hyperplasia
CNN:	Convolutional neural network
CT:	Computed tomography
DIA:	Digital image analysis
DICOM:	Digital imaging and communications in medicine
DL:	Deep learning
EAH:	Endometrial atypical hyperplasia
EC:	Endometrial cancer
EEC:	Endometrial endometrioid carcinoma
EH:	Endometrial hyperplasia
EHwA:	Endometrial hyperplasia without atypia
EIN:	Endometrial intraepithelial neoplasia
ER:	Oestrogen receptor
EU:	European Union
FDA:	Food and Drug Administration
FISH:	Fluorescent in situ hybridisation
FOV:	Field of view
GLOBOCAN:	Global Cancer Observatory
GPU:	Graphics processing unit
HE:	Haematoxylin and eosin
HPF:	High power field
HR	Hormone receptor
ICC:	Intraclass correlation coefficient
ICGI:	Institute for Cancer Genetics and Informatics

IHC:	Immunohistochemistry
IKWG:	International Ki67 in Breast Cancer Working Group
IMDRF:	International Medical Device Regulators Forum
ISH:	In situ hybridisation
kNN:	K nearest neighbour
LIMS:	Laboratory information management system
LN:	Lymph node
MDT:	Multidisciplinary meeting
MIDOG:	Mitosis domain generalisation challenge
ML:	Machine learning
MRI:	Magnetic resonance imaging
MMR:	Mismatch repair
MSI:	Microsatellite instability
NIH:	The National Institutes of Health (USA)
NHS:	National Health Service (UK)
NHSx:	NHS Transformation Directorate
NPI:	Nottingham prognostic index
NPV:	Negative predictive value
NST:	No special type
OSD:	Outer surface density
PACS:	Picture archiving and communication system
PAM50:	Prediction Analysis of Microarray 50
PPV:	Positive predictive value
PR:	Progesterone
REK:	Regional Committees for Medical and Health Research Ethics
RGB:	Red green blue
ROC:	Receiver operating characteristic
R-CNN:	Region-based convolutional neural network
SAH:	Simple atypical hyperplasia
SEER:	The Surveillance, Epidemiology and End Results Programme
SH:	Simple hyperplasia
SUH:	Stavanger University Hospital
SVM:	Support vector machine
TCGA:	The Cancer Genome Atlas programme
TCGA-BRCA:	The Cancer Genome Atlas Breast Invasive Carcinoma data collection

TNM:	Tumour, node, metastasis staging
TUPAC16:	Tumour proliferation assessment challenge 2016
VPS:	Volume percentage stroma
WHO:	World Health Organisation
WHO94:	WHO 1994 Guidelines (for endometrial hyperplasia)
WHO03:	WHO 2003 Guidelines (Pathology and Genetics of Tumours of the Breast and Female Genital Organs)
WHO14:	WHO 2014 Guidelines (Tumours of Female Reproductive Organs)
WHO19:	WHO 2019 Guidelines (Breast Tumours)
WHO20:	WHO 2020 Guidelines (Female Genital Tumours)
WSI:	Whole slide image

Summary

In 2021, digital pathology was deployed in the hospitals of the western health region of Norway. Histological tissue specimens previously viewed under the microscope on glass slides, are now being scanned, and whole slide images (WSIs) are viewed digitally. Digitisation enables the use of advanced technologies to take over repetitive and time-consuming tasks such as biomarker quantification. Furthermore, digital image analysis (DIA) and artificial intelligence (AI) can be used to perform complex tasks such as pattern recognition and classification, to assist healthcare professionals.

The research presented in this thesis aims to explore methods which may improve current diagnostic and prognostic guidelines for breast cancer and endometrial hyperplasia in pathology. To challenge current limitations of visual assessments and investigate if addition of quantitative methodology and AI-assistance tools can improve reproducibility and accuracy of diagnosis and prognosis. The end goal to reduce the risk of under- and over-treatment of these patients.

In Norway, 3,000 to 4,000 women will be diagnosed with endometrial hyperplasia every year. This condition is characterised by the excessive proliferation of endometrial glands in the uterine lining. The diagnosis of endometrial hyperplasia has undergone several important evolutions in recent decades. However, the prognostic evaluation, to assess the likelihood of this condition to progress to endometrial cancer, is still limited by subjective visual assessment of tissue morphology. In the first study, the biomarkers PTEN and PAX2 were evaluated for their prognostic value in endometrial carcinogenesis. A quantitative method assessing PAX2 protein expression revealed prognostic separation of patients diagnosed with endometrial intraepithelial neoplasia with low-

and high-risk of progression to cancer. In a second study, an AI-based tool was developed, to detect and quantify morphological features of endometrial hyperplasia. The tool (ENDOAPP) was able to identify patients with low-risk and high-risk for progression. Furthermore, its accuracy was equal to and marginally superior to a semi-quantitative morphometric method (D-score) and traditional visual classifications (WHO94, WHO20, EIN), respectively.

To state that the diagnosis and treatment of cancer has a long history would be an understatement. The arrival of new technology, molecular advances and AI continues to revitalise the way cancer is viewed in the clinic. The measurement of proliferation in breast cancer has undisputed prognostic implications. However, quantification of proliferation markers is controversial citing lack of standardisation. AI may provide a promising solution for the establishment of improved methods for objective, automated, reproducible quantification of proliferation markers such as mitotic count and Ki67. In the third study, in-house and commercial DIA tools were investigated alongside manual Ki67 quantification methods for their prognostic capability and variability. It was observed that DIA tools were superior to their manual counterparts with regards to their discriminative ability for separation of low-risk and high-risk for distant metastasis free progression. Furthermore, the cut-offs currently used for binary risk categorisation of proliferation markers should be carefully re-evaluated if we wish to standardise quantification of Ki67. In the final study, a deep learning tool was investigated for the detection and quantification of mitotic count in several cancers, including breast cancer. It was observed that automated mitotic count was prognostic in multiple cancer types in addition to breast cancer, where it is routinely performed.

It is important to emphasise that the results presented in this thesis are limited to the datasets presented. These were retrospective datasets,

and often confined to a single hospital, with the exception of the fourth study. Therefore, the methods run the risk of overfitting and hidden bias. It is therefore imperative that these tools are validated in external datasets to ensure their robustness, uncover any bias or overfitting, and to confirm their prognostic validity. Although the studies presented in this thesis suggest the validity of investigating AI-tools for clinical use, further study to critically evaluate their worth is still required.

List of Publications

I **Assessing the prognostic value of PAX2 and PTEN in endometrial carcinogenesis.** Emma Rewcastle[*], Anne Elin Varhaugvik[*], Einar Gudlaugsson, Anita Steinbakk, Ivar Skaland, Bianca van Diermen, Jan P Baak, Emiel A M Janssen. (2018) Endocrine-Related Cancer 25(12):981-991

** Authors contributed equally.*

II **Automated Prognostic Assessment of Endometrial Hyperplasia for Progression Risk Evaluation Using Artificial Intelligence.** Emma Rewcastle, Einar Gudlaugsson, Melinda Lillesand, Ivar Skaland, Jan P.A. Baak, Emiel A.M. Janssen. (2023) Modern Pathology 36(5):100116

III **The Ki67 Dilemma: Investigating Prognostic Cut-Offs and Inter-Platform Reproducibility for Automated Ki67 Scoring in Breast Cancer.** Emma Rewcastle, Ivar Skaland, Einar Gudlaugsson, Silja Kavlie Fykse, Jan P.A. Baak, Emiel A.M. Janssen. (2024) *Manuscript submitted*

IV **Applicability of mitotic figure counting by deep learning: a development and pan-cancer validation study.** Tarjei S. Hveem[*], Maria X. Isaksen[*], Joakim Kalsnes[*], Frida Julbø, Manohar Pradhan, Andreas Kleppe, Sepp De Raedt, Ole-Johan Skrede, Turid Torheim, John Arne Nesheim, Hans Martin Mohn, Hanne A. Askautrud, Karolina Cyll, Wanja Kildal, Emma Rewcastle, Melinda Lillesand, Vebjørn Kvikstad, Emiel Janssen, Robert Jones, Odd Terje Brustugun, Håkon Wæhre, Bjørn Brennhovd, Erik Skaaheim Haug, Lill-Tove Rasmussen Busund, Kristina Lindemann, Gunnar Kristensen, Neil A. Shepherd, Marco Novelli, Knut Liestøl, David Kerr, Håvard E. Danielsen.
(2024) Manuscript

** Authors contributed equally.*

Table of Contents

Table of Figures

List of Tables

Appendices

1 Introduction

1.1 Digital Pathology

1.1.1 A Brief Overview of Pathology

Pathology can be traced in time throughout the records of the civilized peoples of antiquity, its foundation is built from our desire to understand disease. Pathology, from the greek '*πάθος*' – '*pathos*' meaning condition or suffering, is a branch of medicine that studies the origins and causes of disease[1]. The knowledge collated throughout the centuries forms the basis for how we view, study, and practice medicine today.

Pathology may be considered as a central hub in modern medicine. Any tissue or cytological sample removed in an operating theatre, or sampled by a general practitioner are sent to a pathology laboratory for evaluation. The specimens are handled and prepared by a team of medical laboratory personnel and a pathologist will assess and diagnose the tissue. This workflow is outline in Figure 1.

New technology has been at the forefront of paradigm shifts in medicine. In the mid-nineteenth century, the microscope was responsible for huge advances with Virchow's "*omnis cellula e cellula*" forming the basis of how we view disease today: all cells arise from other cells[2]. Recently, the field of pathology has undergone significant changes. The molecular advances made since the 1990's has played a key role in the evolution of precision medicine[3]. More recently, pathologists are transitioning from viewing histological slides under a microscope to inspecting digitized images on screens. This paradigm shift is referred to as Digital Pathology.

"I am inclined to think that the limit of what the microscope could and has done for us is now approaching and that for a further penetration into the important, all-governing problem of cell life even the most highly refined optical aids will be of no use to us"[4] – Paul Erlich, 1908

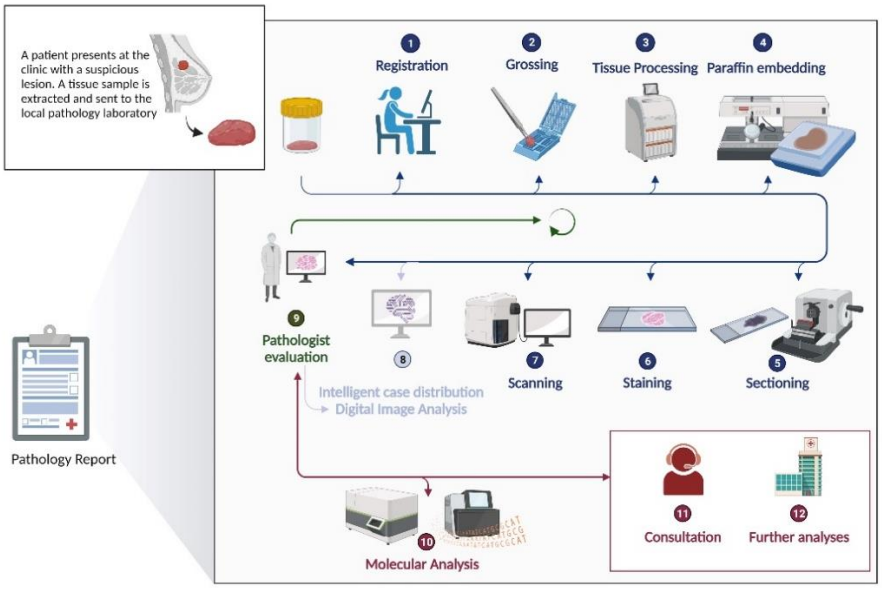


Figure 1: **Workflow of a Modern Pathology Laboratory**. *Created with BioRender.com*.

1.1.2 Digital Pathology

Digital Pathology is defined as the acquisition, management and assessment of digitised pathological information often in the form of images and data[5]. Glass slides are digitised by whole slide scanners and then stored and retrieved by a picture archiving and communication system (PACS). Medical personnel can view these images, annotate, and run digital image analysis (DIA) algorithms. Previously, examination of glass slides was performed using a microscope and any annotations were limited to the size of the slide itself and much less precise than digital annotations. Furthermore, the use of DIA was limited. With the digitisation of images, advanced technologies such as artificial

intelligence (AI) can expand the number of tasks available for DIA. This includes complex tasks such as object detection, pattern recognition, and even classification and prediction of treatment response and prognosis.

1.1.3 Strengths and Drawbacks of Digital Pathology

The primary goal of implementing digital pathology is to improve the accuracy, reproducibility, efficiency, and availability of diagnosis. It provides pathologists and technicians with the tools to better tackle the increasing workload without an equivalent increase in personnel[6-9]. The number of samples being processed in a pathology department each year is increasing due to growing and aging populations. Additionally, the complexity of diagnosis is increasing. In developing regions, where the number of pathologists and resources is severely lacking in comparison to developed regions[10,11], digitisation enables remote access to personnel and specialists[12,13].

1.1.3.1 Strengths

Numerous benefits have been cited in favour of implementing a digital workflow in pathology. Firstly, institutions can save on time and cost of shipping glass slides to other sites for specialist consultation[14]. Moreover, pathology departments no longer require internal transportation of slides from the laboratory to a pathologist's office; there is complete control over the whereabouts of physical slides. The time saved from this allows the lab to focus more on quality control of the slides themselves than on transport. The samples entering the laboratory are registered with a unique patient identifier so they can be easily tracked all the way through the lab and into the PACS. Within the PACS viewer, multiple examiners can view an image and give feedback in tandem, saving on time[15]. Collaboration within and between medical

fields is enhanced, particularly during multi-disciplinary (MDT) meetings, for training or simulation and for research purposes.

Other benefits of digital pathology include that annotations and measurements can be performed quickly, accurately, and stored for future reference. Risk of misplacing slides in the local archive is greatly reduced due to automatic assignment and storage in the local PACS according to the patient identifier. Observations from an early digital pathology implementation study included remarks from participants that areas of concern were easier to identify at a low power magnification on a whole slide image (WSI) than on glass slides. Furthermore, it was easier to assess a multi-slide case digitally owing to ease of transition between immunohistochemistry (IHC) slides in comparison to physical[14].

1.1.3.2 Drawbacks

The drawbacks of implementing digital pathology are often due to technical limitations of available technology. To start with when creating a digitised image from a glass slide it is expected that the WSI can replicate the tissue section on the glass slide, completely[16]. Any areas that are out of focus or missing in the scan could result in important information being lost[16]. WSIs should be evaluated to assess the quality of the scan and out of focus regions may cause a case to be sent for rescanning, which may delay diagnosis. Many scanners have integrated algorithms for detection of such regions, and such errors occur less frequently as technology improves.

There is yet currently no consensus on a standard image file format in Pathology. Each scanner will usually produce a WSI image file using a proprietary format and visualisation of the WSI using different image viewers or a PACS viewer can result in variations in colour and quality.

This can be equated to the staining variation observed between automated staining machines and institutions. This lack of standardisation is not optimal and requires time and resources to normalise the images. DICOM is an international standard for medical images and associated information[17]. It professes interoperability amongst different imaging hardware and viewing platforms and complies with the ISO 12052 standard[17]. It has been deployed in Radiology, and a working group (WP-26) has been established to develop this format for use in Pathology[17]. At present, it is one of the least represented formats in the field of pathology, although more and more scanner vendors offer DICOM format[18]. Not only file format can affect visualisation and thus interpretation of an image, but the display screen can also be influential. However, this has not been as thoroughly investigated as it has been in Radiology[19].

Most microscopes have a turn dial for coarse focus adjustment and fine focus adjustment. Translating the latter to a digital platform has been a challenge. Some scanners may be equipped with a multiplanar focusing option, referred to as z-stacking, which is mandatory for cytological slides and also beneficial for mitosis counting[20]. Z-stacking is a digital processing method where the tissue is scanned at different focal planes and can be combined into one composite image that has a greater depth of field[21]. However, there are limitations associated with z-stack acquisition, firstly a longer scanning speed and secondly larger file sizes[22,23]. Studies by Sturm *et al.*[24] and Kim *et al.*[25] have observed that although z-stacking provides some benefit regarding diagnostic accuracy in specialised cases, it is modest and does not necessarily justify its widespread use. Furthermore, Tabata and colleagues, observed that a higher pixel resolution and numerical aperture is sufficient to resolve this issue in the context of digital mitotic counting[23]. With time as scanner technology improves, this issue may be overcome as pathologists adjust to digital microscopy.

Deployment of digital pathology demands new skill sets. The expertise required to maintain a PACS system and its connection to the local laboratory information and management system (LIMS) is beyond medical training and requires IT competency. Creating a close partnership between IT and medical personnel is crucial for successful implementation. This ensures successful communication between parties with different skill sets and safeguards that each other's needs are met. This is particularly essential for development, validation, testing, and implementation of DIA.

In summary, although the deployment of digital pathology may face several technical challenges, these may eventually be overcome with advances in technology and experience.

1.2 Digital Image Analysis

Digital image analysis (DIA) refers to the examination of a digitised specimen by using a computer to extract data, often using algorithms[26]. Basic DIA tasks include counting objects or measuring object properties such as area or quantification of pixel values. More recently, DIA by means of AI has begun to tackle more advanced tasks such as class prediction i.e., tumour vs. non-tumour and histological grading.

1.2.1 What is an image?

To understand how DIA works we need to go back to the basics. An image is a representation of something[27]; a digital image is a 2D rectilinear array of pixels each consisting of a spatial coordinate and a value[28,29]. A digital image will exist in a colour space, pixels in a greyscale image will have a coordinate (x, y) and a value between 0-255, whilst an RGB image will have a coordinate and value in each colour space: red, green and blue[28] (Figure 2). Whilst a pixel is the smallest unit of an

image or display, a bit is the smallest increment of data, existing as 0 or 1[28]. The bit depth or number of bits per pixel in an image describes the colour information (Figure 3). A bit depth of 1 can display two shades: 0 or 1 (2^1). A bit depth of 3 (2^3) can display 8 shades, with 8 different combinations of 0's and 1's: 000, 001, 010, 011, 100, 101, 110, and 111. In an 8-bit greyscale image, there are 256 possible combinations (2^8) and a single pixel will have a value of 0 (black) to 255 (white)[28]. This conveniently closely represents what is perceivable by the human eye (Figure 2). In an RGB image, each colour space (red, blue and green) will have 256 possible combinations and together create a 24-bit image with 16,777,216 possible combinations[28,29] (figure 2).

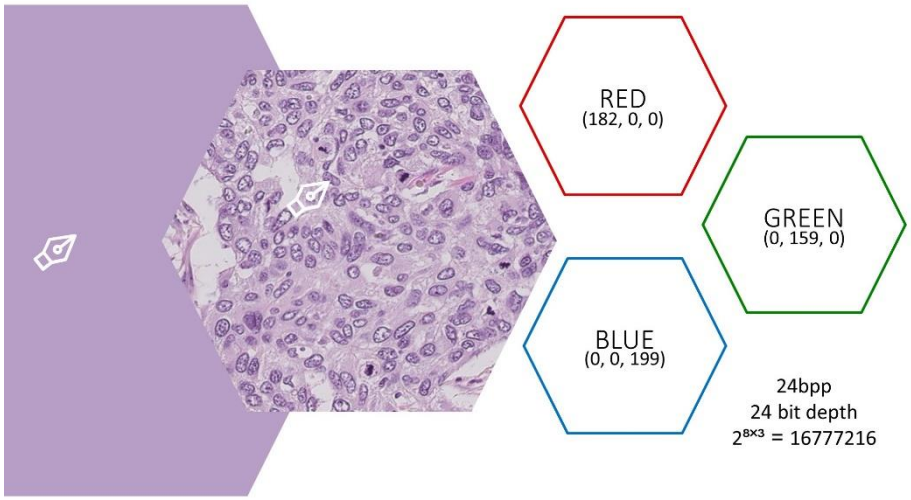


Figure 2: **Bit depth in an RGB image**. A specific pixel (arrow) has a unique value from 0-255 in each colour space. An image with a bit depth of 24 has 16,777,216 possible combinations.

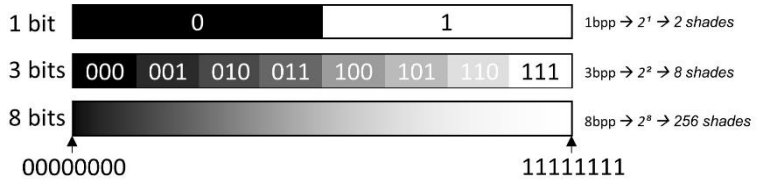


Figure 3: **Bits and bit depth** explained in greyscale.

1.2.2 Image Processing

Image processing refers to the use of algorithms to manipulate the attributes of an image for certain tasks. These tasks include, amongst others, classification, feature extraction and segmentation. Filters are tools used in imaging processing tasks such as edge detection or noise removal. Examples of such filters include mean or median filters often used for segmentation tasks.

1.3 Artificial Intelligence

1.3.1 Background

Artificial intelligence (AI), a term coined by John McCarthy in 1955, has become a current topic of interest and debate in the media. Defined as a computer's ability to mimic human intelligence or behaviour, AI encompasses the disciplines machine learning which in turn covers deep learning (Figure 4). In machine learning (ML), algorithms are developed to perform specific tasks without being explicitly programmed to do so[30]. A ML algorithm is designed to learn from a dataset, extracting information from that data whilst also learning the context of the data in order to make decisions or define outcomes[31].

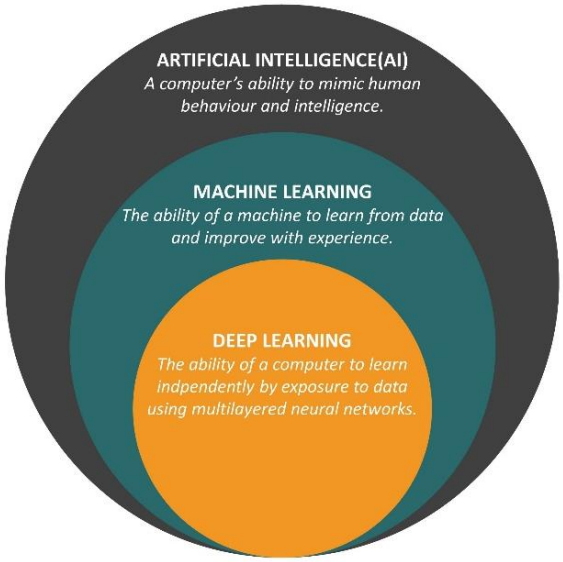


Figure 4: **Artificial intelligence definitions**: AI, Machine Learning, and Deep Learning.

ML approaches vary from strongly supervised learning to unsupervised learning. The type of learning method describes whether labelled or unlabelled data is used to train the algorithm. A summary of the types of learning can be found in Table 1. Both supervised and unsupervised learning were used in the works presented in this thesis.

Deep learning (DL) relies on multi-layered neural networks. Neural networks are paralleled to the neuronal network of the human brain, whereby an intricate layer of neurons is activated by a signal, which activates downstream layers resulting in an output. One of the most popularised networks in the field is the convolutional neural network (CNN), which will be discussed later in this section. Deep learning (DL) has made notable progress in recent years due to advancements in computer hardware, such as the graphics processor unit (GPU), breakthroughs in research on neural network architectures, and the acknowledgement that large datasets are required for DL training tasks[36]. As a result, it has been increasingly used for problem-solving tasks in medicine.

Table 1: Types of Machine Learning (ML) Methods used to Train ML models, as defined by the International Medical Device Regulators Forum[32]

Type of Learning	**Definition**
Strongly supervised	During training, an algorithm will learn the relationship between independent features and a known defined attribute (the label). This is considered the most time-consuming approach due to the requirement of large amounts of annotated data.
Semi-supervised	This approach uses both labelled and unlabelled data to train. "It is a hybrid technique between supervised and unsupervised learning"[33]. This is a less label-intensive method than strongly supervised learning. The algorithm will use traditional supervised learning to train the model, whilst for unlabelled data it will minimise the difference the predictions made between other training examples[33].
Unsupervised	The unsupervised ML approach is often used to handle more complex tasks than the other approaches. What distinguishes it from the other approaches is its use of unlabelled data. Clustering is a type of unsupervised ML where a model will find patterns in the data and natural clusters[34]. Association is also another method, where the association between variables can be used to map them in a way that is useful[34]. Unsupervised learning is the least labour-intensive approach as it does not require annotations; however, it does require large, complete datasets that are representative of the context.
Reinforcement	Another type of ML approach is reinforcement learning. As its name suggests the algorithm will learn and adapt according to feedback or to maximise the return/reward. This process can be described as trial and error[35].

1.3.2 Machine Learning

Several examples of ML methods can be found in the medical literature. Linear and logistic regression, familiar to those working in the field of statistics, are popular choices. Other types of supervised machine-learning methods include support vector machines (SVMs), Naïve Bayes, K-nearest neighbour (kNN), decision forest and random forest. Each of these methods has their own unique advantages and disadvantages. The methods to be discussed in this thesis are K-means clustering and deep learning networks that use a convolutional neural network (CNN).

1.3.3 K-Means Clustering

K-means clustering is an unsupervised ML method. Clustering is a method of unsupervised learning where the model finds patterns and natural clusters in data[34]. K-means will aggregate collections of data points according to similarities. This clustering will generate centroids according to the mean of the datapoints. Each centroid will represent a class and the number of centroids is determined by k [37] (Figure 5). A new datapoint will be compared to these centroids and a class assigned based on the centroid with the closest proximity.

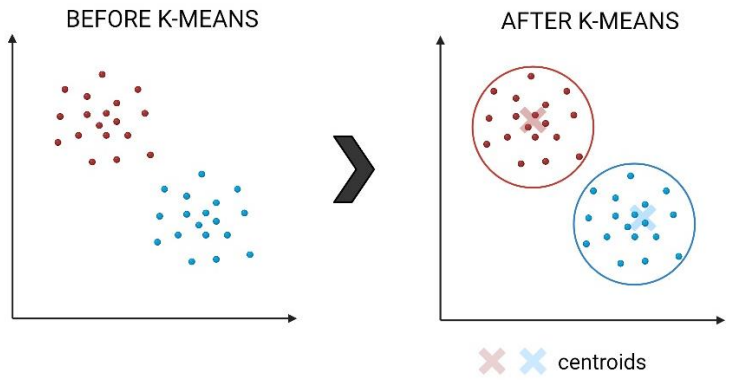


Figure 5: **K-means Clustering.**

1.3.4 Deep Learning: Neural Networks

1.3.4.1 Artificial Neural Network

An *artificial neural network* (ANN) is one of the simplest forms of a neural network. An ANN consists of artificial neurons, modelled after the biological neuron (Figure 6). An artificial neuron consists of inputs (x_i) and weights (w_i)[36]. Each input will be weighted and the *weight* indicates the strength of the connection between the nodes[38]. The activation function will calculate the sum of the weights and add the bias, if these exceed a specified threshold, then the node is activated and a signal passes on to the next layer[39] (Figure 6).

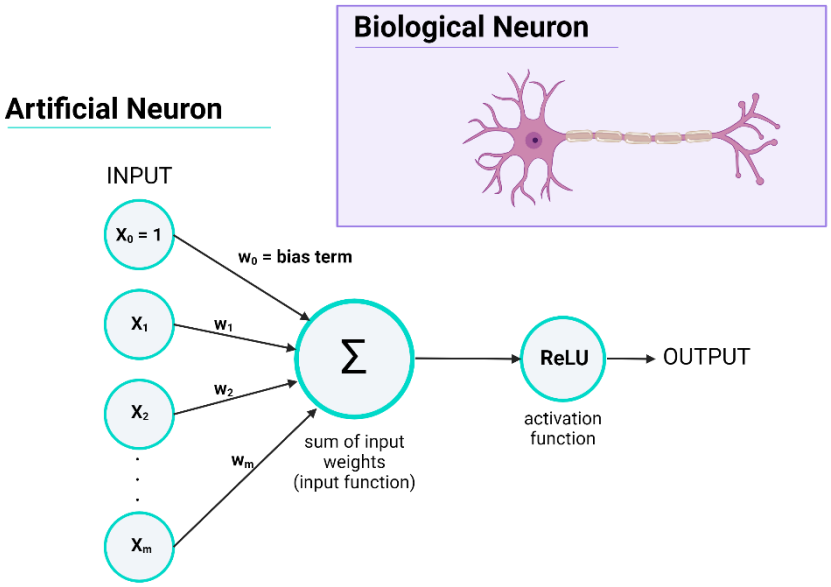


Figure 6: **Example of an artificial neuron**[36]. *Created with BioRender.com*

In an ANN model, multiple neurons are linked together in layers (Figure 7). The first layer is the input layer, followed by a hidden layer. Each

layer consists of *n* number of nodes. In *a fully connected network*, each node will have a connection to all the nodes in the previous layer and all the nodes in the next layer[39]. The strength of an ANN comes from many nodes working in parallel[40]. The final layer of the ANN is the output layer, where the network generates an outcome/prediction. The weights of each neuron are iteratively adjusted to increase the accuracy of the outcome prediction.

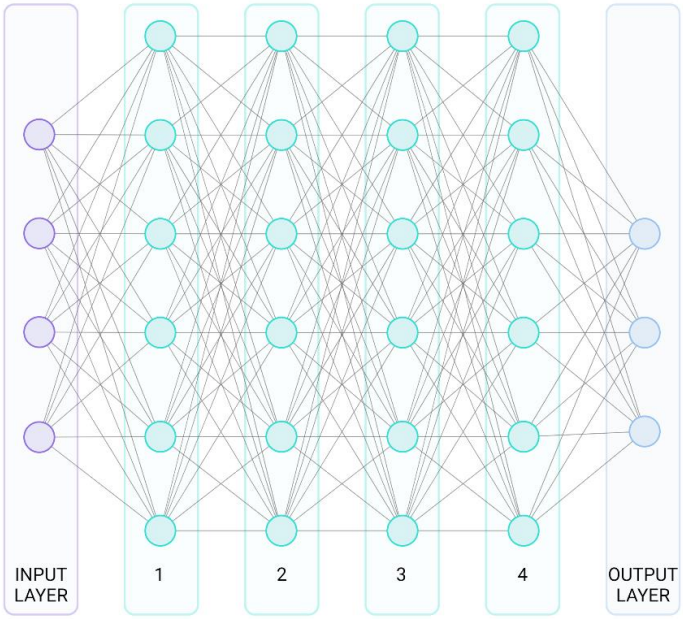


Figure 7: **A multi-layer neural network**. A multilayer perceptron (MLP) is a type of feed-forward ANN consisting of fully connected neurons, an input layer, hidden layers (1-4 in this example), and an output layer. *Created with BioRender.com*

The training of an ANN begins with random weights[36]. Training of a neural network is an iterative process where these weights are optimised. This process requires the calculation of the error using a loss function. For example, in a feed-forward network, the loss function generates an error from a forward pass through the network[40]. The

error is obtained by calculating the difference between the expected outcome and the model prediction[41]. The weights are then adjusted to minimise this error. *Backpropagation* is a commonly used optimisation algorithm that aims to reduce the loss function by calculating the gradient of error (direction and magnitude) that is used to adjust the network weights[42]. It requires that the error rate produced from the forward pass of the network is then fed back through the neurons of the network layers from back to front. A technique referred to as *gradient descent* is used to find the most optimal adjustments that minimises the error as much as possible[43]. The *learning rate* describes how large of a change is made to the weights, referred to as step size[43]. If the learning rate is too high, the changes might be extreme, and the minimum error overshot (Figure 8). If the learning rate is too small, the changes to the error are inconsequential and it take a long time before the minimum error rate is achieved[44]. The learning rate of a neural network can be manually adjusted during the training process and the resulting error rate monitored after each iteration (Figure 8).

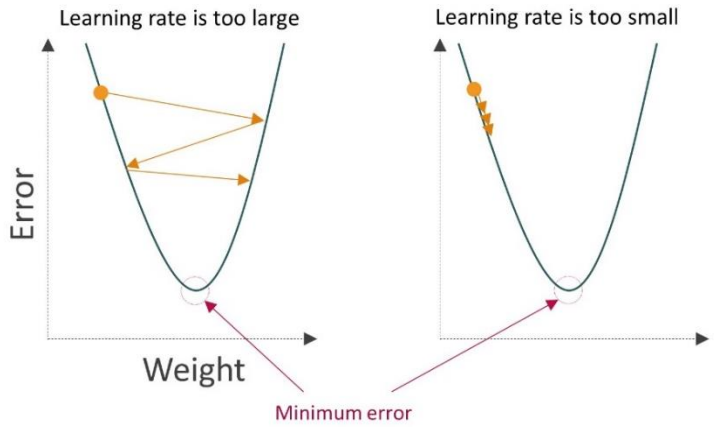


Figure 8: **The learning rate of a neural network**. The learning rate regulates the level of change to the network weights. A learning rate that is too large will oscillate and overshoot the minimum error. If the learning rate is too small it will take a long time before reaching the minimum error[43,44].

1.3.4.2 Convolutional Neural Network

A convolutional neural network (CNN) is a type of neural network that consists of three main types of layers: a convolutional layer, a pooling layer, and a fully connected layer[39]. The convolutional layer is what makes a CNN unique. Its ability to analyse an image using specific filters for feature recognition have made it an attractive model for object detection and classification tasks on pathological images[45]. In each convolutional layer, the input is convolved with a set of filters (kernels) which produces feature maps[39]. Different filters will generate different feature maps and capture different features within the same image. The second type of layer is a pooling layer and results in downsampling of the feature maps. The purpose of this layer is to reduce the computational power required by downsizing the feature map but retaining the essential information[46]. This is performed for example by max pooling or average pooling. Prior to the final layer, the 2D arrays from the pooled layer are 'flattened' into a single continuous linear vector. This is then fed into the final layer: the fully connected layer. The final layer resembles a feed forward MLP, with input features, N-hidden layers, and an output layer. Prior to the output layer, the softmax activation function will normalise values by classifying inputs into a probability from 0 to 1 for each prediction outcome[46]. Popular examples of neural network architectures, that use a CNN, are U-Net, R-CNN, DeepLabv3, VGG, GoogLeNet, and ResNet[45]. The U-Net, DeepLabv3 and R-CNN models were utilised in two of the studies presented in this thesis.

1.3.5 AI in Medicine

Although the term AI originated in the 1950's, shortly followed by the term machine learning by Arthur Samuel and later deep learning in 1986 by Rina Dechter, it would take many years before the medical field

would take an active interest[47]. In the 1970's and 80's, AI was slated to take a prominent role in medicine in the near future, acting as "a powerful extension of the physician's intellect"[48]. Early algorithms relied heavily on rule-based systems; specific patterns and situation dependent features could be organised by a machine to make logical conclusions[48]. The original conviction began to weaken as researchers realised medical patterns were not so simple and the technology not yet sufficient to handle them. The intricacies of medical reasoning revealed that diagnosis is too complex a task to develop a single set of rules to cover all variables.

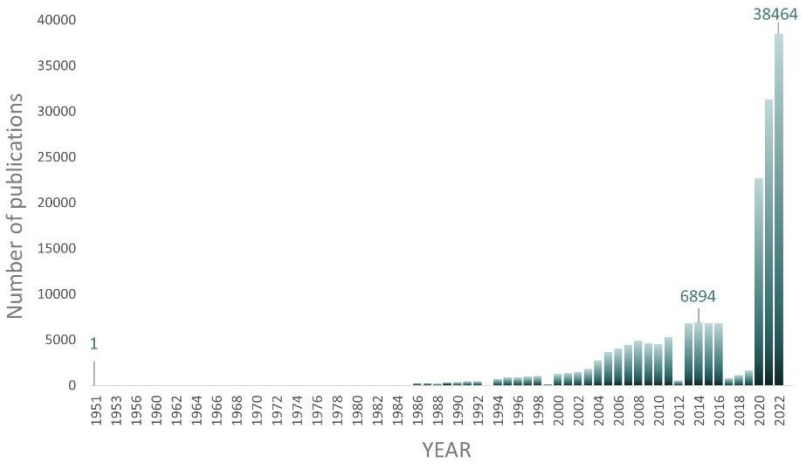


Figure 9: **AI Publications on PubMed**. The number of publications per year up until 2022 using the search query: "artificial intelligence" on PubMed.

At the turn of the millennium, technology was making significant advances. Radiology was well into the transition from analogue film to digital radiography, additionally the first whole slide scanners were introduced allowing for the digitisation of pathology slides[47,49]. The last two decades has witnessed a rise in the number of public repositories, from the sequencing of the human genome to the establishment of public image databases such as the TCGA-BRCA, CAMELYON 16 & 17

breast cancer WSI datasets and many more[50-52]. The prominence of AI in medical research may be visualised in the context of the number of publications registered on PubMed from the 1950's up until present day (Figure 9).

One of the major obstacles developers have faced in the past is the lack of adequately labelled datasets. The performance of deep learning models is dependent on the availability of such data to create sufficient training and validation datasets[53]. These datasets should be accompanied by complete and accurate documentation. If labels are missing or incorrect then a model will be incorrectly trained.

Private datasets may also be used to train AI models and may be combined with public datasets to improve diversity. Private datasets usually contain more information because of stricter confidentiality agreements and therefore, may contain treatment and survival data, which can provide a more accurate insight into how the AI model is performing[53].

Validation of AI tools for use in a clinical setting is often performed in the form of clinical trials. Many such trials compare the performance and accuracy of diagnosis with and without the use of AI tools. AIMS Norway is a national, randomised controlled trial investigating interpretation of mammography scans by radiologist(s) with or without the assistance of the AI tool Transpara[54]. CONFIDENT-B, -P are controlled clinical trials performed at the University Medical Centre Utrecht[55]. Prostate cancer specimens (-P) and breast cancer specimens (-B) will be assessed by pathologists with or without an AI assistance tool[55]. However, before any implementation it is important to

understand the development process of an AI tool, why it was developed and how and what data was used to train and validate it.

1.3.6 Developing an AI tool

Prior to development of an AI tool for medical purposes, it should be ascertained if AI is indeed the correct choice for tackling the problem at hand[56]. If a task is considered a burden and is repetitive in nature, AI may be the appropriate choice. The National Health Service Transformation Directorate (NHS^{X}) in the United Kingdom define several criteria that should be met if an AI tool is an appropriate solution for a medical task:

1. *The data is repetitive in nature or on such a large-scale that there is an above-average risk for human error during processing.*
2. *Any outcomes can be tested against a ground truth*
3. *Outcomes should have real-world implications, i.e. will have implications on patient treatment, workflow efficiency, etc.*
4. *The data is real and not synthetic*
5. *Using the data is ethical and safe*

A Buyer's Guide to AI in Healthcare, 2020[56]

1.3.6.1 Clinical demand

The purpose of the AI tool needs to be clearly defined and there needs to be a clinical demand or benefit that the tool must meet. The context in which the AI tool is to be used should also be clarified early on, as this will affect the type of training data that should be included in development. Lastly, there needs to be a measurable benefit to using the tool.

1.3.6.2 Ground truth

The purpose of using AI may be with the intent to improve on existing methodologies and workflows. To measure this, the outcome must be compared to a ground truth or reference standard. Ground truth is defined as information that is known to be real, often in the form of empirical evidence or observations[57].

However, when defining a ground truth, a problem arises when it is generated by a method with intrinsic weaknesses. It is likely that the ground truth in pathology will be a pathologist as there is often not any other comparable reference standard. A pathologist's diagnosis, though expertly trained, is based on a subjective visual assessment of the tissue and therefore the ground truth falls victim to their inherent irreproducibility[58]. Factors influencing intra- and inter-observer variability include the type of training received, years of experience, recall bias and fatigue. It is by no means a poor ground truth, but its inherent imperfections should not be ignored, and its limitations expressly considered. A way to combat this is to use an expert panel. A predefined, minimum consensus agreement amongst a panel of pathologists, preferably from different laboratories can be required to affirm a ground truth[8,58]. Other ground truths may be worth investigation such as molecular testing or even survival data or data on treatment response.

1.3.6.3 Datasets

The Development Dataset: Training & Tuning

The data used to train, tune, and validate an AI algorithm is perhaps one of the most important considerations during study design. Datasets to be covered are the development and test dataset. The definitions from

Liu et al. (2019) will be used here[59]. The development dataset may consist of a training and tuning set, often only a training set.

The training set consists of cases used to train the model. The input data may consist of unlabelled images or annotated images. The tuning set will be used to adjust the hyperparameters such as the learning rate. The tuning set is used to optimise the model selection[32].

To produce the best possible AI model, the development dataset should be representative of the clinical setting. It should include unique challenges such as rare subtypes, artefacts, and diagnostic mimics. The proportion of cases should reflect the real-world scenario. Considerations should be made for inclusions of images scanned by different scanners and tissue stained with different antibodies, using different automated staining instruments for the detection of the same markers. A model will only ever be as good as the data you put in (Figure 10).

Figure 10: **Garbage in → Garbage out**. An AI model will only ever be as good as its training data. If the training data is of poor quality and unrepresentative of the task the result will reflect this. *Figure created using biorender.com*.

Transparency is a stepping-stone to credibility. It is important to describe which dataset is being used whether it be a public or private dataset and to acknowledge any limitations and how the data has been annotated. Case studies have shown that an algorithm will learn specific

patterns in images and assigns an outcome, but in some cases further examination reveals that these patterns are not in fact related to the condition at all and are essentially artefacts. Radiologist Lauren Oakden-Rayner questioned the validity of the ChestXray14 dataset, in 2017, to train a model to detect pneumothorax[60]. She observed that a model was performing acceptably, however, further examination of the images, revealed that all the patients classified with a pneumothorax by the algorithm, had chest drains. This meant the patient had already been treated for pneumothorax. Therefore, the AI had simply learnt to detect the chest drain and not the pneumothorax, which "isn't a medically important problem. We want to avoid missed pneumothoraxes, and by definition these have not been missed"[60]. This emphasises the need for explainable AI, to ascertain what is being detected. Lauren's observation also raises the potential pitfalls of using open datasets to train medical AI systems.

The Test Set

The test set is sometimes referred to as the validation set. The test set is used to evaluate model performance before it can be applied in the clinic. It will test that the model performs as intended and uncover any overfitting or bias in the model[61]. The model can only be run on the test set once. It should also be representative of the clinical setting[61]. Confusion arises when this term is mistakenly used to describe the tuning set. The most important distinction between the development dataset and the test set is that they are independent of each other[61]. The algorithm should not have been exposed to any of the data in the test set during training. Ideally, the data used in the test set should also be temporally and geographically distinct from the data used in the development dataset[61]. This will evaluate the robustness of the model and its clinical applicability. Furthermore, the test set can confirm any limitations of the model, for example: certain scenarios such as poor

staining or inadequate fixation where the algorithm will not perform accurately. Such limitations must be determined as either "catastrophic", the algorithm is not approved; or "acceptable", the algorithm is approved with recognition of its limitations.

1.3.6.4 Overfitting & Bias

Overfitting describes when a model is trained so perfectly on the training dataset that it cannot be applied to any other dataset[59]. It has therefore learnt parameters specific to the training dataset that have no relationship with clinical outcome[59]. Techniques exist for avoiding this such as early stopping, where training is stopped early enough that the algorithm reaches an acceptable error rate before becoming too fitted to the training data (Figure 11).

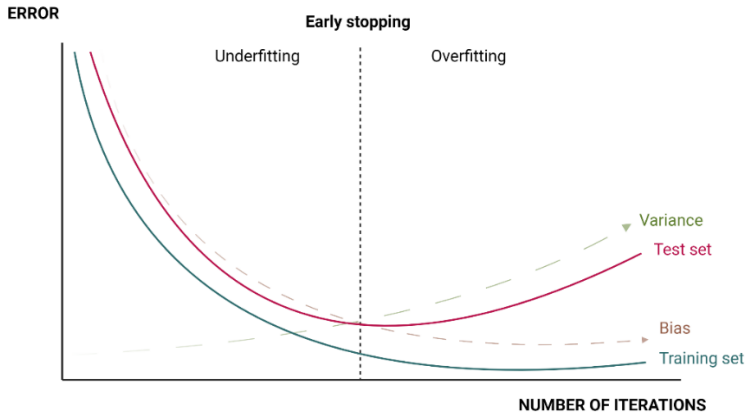


Figure 11: **Early stopping and the bias-variance trade-off**. *Figure created using BioRender.com*.

Bias refers to a prejudice resulting in favour of or against a group or person[62]. In ML bias occurs when a model is prejudiced due to faulty datapoints[63]. Variance, is a type of error which describes a model's sensitivity to changes in the features used to train the model, thereby if overfit, the model is very sensitive to small details in the relationship

between variables[63]. In ML, to achieve the most optimal model, both bias and variance should be low[63] (Figure 11).

Bias in medicine is usually implicit: not intentional or with ill intent[64]. It is usually introduced by unforeseen differences between groups, either due to poor study design or the assumption that there are no differences. In medicine, women and minorities have been historically under-represented[65-68]. Research over time has exposed a multitude of variables that are highly relevant to clinical outcomes including biological sex, socioeconomic status, ethnic background, etc. One should strive to account for and acknowledge this, where applicable, in datasets used to train and/or test AI algorithms[35]. Systemic bias occurs when a model will consistently over- or under-estimate an outcome as a result of training[35]. It often refers to how the data was collected.

Several forms of systemic bias exist including selection bias, gender bias, measurement bias, and racial bias. Selection bias can occur when the target population is not accurately represented in the training data whilst measurement bias is a result of an error made during collection[69]. Selection bias is particularly important to be aware of when using open datasets. For example, the TCGA dataset, consisting of 1133 H&E stained WSIs from 1062 breast cancer patients, was collected from patients solely in the USA, majority white women and a large percentage younger women[52,53]. This may not be representative of the US population, let alone in another demographic. One should always be aware of potential bias by critically assessing inclusion criteria, when using open datasets.

> *"The introduction of AI to diagnostics seems to be accompanied by little to no acknowledgement of the well-documented and chronic gaps in medical data when it comes to women. And this could be a disaster… particularly given what we know about machine-learning amplifying already-existing biases."*[65] – Caroline Criado Perez, 2019

Gender bias, disparate inclusion, or lack of necessary representation of sexes in training datasets, can result in certain sex-specific patterns being overlooked or overshadowed. In many clinical evaluation studies and even animal studies, one sex is disregarded entirely or inadequately represented[65]. This can result in an underestimation of outcomes in one sex and even life-threatening consequences[65]. For example, an algorithm trained on a dataset overrepresented by male patients, may lead to accurate detection of male-specific patterns of a disease but those more frequently manifested in females are not. Therefore, women risk being underdiagnosed. Gender differences in disease has been reported in the literature for cardiovascular disease[70,71], Parkinson's disease[72] and cancer[73]. Therefore, it is important to consider potential differences between genders and account for them where necessary.

Racial bias may occur when an algorithm is not adequately trained on data collated from multiple ethnicities. For example, in dermatology concerns have been raised regarding ML algorithms developed to detect melanoma. Melanoma will present differently in white skin types than darker skin types. Therefore, if one or the other is not adequately represented in the training data this will limit the algorithm to accurately detect melanoma in a range of skin colours[74].

A well-collated test set may be a method of uncovering potential biases in the training of the AI. An adequate sample size for training AI, in addition to balanced datasets, can also help to reduce bias. Effectively, AI technology has the ability to mitigate systemic bias if designed properly, but may also magnify existing disparities if a bias is not corrected for[75]. To summarise, it is important to address the need for equitable AI and to express any limitations of datasets.

1.3.6.5 Performance metrics

To assess the performance, quality, and selection of the most optimal model, statistical performance metrics are measured. Two of the most common methods used are a receiver operating characteristics curve (ROC curve) and a confusion matrix.

A ROC curve is used to determine the efficacy of a model by measuring the area under the curve (AUC). An AUC above 0.8 is preferred, with 1.0 indicating perfect prediction, and less than 0.5 indicates an inadequate model[76]. The ROC curve itself is a graphical representation of the true positive rate (sensitivity) and the false positive rate (1-specificity)[76]. This metric can be used to determine optimal cut-off points often where there are binary outcomes, such as disease is present or absent or low- and high-risk of progression of disease.

The confusion matrix is a contingency matrix that describes the performance of an algorithm, by assessing predicted values against actual values[77]. A variety of metrics can be calculated from the confusion matrix including sensitivity (recall), specificity, positive predictive value (PPV or precision), negative predictive value (NPV), F1 score and accuracy (Table 2). Each of these metrics is used to evaluate the predictive ability of the test or model as defined in Table 2.

A perfect test will have 100% sensitivity and 100% specificity or a PPV of 1 and NPV of 1. However, in a realistic clinical setting, this is extremely unlikely. There is often a trade-off where an increase in sensitivity may lead to a decrease in specificity. Therefore, when designing a test, it is important to consider if it should be more sensitive or specific. This will depend on the consequence of a false positive or false negative result. For example, a **highly sensitive test** will correctly identify those with a condition i.e., few false negatives. This will ensure the ability **to rule out disease** if the test is negative[78] (Figure 12). A

highly specific test will correctly identify individuals who do not have the condition i.e., few false positives. This will ensure the ability to **rule in disease** if the test is positive[78] (Figure 12). The mnemonic SNOUT and SPIN can be useful in remembering this difference[78]:

SNOUT: Sensitivity, negative, out
SPIN: Specificity, positive, in

Table 2: Overview of performance metrics used to evaluate the efficacy and quality of a test[77].

Performance Metric	Definition
True positive (TP)	The number of cases scored *positive* by the test that *have* the condition.
True negative (TN)	The number of cases scored *negative* by the test that *do not have* the condition.
False positive (FP)	The number of cases scored *positive* by the test but *do not have* the condition.
False negative (FN)	The number of cases scored *negative* by the test but *do have* the condition.
Positive predictive value (PPV)	The PPV describes the odds that an individual with a *positive result* actually *has the disease*/condition.
Negative predictive value (NPV)	The NPV describes the odds that an individual with a *negative result* does *not have the disease*/condition.
Sensitivity	The sensitivity of a test is an indication of the proportion of individuals that have a *positive* result among those that *have* the condition.
Specificity	The specificity of a test indicates that the proportion of individuals with a *negative* result among those that *do not have* the condition.
Prevalence	The number of cases in a defined population with a condition at a single point in time.
F1 Score	*A measure of a test's accuracy* based on the PPV (precision) and sensitivity (recall) of a test. The F1 score ranges from 0 to 1, with 1 indicating perfect precision and recall.
Accuracy	Accuracy is a measure of a test's ability to distinguish between individuals with or without a condition – *how precise is the test*

Ideally, a test with a high specificity, or high PPV, is recommended for tests where a positive result carries the risk of harm from subsequent treatment or diagnosis i.e. surgery or a terminal diagnosis. A test with a high sensitivity, or high NPV, is adequate for screening, where tests are usually minimally invasive and cheaper. For example, an HPV test is used as a screening tool for cervical cancer. The test itself is minimally invasive, and if positive will lead to further examination and/or a biopsy. A test with high specificity is adequate for a diagnostic test.

It is also important to note that PPV and NPV define a slightly different context to sensitivity and specificity. The latter is used to indicate the credibility of a test in *comparison to a trusted reference standard*. Whilst the former indicates the probability of the test to successfully identify individuals that *do or do not have a condition according to the test results*[78]. In summary, sensitivity, specificity, and predictive values may be used to describe the usefulness and adequacy of a model or test, however, a clear description of how they have been determined should be stated. Furthermore, no test is perfect, and therefore the consequences of under- or over-treatment should be evaluated to determine and adjust the weight of different performance metrics accordingly.

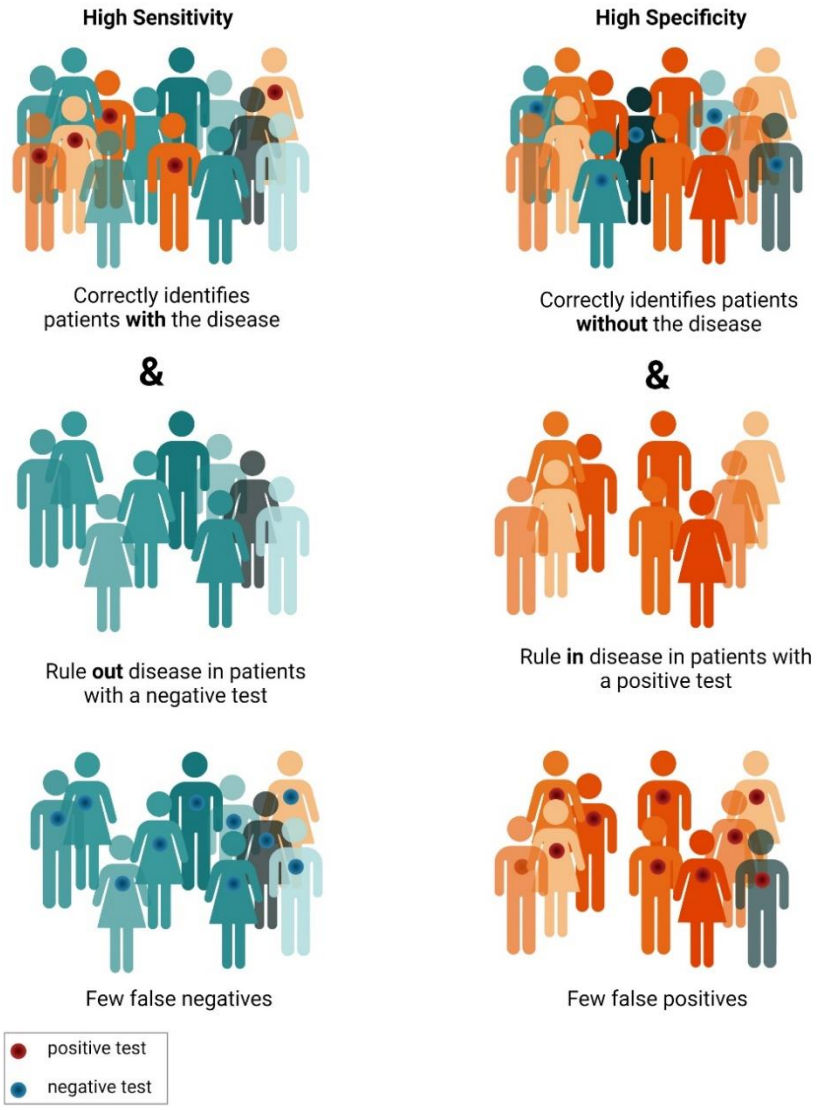


Figure 12: **Sensitivity and Specificity**. A test with a high sensitivity will correctly identify patients with a disease and is used to rule out patients with a negative test i.e. few false negatives. A test with a high specificity correctly identifies patients without the disease and is used to rule in patients with a positive test i.e. few false positives. Blue individuals indicates without the condition, orange indicates individuals with the condition. *Created with BioRender.com*

1.4 Cancer

> *"Cancer is a flaw in our growth, but this flaw is deeply entrenched in ourselves. We can rid ourselves of cancer, then, only as much as we can rid ourselves of the processes in our physiology that depend on growth – aging, regeneration, healing, reproduction"*[2] – Siddhartha Mukherjee, 2010

1.4.1 The Hallmarks of Cancer

Cancer has long been present in the history of humankind. It's designation as a modern disease is likely owed to the fact that as humans begin to live longer, civilisation and modern medicine has "unveiled it"[2,79].

As described by Virchow in 1859, all cells arise from other cells. Cancer arises from the normal cells in the tissue, from which it is discovered[80]. Cancer is the uncontrolled growth and proliferation of cells; a mass of cells described as a tumour in solid cancers[81]. This is the result of several biological capabilities that are acquired during a multi-step process, which are referred to as the hallmarks of cancer (Figure 13)[82]. These capabilities also contribute to a cancer's ability to spread throughout the body to establish new growths (metastases) in a process known as metastatic dissemination[81]. From the original six hallmarks described in 2000 by Weinberg and Hanahan[83], two functional capabilities were added alongside two enabling characteristics in 2011[82], and in 2022[84] a further four were added (Figure 13).

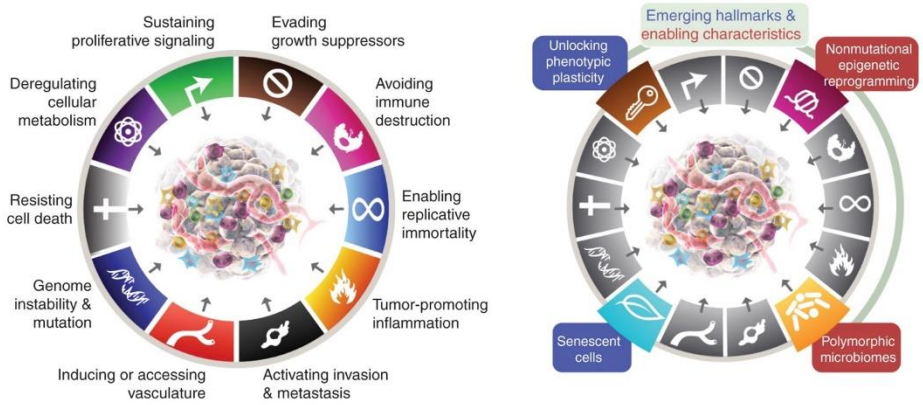


Figure 13: **The Hallmarks of Cancer** (left) and the newly proposed emerging hallmarks (right). Reprinted from Cancer Discovery, 2022, 12 (1), 31-46, Hanahan, Hallmarks of Cancer: New Dimensions, with permission from AACR.

Perhaps the most fundamental characteristic of cancer cells is their ability to proliferate unrestricted[82]. In normal tissue, the growth of cells is carefully regulated with the controlled release of growth-promoting signals initiating entry into the cell cycle[82]. This ensures a balance in the number of cells present in the tissue at any one time[82]. Cancer cells gain the ability to manipulate such signals and initiate cell proliferation unencumbered by normal checkpoint regulators. This capability often coincides with changes to mechanisms in cell survival and energy metabolism[82]. To sustain proliferation, cancer cells must also avoid growth suppressors. These suppressors are either encoded or dependent on tumour suppressor genes. such as RB (retinoblastoma-associated) and TP53 proteins. These proteins play a key role in systems that initiate proliferation or activate senescence and apoptosis[85].

As affirmed by Hanahan and Weinberg, the complexities of cancer cannot be understated[82]. These hallmarks provide researchers with the tools to better understand cancer; its pathology and to generate new targeted treatments.

1.4.2 Biomarkers

A biomarker is defined as a "biological molecule found in blood, other bodily fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease"[86]. Biomarkers are commonly used for diagnosis, monitoring, prognosis, prediction of response to treatment, susceptibility, pharmacodynamics, and assessing of toxicity of exposure to a medical product[87]. A variety of classes exist and are outlined in the BEST (Biomarkers, EndpointS, and other Tools) Resource published by the FDA-NIH Biomarker Working Group[87], as summarised in Table 3.

Biomarkers may exist in the form of proteins, molecular profiles, a biological state, or hormones but most importantly, they can be measured[87]. Biomarkers are a cornerstone of precision medicine as research shows that "no two patient's cancers are exactly the same"[3]. Despite sharing many general similarities, different cancers behave and evolve in different ways, possess different genetic profiles, and have variable responses to treatment[3]. Precision medicine is designed to optimise patient treatment by assigning treatment strategies according to specific traits to maximise therapeutic benefit (Figure 14). This is in contrast to the current one treatment fits all approach[3].

> *"that's the promise of precision medicine -- delivering the right treatments, at the right time, every time to the right person. And for a small but growing number of patients, that future is already here"*[88]
> - Former President Barack Obama, 2015

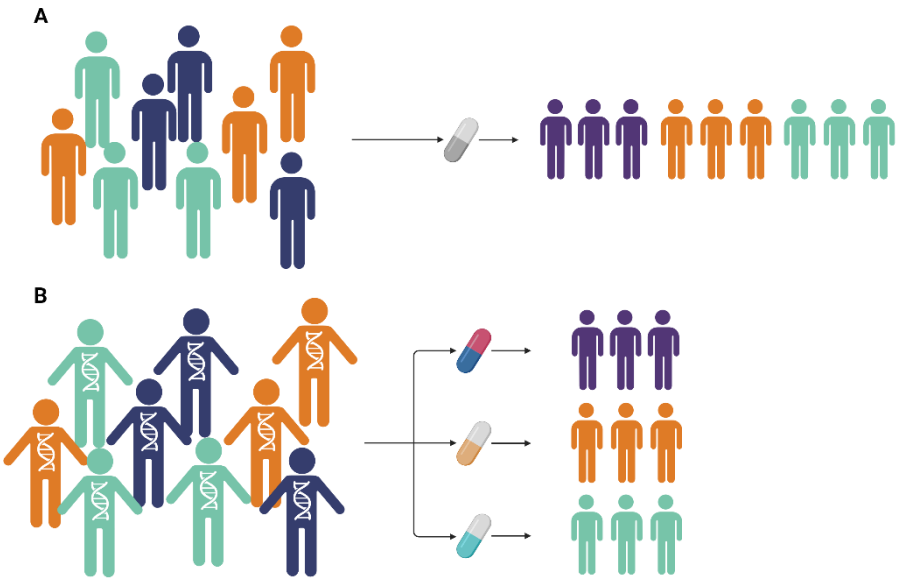


Figure 14: **Precision medicine**. A) A one treatment fits all approach. B) A precision medicine approach. *Created with BioRender.com.*

To evaluate the clinical performance of a biomarker it is compared to a reference standard (ground truth). Typically, calculation of a test's performance metrics (Table 2) indicates the efficacy of the biomarker. A perfect diagnostic biomarker has 100% sensitivity and 100% specificity, but this is unlikely in a real-world setting. Additionally, a biomarker test, like an AI model, should have defined conditions of use[87]. For example, a patient must be in a fasting state prior to sampling blood, or a biomarker only demonstrates benefit in certain subtypes of a disease, like Ki67 measurement is only recommended in luminal-like breast cancer[89].

Table 3: Classes of Biomarkers as defined by the FDA-NIH Biomarker Working Group[87]

Class	Definition	Example
Diagnostic	To confirm the presence or absence of a disease/condition	C-reactive protein (CRP) –a measure of inflammation, indication of infection.[90]
Monitoring	A biomarker is measured at defined intervals to assess the status of a condition/disease in response to treatment or exposure to an environmental agent	HIV RNA – can be measured to assess the effectiveness of anti-HIV therapies.[91]
Prognostic	To predict the likelihood of a clinical event e.g. disease recurrence, progression or mortality	Ki67 – a marker of cell proliferation commonly used in breast cancer to assess outcome[89,92]
Predictive	To evaluate if an individual has a greater response to a treatment or agent in comparison to others.	HER2 – breast cancer patients with HER2 protein overexpression may benefit from treatment with trastuzumab[89,92]
Susceptibility /Risk	A biomarker that indicates if an individual is more likely to develop a disease or condition, who does not yet have the disease or condition.	BRCA – A mutation in the BRCA1/2 gene has been associated with increased risk of developing breast cancer.[89]
Pharmacodynamic /Response	To show that a biological response is present in an individual following exposure to a treatment or environmental agent. This class is often used in clinical trials.	Blood pressure reduction – indication of a desired response to anti-hypertension treatments[93].
Safety	A biomarker that is measured before and after an individual has been exposed to a treatment or agent. This is usually to assess the toxicity of the exposure.	Neutrophil count – to assess the cytotoxcitiy of chemotherapy in order to determine correct dosage or intervene.[94,95]

1.4.3 Cancer in Norway

The Cancer Registry of Norway has systematically collated information on cancer occurrence in the Norwegian population since 1953[96]. The Norwegian population has since grown by 64%, with 5 488 984 inhabitants recorded as of 1 January 2023[96]. Furthermore, the population is aging, with 13% of the population in 2022 over 70 years of age and predicted to rise to 22% in 2060[96]. Most cancers in Norway are diagnosed in persons over 50 years of age (90% for males, 86% for females) and 55% (49%) of new cancers are diagnosed in males over 70 (females over 70)[96]. An aging population is not only associated with an increased incidence and risk of cancer but is also set to impact healthcare in other ways. This includes increases in cost of treatment and long-term management, shortage of qualified workers and an increasing dependency ratio. Therefore, prevention, early detection and targeted therapies are key in the fight against cancer, particularly in the face of an aging population.

In 2022, 38 265 new cases of cancer, in 37 277 individuals, were recorded in Norway[96]. The four most common forms of cancer were: prostate, female breast, lung and colon cancer (Figure 15)[96]. Incidence rates for most forms of cancer have increased, with the exception of stomach cancer and, until recently, cervical cancer[96]. The cumulative risk for prostate cancer in males (16.1%) and breast cancer in females (10.7%) is highest of all cancer types[96]. One in ten Norwegian females is expected to be diagnosed with breast cancer before their 80th birthday and four in ten Norwegians are expected to be diagnosed with a cancer before the age of 80[96].

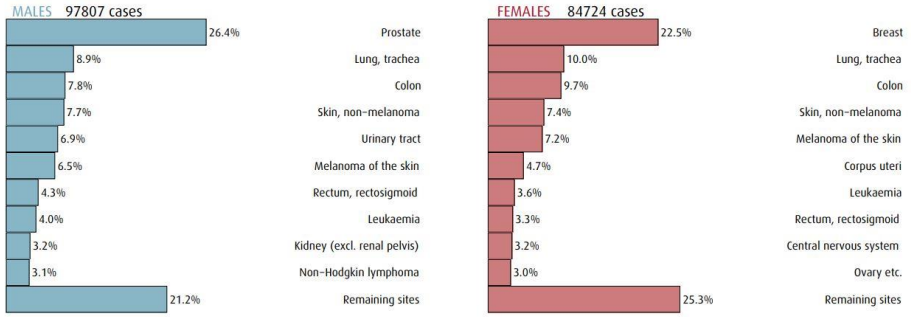


Figure 15: **Cancer in Norway**. The most frequent types of cancer by age and sex, 2018-2022 (all ages)[96]. Reproduced with permission from the Cancer Registry in Norway.

1.5 Breast Cancer

1.5.1 Epidemiology

In 2020, the GLOBOCAN global estimate for the number of new cases of female breast cancer was 2 261 419[97]. This was the highest of all cancer types recorded[97]. It is also the leading cause of cancer death in women and the fifth leading cause of cancer deaths when disregarding sex, with a reported 684 996 cancer deaths in 2020[97]. The highest incidence rates have been observed in developed countries such as Belgium and Australia[97]. However, mortality rates are higher in developing countries in comparison to developed (15.0 vs. 12.8 per 100,000, respectively)[97]. Contrast in incidence rates reflect differences in reproductive trends and lifestyles factors e.g. fewer number of children, advanced maternal age at first birth and excess body weight are more commonly observed in developed countries than developing[97]. The difference in mortality rates, may be attributable to diagnosis of breast cancer at a later stage in developing countries perhaps due to more limited access to healthcare. Several risk factors for breast cancer are reported (Table 4).

Table 4: Established risk factors for breast cancer.

Associated risk factors
Lack of exercise[98]
Obesity[99]
Alcohol consumption[100]
Smoking[100]
Lower parity[101]
Reduced length of lactation[101]
Late age for first birth[102]
Early menarche[103]
Late menopause[103]
Family history of breast cancer[104]
BRCA 1/2 mutations[105]

In Norway, 4 247 new cases were reported in 2022, and breast cancer was the third leading cause of cancer-related deaths in women[96]. Although, 5-year survival rates have increased in recent decades (89.3% in 2009-2013, 90.7% in 2014-2018), the incidence is still rising[96,106].

1.5.2 Anatomy of the breast

The breast of an adult woman is primarily composed of glandular tissue, fibrous connective tissue and fatty tissue[107] (Figure 16). The connective tissue also includes supportive ligaments that keep the breast tissue in place through connections between the skin and chest wall. Although the breast itself does not contain any muscle, the pectoral muscle lies against the chest wall and beneath both breasts providing support[107]. The fatty tissue fills the space between the connective tissue and glandular tissue and is the main determinant of breast size[107]. The glandular tissue refers to the glands, called lobes, embedded in the fatty and fibrous tissue. Each breast contains between 15 and 20 lobes, each of which comprises many smaller lobules which are the milk producing centres[108]. Lobules are arranged in clusters. Connecting the lobes to the

nipple are ducts, thin tubes that carry milk to small openings in the nipple of nursing mothers[108]. The breast tissue is sensitive to three major hormones: oestrogen, progesterone and prolactin[108].

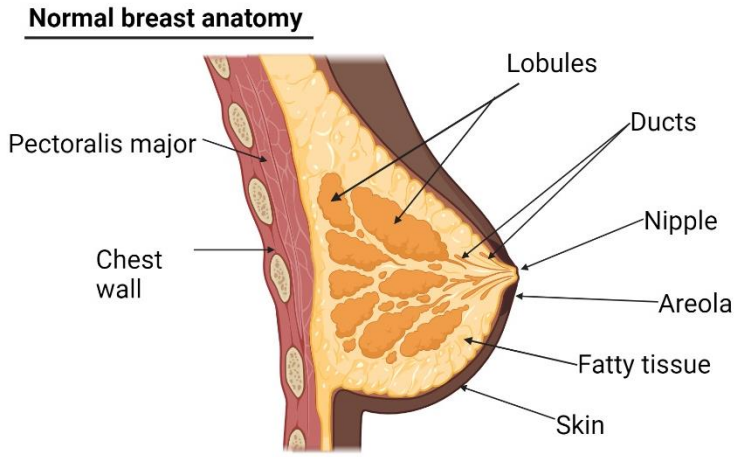


Figure 16: **Breast anatomy.** Anatomy of a normal breast in an adult woman. *Created with BioRender.com.*

1.5.3 A Brief History of Breast Cancer

On a poorly copied papyrus, dating from 2625BC, Ancient Egypt, the physician Imhotep describes a patient with bulging masses of the breast[2,79]. Designated case number 45, it was the only case of which no treatment or therapy was prescribed[2,79]. Accepted as one of the first clinical descriptions of breast cancer, over time its treatment would expand to include primitive therapies and brutal surgeries, up until today's modern precision medicine[79,109].

In the 1890's, in the male-dominated medical field, the breast was viewed by many physicians as "one of the most dispensable parts of the body" and easily disposed of[2,79]. William Halsted (1852-1922), an

American surgeon, was a pioneer of radical surgery. He believed that breast cancer would arise in a small cluster and spread outwards in a radial fashion, thus he conjectured that not only removal of the primary tumour, but the surrounding tissue and pectoralis major muscle could completely cure the patient[2,79]. Halsted reported a recurrence rate of 6%, a great improvement compared to other surgeries of the time which reported more than 9 times this number[110,111]. However, these surgeries would leave women traumatised and disfigured, resulting in fewer women seeking treatment early out of fear[112]. Furthermore, more than half of Halsted's patients succumbed to the disease within three years post-surgery[2]. Even so, this method would go relatively unquestioned as the standard treatment of breast cancer for nearly 50 years[2,79].

> *"Standardization must not, however, be allowed to create a fixed belief that no further improvement is possible and that any suggested change is necessarily to be regarded with disapproval"*[112] – Geoffrey Keynes, 1937

In 1937, English surgeon Geoffrey Keynes (1887-1982), expressed his doubt on the radical mastectomy[112]. Keynes trialled and advocated that some patients would benefit more from treatment with irradiation, whilst others benefit from a combination of conservative surgery and irradiation[112]. His results were comparable to the radical method (three-year survival rates: 83.5%, irradiation; 79.2% radical mastectomy)[109,112]. Keynes' method came with obvious advantages: if patients could receive the same results and be spared from disfigurement and risk of operative mortality, would this not be an avenue worth pursuing?[112]. Unfortunately, the reception of his voice of caution was met with disbelief and he was regarded as an eccentric[109].

The concept of categorising patients was popularised in the early 1900's. Boston surgeon, Robert B. Greenough (1871-1937) and German

physician Steinthal (*dates unknown*), would come to categorise breast tumours into three groups[109]. Greenough assigned classes based on microscopic examinations, whilst Steinthal assigned stages on the basis of invasiveness and involvement of lymph nodes[109]. Steinthal advocated that stage III tumours should not be operated on. Staging of breast tumours would become popularised in Germany and Scandinavia, but was not commonplace in the USA before the 1950's[109]. Today, staging and grading of breast tumours remains a prominent feature of breast cancer diagnosis.

Throughout the 20th century the rate of discovery would influence the way breast cancer was treated. Preceding the fall of radical surgery in 1981, the era of drug discovery brought about a new approach to treating cancer. The discovery of oestrogen and its receptor (ER) would eventually lead to the separation of breast tumours that do or do not express ER[2,113-115]. The discovery that tamoxifen, originally developed as a form of birth control, was a potent modulator of oestrogen, demonstrated the first use of a chemopreventative drug in cancer treatment[116]. Patients with ER+ breast tumours treated with tamoxifen reduced relapse rates by almost 50%[2].

The diagnosis and treatment of breast cancer in the 20th century was a period of heated debates, the collapse of "an entire culture of surgery"[2] and molecular and genetic discovery. Today, breast cancer may be treated with surgery, chemotherapy, hormone therapy, targeted HER2 therapy, radiation, immunotherapy, or combinations of the aforementioned. History has paved the way to the treatments in use today, but it has also done so for the methods in which to guide these treatments, all derived from the painstaking labours of researchers before us. Perhaps, in turn, we too will one day be humbled by the inevitable future of discovery.

1.5.4 Clinicopathological Features of Invasive Breast Cancer

Clinicopathological features are *manifestations and symptoms directly observable by a physician through laboratory examination*[117]. Traditionally breast cancer is grouped according to these features which will influence the way a patient is diagnosed and treated. For breast cancer, clinicopathological features include but are not limited to:

- Tumour size
- Lymph node status
- Morphological subtype
- Histological grade
- Hormone receptor status
- HER2 status
- Proliferation status
- Molecular profile

In Norway, these features, in addition to age, will indicate response and the type of adjuvant systemic treatment a patient will receive[118]. The focus of this thesis is breast cancer in women.

1.5.4.1 TNM Classification

The TNM classification is an anatomic classification used in breast cancer[119]. It consists of the prognostic factors: tumour size (T), regional nodal involvement (N), and distant metastasis (M) to describe the extent of disease and to group patients into stages with comparable outcomes[119]. The tumour size (T) may describe not only the size of the tumour but how invasive it is. Regional nodal involvement (N) indicates if cancer is present in the lymph nodes. The presence or absence of

distant metastasis (M) is also indicated. The TNM system is outlined in the WHO19 Breast Tumours Guidelines[89].

All three factors are strong prognostic indicators. Increasing tumour size is associated with decreasing survival[120-125]. In a node negative cohort, patients with tumours <1cm were observed to have a higher 10-year recurrence free survival (82%) than patients with tumours between 2 and 5cm (66%)[126]. This was also apparent for overall survival (79% vs. 66%, respectively)[126]. Projected relapse survival rates over a 20-year period after initial treatment were reported by Rosen and Groshen[124,125] (Table 5). Furthermore, although tumour size is independently prognostic, it is also correlated with lymph node involvement[89,120,127]. Data from the SEER programme in the United States, indicated that patients with tumour diameters ≥5cm with lymph node involvement (LN+) had a survival rate from 45.5% compared to 96.3% for patients with tumour diameters <2cm and with no lymph node involvement (LN-)[127]. Lymph node involvement is also independently prognostic, the higher the number of positive nodes, the worse the prognosis[128-131]. Presence of distant metastases is an indication of the primary tumour's ability to disseminate, survive, and grow in other regions of the body. As for all cancers, presence of distant metastasis is associated with poor survival outcomes.

Table 5: Projected relapse survival rates according to tumour size, over a 20-year period following initial treatment, as reported by Rosen and Groshen (1990)[124,125].

Tumour size	Projected relapse survival rate
<1cm	88%
1.1 – 1.3cm	73%
1.4 – 1.6cm	65%
1.7 – 2.2cm	59%

1.5.4.2 Morphological subtype

Breast cancer is categorised into different types according to clinicopathological features. Each type has distinct morphologies and clinical implications[132]. Approximately 70-80% of all invasive breast cancer is of the type *invasive carcinoma of no special type* also referred to as *invasive carcinoma NST* or *invasive ductal carcinoma*[118]. Other specialised types are classified if more than 90% of the morphology correspond to, for example, lobular, mucinous, tubular or cribiform[89]. If 50-90% of the tumour displays specialised morphology then the tumour will be classified as mixed non-specific and the specific type[118]. Approximately 5-20% of all invasive breast cancer is lobular, and the remaining specialised types account for between 1-2% each[89,118].

1.5.4.3 Histological Grade

Histological grading of breast tumours for prognostic stratification of patients is widely performed using the Nottingham Grade. The Nottingham Grade was modified from the Scarff-Bloom-Richardson grading system, which can be traced back to the principles of Greenough in 1925[133,134]. The Nottingham Grade is composed of three features: tubular formation, nuclear pleomorphism and mitotic count (Table 6). These features describe the degree of tumour differentiation. Tubular formation refers to the percentage area within the tumour that display tube-shaped structures, whilst nuclear pleomorphism refers to the degree of irregularity of nuclear outlines and number and size of nucleoli[89]. Mitotic count is an indicator of the level of proliferation in the tumour as result of assessing the number of actively dividing cells (mitoses). Each feature is scored, and the score combined determines the grade of the tumour (Table 6). Patients with a grade 1 tumour have a higher 10-year survival rate (85%) in comparison to patients with a grade 3 tumour (45%)[124].

Table 6: Nottingham Grade Scoring System[89]

Feature	**Score**	**Description**
Tubular formation	1	>75%
	2	10-75%
	3	<10%
Nuclear pleomorphism	1	- Nuclei similar in size and shape to benign epithelial cells. - Nucleoli are not visible or inconspicuous
	2	- Nuclei are larger than benign epithelial cells (1.5-2x). - Nucleoli are small or inconspicuous
	3	- Nuclei are larger and vary in size and shape (>2x). - Nucleoli are prominent.
Mitotic Count	1-3	See Appendix 2 –
Grade	**Score**	**Description**
Nottingham Grade 1	3-5	Well differentiated tumour
Nottingham Grade 2	6-7	Moderately differentiated tumour
Nottingham Grade 3	8-9	Poorly differentiated tumour

The Nottingham Grade has been combined with lymph node status and tumour size to form the Nottingham Prognostic Index (NPI) which is routinely used in the United Kingdom[135]. The NPI is calculated using the following formula[136]:

$$NPI = (0.2 \times TS) + N + G$$

TS: tumour size, the maximum diameter of the tumour (cm)
N: number of lymph nodes involved (0 nodes = 1, 1-3 nodes = 2, >4 = 3)
G: grade of the tumour (grade 1= 1, grade 2 = 2, grade 3 = 3)

The NPI identified three prognostic groups with decreasing 15-year survival rates: good prognosis (80%), moderate prognosis (42%) and poor prognosis (13%), as reported in a cohort of 1629 patients with operable breast cancer <70 years of age[136]. Similarly, a study on a Danish cohort (n= 4,791 patients) reported 10-year survival rates as

79%, 56%, and 25% for good, moderate and poor prognosis groups, respectively[137].

The Nottingham Breast Pathology Research Group, who developed the NPI, is also working on developing and evaluating two updated versions of this index. These are the NPI+, which factors in molecular subtypes, and the Nottingham Px (NPx) which aims to be more cost-effective and consists of the tumour size, grade, progesterone receptor status, and Ki67 score[138,139].

1.5.4.4 Hormones and their Receptors

Oestrogen

Prior to its purification in 1929, by Edward Doisy, oestrogen had already been linked to breast cancer in 1896[2,140]. Scottish surgeon, George Beatson, observed that when he removed the ovaries of three breast cancer patients, the tumours shrank in response[79]. However, when he tried to replicate these experiments, not all the tumours responded[2]. In an interview with Elwood Jensen (1920-2012), credited for the discovery of the oestrogen receptor (ER), he explained that a common treatment of breast cancer in the 1950's was the removal of the ovaries (source of oestrogen in pre-menopausal women) or adrenal glands (source of oestrogen in menopausal women)[109,114]. However, only around 1/3 of patients responded to this therapy, and clinicians would have to wait around 8 months to see if they did[114,141,142]. During this time, the tumours of the patients that did not respond to the surgery, would continue to grow[114]. Yet, if the clinicians could identify which patients would respond and which would not, non-responders could receive chemotherapy early, and responders could potentially avoid chemotherapy[114,143]. This observation is used in precision treatment strategies today.

The majority of all invasive breast cancers are ER positive (ER+)[89]. The sex steroid hormone oestrogen exerts its action by binding to it's receptors (ERα and ERβ)[144]. This activates a downstream signalling pathway that controls gene expression of a large number of target genes[144,145]. These targets play a role in the processes of cell proliferation, differentiation, and apoptosis[146]. Oestrogen is primarily synthesised in the ovaries but is also produced by the adrenal glands and adipose tissue[144]. Oestrogen refers to four female hormones estrone, estradiol, estriol and estretrol. Estradiol is the predominant oestrogen circulating in humans, whilst the others are synthesised during specific conditions such as pregnancy (estretrol, estriol) and menopause (estrone)[144].

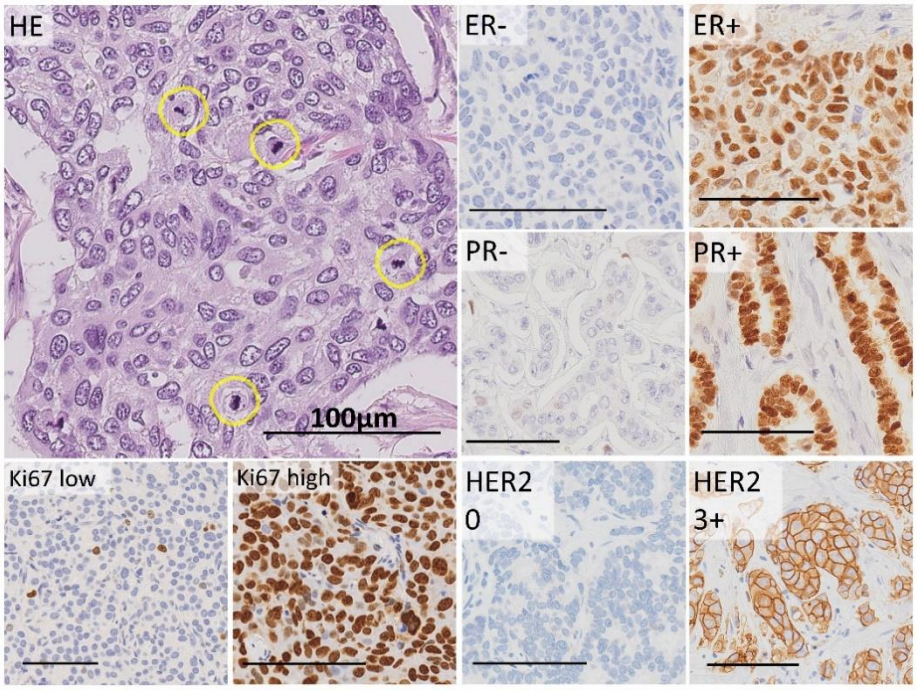


Figure 17: **Stains used in breast cancer histopathology.** Haematoxylin and eosin is the gold standard stain for visualisation of tissue morphology (yellow circles highlight mitotic figures). Immunohistochemical stains allow for the detection of specific proteins such as ER, PR, HER2 and Ki67.

Progesterone

The sex steroid hormone progesterone was first isolated by William M. Allen and George W. Corner in 1929 and eventually named in 1935[147,148]. Like oestrogen, progesterone is derived from cholesterol, but is also an important upstream intermediate in the biosynthesis of estradiol and other hormones[148]. Purification of the progesterone receptor (PR) occurred in the early 1970's, and its synthesis is regulated by none other than oestrogen[149]. In addition, PR is a dominant modulator of ER activity[149]. PR expression is an independent prognostic variable in breast cancer, where absence of expression is significantly associated with poorer overall and disease-free survival, in both ER+ and ER- disease[150].

Histopathological Examination of ER and PR

ER+ status is determined by examining a tissue section stained using IHC for detection of ER (Figure 17). In Norway, if ≥1% of tumour nuclei are stained, the tumour is considered ER+, and if <1% it is considered ER-[118]. For PR status the tissue is stained for detection of PR, and if ≥10% of tumour nuclei are stained, the tumour is considered PR+, if <10% it is considered PR-[118] (Figure 17). Additionally it is reported if ER or PR is >50% or an estimated percentage is reported[118].

Effect of ER and PR Status on Treatment Decisions

Presence of ER and PR indicates the likelihood of response of breast cancer to endocrine therapy. ER+ tumours are receptive to anti-oestrogen therapies* and these are usually prescribed for 5-10 years post-diagnosis[118,151]. Adjuvant treatment with tamoxifen has demonstrated improved survival over a 5 year period, and reduced

*Some patients will inevitably display intrinsic or acquired resistance.

recurrence rates over a 10 year period[152]. Tamoxifen is primarily used in pre-menopausal women to block the action of ER, whilst aromatase inhibitors are used in post-menopausal women or in combination with tamoxifen[151]. Aromatase inhibitors alone is not recommended for pre-menopausal women as the ovaries in these women are highly active and produce such a quantity of oestrogen that aromatase inhibitors are inefficient at blocking its production[151]. Tamoxifen response has a high degree of variability between patients, due to differences in metabolism[153]. Tamoxifen is a prodrug, whereby its metabolites are the causative factors[153]. Measurement of these metabolites has suggested prognostic and monitoring potential, which can be used to adjust dosage according to a patients ability to metabolise tamoxifen[154].

Whilst tamoxifen has an antagonistic effect in the breast by competitive inhibition of oestrogen binding to ER, it has a agonistic effect on ER in the uterus in postmenopausal women[155,156]. The biological mechanism behind the contradictory roles of tamoxifen in the breast and endometrium, is not yet fully understood. It has been theorised that it may be due to recruitment of different co-activators and co-repressors in these tissues, compensatory expression of ERα isoforms, or activation of an orphan G-protein couple receptor (GPR30) in the endometrium[157].

Although ER status is the primary predictor for endocrine therapy decisions, PR status is also indicative for therapy response[118]. The role of PR in breast cancer is primarily as a prognostic and predictive marker of response to endocrine therapy by predicting the existence of functional ER[158,159]. ER+ tumours that display a loss of PR expression during tamoxifen therapy have a worse prognosis[149,160,161]

PR is expressed to varying degrees in ER+ breast cancers. While ER+/PR+ tumours are more common than ER+/PR-, ER-/PR+ tumours are

rare[162,163]. Several studies have demonstrated differences in survival in tumours with double, single, or null receptor positivity[159,164-168]. Tumours positive for both ER and PR have a greater response to tamoxifen therapy, and a better overall and relapse-free survival than single or null receptor positive tumours[159,164,166,167]. One study reported that ER-/PR+ tumours benefited from tamoxifen therapy compared to no tamoxifen[168]. Two studies report decreasing breast cancer specific survival from double, to single, to null receptor positive tumours[165,169]. Also, single receptor positivity has been demonstrated to be associated with other clinicopathologic markers. For example, ER-/PR+ tumours have been associated with larger tumour size[166].

1.5.4.5 HER2

HER2, also referred to as ERBB2, is a receptor tyrosine-protein kinase, encoded by the *ERBB2* gene[170]. It is a cytoplasmic membrane-anchored protein. ERBB2 does not directly bind any known ligands, but it can be phosphorylated by another receptor, by forming a heterodimer with its preferred partner, another ERBB receptor, such as EGFR (ERBB1/HER1)[171]. When activated, it can induce downstream signalling pathways such as PI3K/AKT, MAPK, PLYγ/PKC and JAK/STAT[170]. These pathways regulate genes that play a role in cell survival, proliferation, differentiation, motility, apoptosis, invasion, migration, adhesion and angiogenesis, all of which are key in carcinogenesis[170].

HER2 in Breast Cancer

HER2 overexpression (HER2+) is commonly observed and accounts for approximately 10-25% of all invasive breast cancer[89,118,172]. Furthermore, it is associated with more aggressive tumours and poorer prognosis[89,173-175]. In 1998, trastuzumab (Herceptin®), a recombinant monoclonal antibody targeting HER2, was approved by the FDA for the

treatment of women with HER2+, metastatic breast cancer[176]. Treatment with trastuzumab has demonstrated delayed disease progression, longer survival and a reduction in the risk of death[177]. In the early 2000's clinical trials reported benefits for adjuvant treatment with trastuzumab in HER2+ patients[178-180]. Other anti-HER2 therapies also report similar results[181-183]. Recently an intermediate category to HER2+ and HER2-, was introduced: HER2-low. Previously these patients were categorised as HER2-, however, recent evidence suggests that patients with an IHC score 1+, or 2+ with negative ISH (see next paragraph) may benefit from treatment with trastuzumab[184]. Treatment of this patient group was recently approved in the USA and the EU[185,186].

Histopathological Examination of HER2

HER2 status should be examined in all newly diagnosed and metastatic breast cancers. HER2 status is determined by IHC and/or in situ hybridisation (ISH) procedures such as fluorescent ISH (FISH). HER2 is first assessed by IHC (Figure 17). The breast tissue section is stained for the detection of the HER2 protein. The membrane staining is scored as 0 (negative: no/incomplete membrane staining in <10% of invasive tumour cells), 1+ (negative: faint/weak membrane staining in >10% of invasive tumour cells), 2+ (borderline: weak to moderate complete membrane staining in >10% of tumour cells), or 3+ (positive: strong complete membrane staining in >10% of invasive tumour cells), which reflects the completeness and intensity of the stain[89,187]. Tumours scored as 2+ and 3+ are recommended for verification of HER2 status by in situ hybridisation (ISH)[118]. While the IHC method assesses HER2 status at the protein level, ISH measures HER2 overexpression at the gene level. Further details on ISH scoring criteria are outlined in the ASCO®/CAP 2019 guidelines[188].

1.5.4.6 Proliferation

Proliferation is one of the most fundamental biological processes in normal mammalian cells. Yet, sustaining proliferation is also one of the hallmarks of cancer (Figure 13). Cellular proliferation is the process of replicating DNA within a cell followed by cell division[189]. The cell cycle describes a series of tightly regulated events which include protein phosphorylation cascades that initiate movement through the cycle and checkpoints that monitor completion of critical events that can delay progression[189]. Progression is heavily regulated by cyclins and cyclin-dependent kinases (CDKs)[190], where the cyclins are the activators and CDKs are the catalytic units[190]. One of the major inhibitors of cell cycle progression is detection of DNA damage.

The cell cycle is separated into five phases: G_0, G_1, S, G_2 and M (Figure 18A). The first phase, G_0, is the resting phase of the cell also known as quiescence. Cells in G_0 can be stimulated to enter the cell cycle at G_1 to begin the process of cellular proliferation. The G_1 phase is characterised as period of cell growth when the cell prepares for DNA synthesis. At the end of G_1 the cell will commit to traversing the entire cell cycle[189]. DNA is replicated during the S phase, following by another period of growth in G_2 when the cell prepares to enter the M phase[191]. The last phase, M phase, is when the cell undergoes mitosis, active cell division (Figure 18B). During mitosis, the chromosomes will condense into two chromatids held together at the kinetochore in prophase, connect to spindle microtubules and align at the equator of the cell in metaphase, separate and be pulled to opposite poles in anaphase, before eventual separation and unravelling in telophase[191] (Figure 18B). Finally, the cell cleaves into two identical daughter cells by a process termed cytokinesis[191].

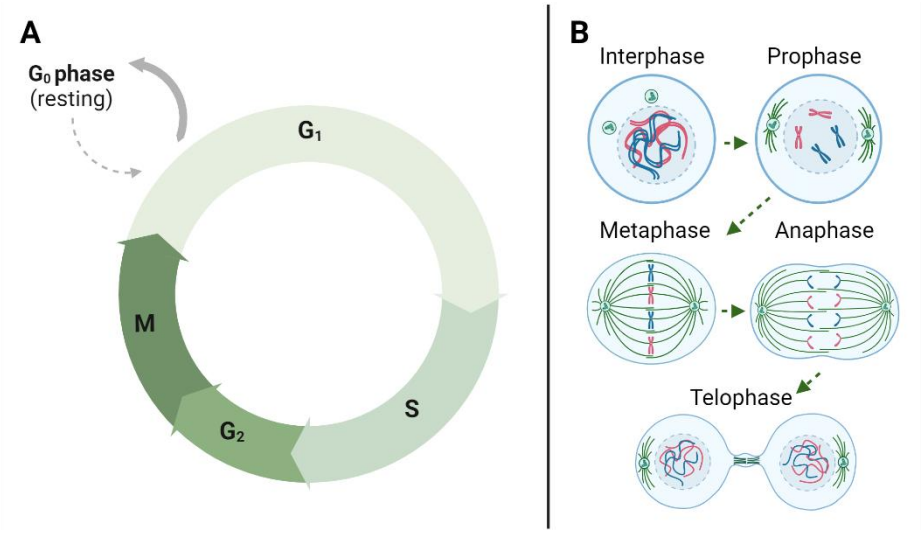


Figure 18: A) **The Cell Cycle**. B) Interphase and the stages of mitosis – M phase. *Created with BioRender.com*.

As previously described, cancer is the result of uncontrolled growth and proliferation. To reach this state, cancer cells develop mechanisms by which they escape the tightly controlled regulation of the cell cycle. These mechanisms include gain-of-function mutations of oncogenes that promote cell proliferation or loss-of-function mutations in tumour-suppressor genes which inhibit checkpoint regulation[189]. For example, the tumour-suppressor gene for p53 is commonly mutated in cancer[85]. The p53 protein is activated in response to checkpoints detecting DNA damage, but due to a loss-of-function mutation in its gene, it is unable to halt progression in the cell cycle[189]. This is a way in which the cancer cell can continue to proliferate but also increase the presence of potentially oncogenic mutations.

In breast cancer, measurement of proliferation has prognostic and predictive value[192-194]. Histopathological examination of proliferation is performed by counting mitotic figures and assessing Ki67 positivity in IHC stained tissue sections.

Mitosis in Breast Cancer

Mitotic count is an important component of histological grading. The quantification of mitotic activity in breast cancer is also independently prognostic[195,196] and often reported separately. The mitotic count is generated by counting the number of mitotic figures in a predefined area and usually reported as the number of mitoses per mm^2. The mitotic count can be performed directly on HE stained tissue sections (Figure 17). Further details regarding quantification of mitoses can be found in *section 3.4.1*.

Ki67 in Breast Cancer

Assessment of proliferation using IHC allows for a clearer distinction between normal, negatively stained cells and cells positively stained for a proliferation marker (Figure 17). The protein Ki67, encoded by the *MKI67* gene, is expressed during all phases of the cell cycle, except for G_0 (Figure 18). It has documented prognostic, predictive, and monitoring value[196-202]. Ki67 is primarily assessed in hormone receptor positive (HR+), HER2- breast cancer, and can be used as a surrogate marker for distinguishing between the luminal subtypes. The most recent St. Gallen (2021) consensus recommend that tumours (ER+, HER2-) with a Ki67 score <5% are not recommended for adjuvant chemotherapy, whilst patients with a Ki67 score ≥30% are recommended for adjuvant chemotherapy[203]. Despite its value, it is a controversial marker due to its lack of standardisation and consequently poor intra- and inter-variability[118,203-206]. Although efforts have been made towards standardisation of quantification methodologies and cut-offs for assigning low and high proliferation, different methods are still used. For example, Norway still advocates the use of hotspot score, whilst Sweden now uses the global score[118,207].

1.5.4.7 Molecular Profiling

At the turn of the millennium, the molecular profiles of breast cancer were being described[208-211]. This led to the classification of five intrinsic subtypes (Figure 19). These pioneer studies emphasised the genomic diversity of these tumours and highlighted how gene expression patterns may be used in precision medicine. Numerous commercial gene panels are available including Oncotype DX®, MammaPrint®, Prosigna® (PAM50), and EndoPredict®. Each panel, as a whole, is unique though there is some slight overlap in gene coverage[212]. Also, several biological and molecular pathways are share and proliferation and HR-related genes are frequently represented[212]. The multi-gene signature panels have demonstrated potential for several clinical applications, in addition to molecular subtyping, such as predicting treatment response, prognosis and staging[210,213-216].

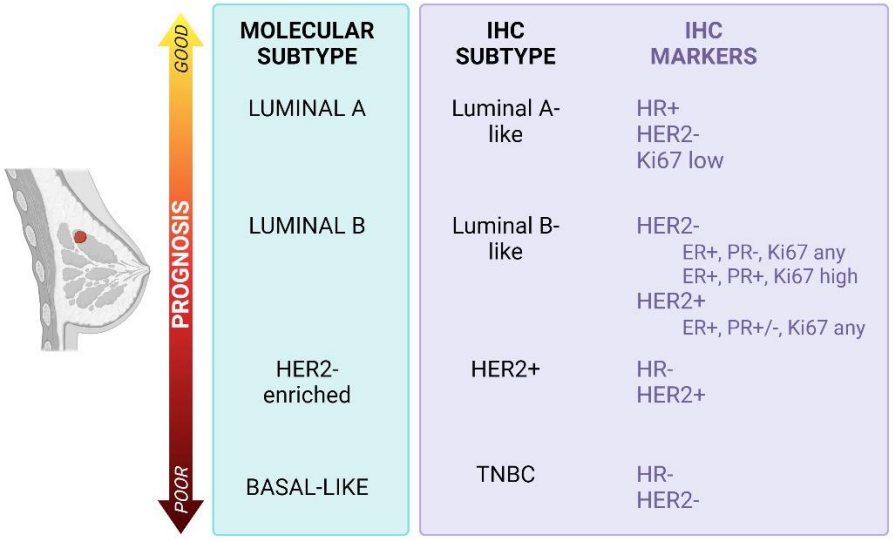


Figure 19: **The PAM50 molecular and IHC (surrogate) subtypes of invasive breast cancer**. *Created with BioRender.com.*

A comparative study assessing the prognostic value of clinicopathological features and gene expression models, revealed that the latter was superior[217]. Despite a correlation between the two, clinicopathological features are not able to fully recapitulate the intrinsic subtypes[199,216,218]. The use of molecular signatures over traditional markers has its own advantages, such as increased objectivity and improved prognostic value, and disadvantages, such as increased costs, and increased processing times.

At the St. Gallen 2021 consensus discussion, 61% of the panel believed that Ki67 should still be tested regardless of multigene signature[203]. Several studies have demonstrated a correlation between Prosigna®, EndoPredict® and Oncotype DX® and Ki67 score[219-224]. As proliferation is a major driver of the molecular profile, this observation is unsurprising. Yet, despite the advantages of these panels, Ki67 is still a relevant and cheaper alternative for prognostic assessment in breast cancer[221,225]. Particularly when molecular panels are unavailable or considered too costly. Ki67 has also been suggested as a screening tool for recommending molecular testing in cases Ki67>5% and <30%[226].

1.5.5 AI in Breast Cancer

> *"[Cancer] has ever been the reproach of the medical art, and the most learned and experienced of the profession have employed their time and attention to but little purpose, towards perfecting its cure"*[227]
> Robert White, 1784

Although breast cancer treatment has vastly improved in recent decades, the consequence has been that diagnosis and treatment decision pipelines have become more complex. Although oncologists and pathologists alike are experienced in these decisions, the influx of new data is constantly changing the way cancer is viewed and treated. Therefore, the use of AI as a decision-support tool may help lift this

burden and allow for the allocation of time on the more complex aspects of diagnosis. It may also expand our knowledge beyond what we know today.

Applications for AI in breast cancer include detection/screening, classification and prediction, and cover a range of imaging modalities such as mammograms, ultrasound, magnetic resonance imaging (MRI), and histopathological images[228].

1.5.5.1 Mitosis Detection

Tools for AI-assisted histological grading of breast cancer have been developed[229-231]. AI-assisted detection of mitoses has been a popular task in the field and has inspired international challenges such as the Mitosis Domain Generalisation (MIDOG) Challenge[232,233], the Assessment of Mitosis Detection Algorithms 2013 (AMIDA13) Challenge[234] and TUPAC16 Challenge[235]. Several studies have developed tools, majority CNNs, for automated detection of mitotic cells in breast cancer[236-243]. The utilisation of AI support tools has been shown to enhance accuracy, sensitivity, and specificity compared to settings without it[244,245]. Furthermore, overall time-saving of 27.8%, in one study, was reported when AI support was used[244]. Implementation of a locally developed AI algorithm for mitotic figure detection, integrated with the local PACS, is reported by the University Medical Centre Utrecht[246].

1.5.5.2 Lymph Node Metastasis Detection

Detection of small clusters of metastatic breast cancer cells in lymph tissue is challenging. AI-assisted detection of micrometastases in lymph nodes may assist a pathologist for more efficient identification and spare IHC. One study reported higher accuracy for pathologists using AI-assistance tools than the algorithm or pathologist alone[247]. An AI-

assistance tool for this purpose has also reported notable time-saving benefits[247,248]. Commercial algorithms for detection of lymph node breast micrometastases are available.

1.5.5.3 Biomarker quantification

Many DIA platforms/applications offer CE-IVD/FDA approved quantitative biomarker solutions (ER, PR, HER2, Ki67, P53). Each differs in design and methodology. Open-source, research use only, solutions are also available, such as QuPath[249]. Automated scoring of ER by DIA reports excellent agreement with pathologists and may eliminate the shortcomings of a subjective evaluation[250]. Some studies have trained deep learning models to predict ER or PR status in HE (haematoxylin and eosin) images[251,252]

Several studies have demonstrated Ki67 quantification using DIA as being equal or superior to manual counting regarding reproducibility and accuracy[198,206,253-256]. Many of these studies have compared automated counts with manual counts and found good to excellent agreement. One study observed increased inter-observer agreement, reduced turnaround time, and decreased Ki67 error when AI was introduced in comparison to no AI[256]. Those studies that have investigated its prognostic potential using survival/recurrence analysis demonstrate improved prognostic value for DIA methods in comparison to manual, specifically in the HR+ subgroup[257,258]. However, no fixed cut-off has yet been recommended for DIAs, and each DIA method differs – some use hotspot quantification, some global, some average. Although reported reproducibility and prognostic value may be sufficient, further investigation for validation and standardization regarding cut-off for low and high-Ki67 is required.

Scoring of HER2-IHC has been a target of DIA algorithms. With HER2-FISH as the ground truth, one study observed that automated DIA of HER-IHC had high specificity (93.2%) and sensitivity (96.4%)[259]. The study reported that DIA would have resulted in a reduction in FISH in the cohort, which would have resulted in time- and cost-savings[259]. Algorithms for quantitation of tumour infiltrating lymphocytes (a prognostic and predictive biomarker) in TNBC have also been reported[260-264].

1.6 Endometrial Carcinogenesis

1.6.1 The Female Reproductive System

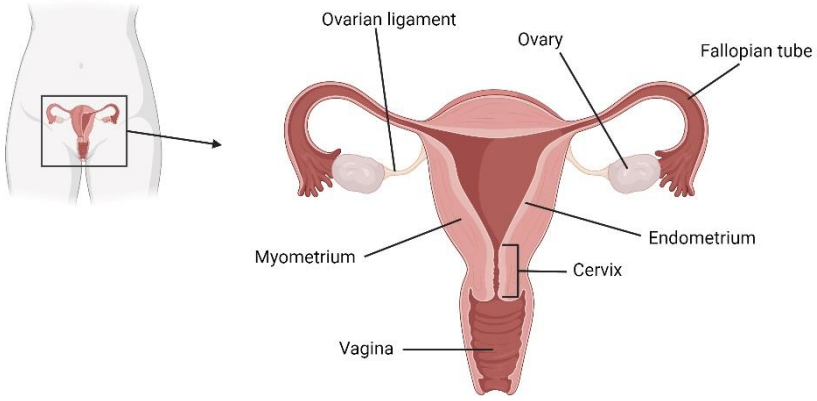


Figure 20: **The Female Reproductive System**. *Created with BioRender.com*

The female reproductive system is composed of the uterus, ovaries, fallopian tubes, vagina, accessory glands, and external genital organs[265] (Figure 20). The primary function of this system is to regulate processes involved in reproduction in females. These processes include the production of female sex hormones, producing and sustaining eggs, and regulating a favourable environment for a growing foetus[265].

The menstrual cycle (Figure 21), is the cyclic process by which the uterine lining is shed, in absence of egg fertilisation, and prepared, in anticipation of fertilisation. This process occurs in response to the interactions of hormones produced by the hypothalamus, pituitary gland, and ovaries[266].

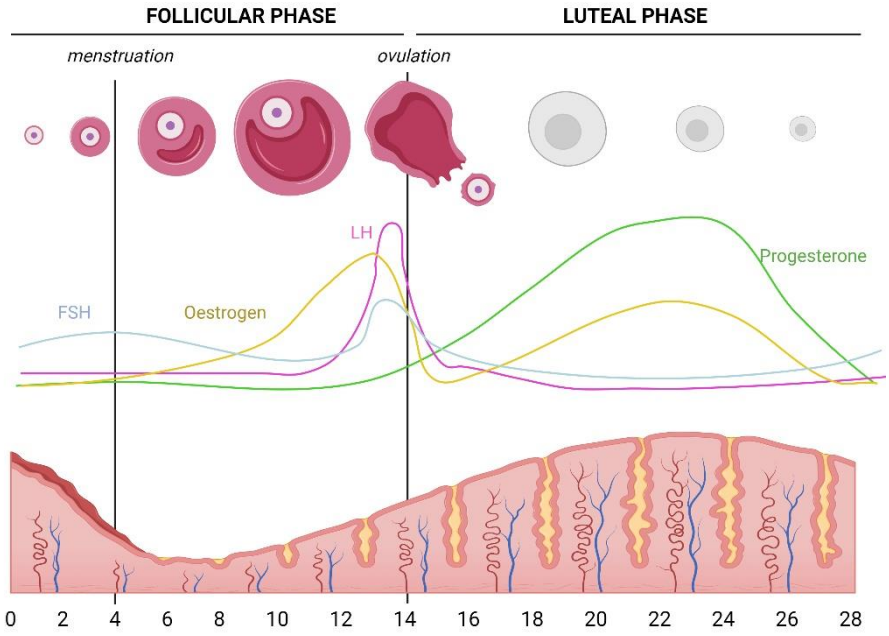


Figure 21: **The Menstrual Cycle**. *Created with BioRender.com*

Four hormones play an essential role in the regulation of the menstrual cycle: oestrogen, progesterone, follicle stimulating hormone (FSH) and luteinising hormone (LH). Oestrogen induces endometrial growth (proliferative phase) during the follicular phase, resulting in a thickening of the uterine lining, to prepare it for implantation of a blastocyst[267]. Peak oestrogen levels are reached prior to ovulation, and production continues during the luteal phase (Figure 21). Progesterone levels increase following ovulation, and stimulates the transition of the endometrium from a proliferative phase to the secretory phase[268]. FSH

regulates oestrogen synthesis and stimulates follicular development[269]. LH stimulates the release of the follicle from the ovary and peaks just prior to ovulation[266].

The endometrium is affected by varying concentrations of oestrogen and progesterone (Figure 21). ER is dominant in stromal and myometrial epithelial cells and PR in endometrial epithelial cells[267]. PR levels decrease in the late luteal phase alongside oestrogen when implantation does not occur. Whilst oestrogen induces proliferation of the endometrium, progesterone inhibits it[267].

As a result of oestrogen and progesterone action, morphological changes in the endometrium occur. These characteristic changes indicate the stage of the menstrual cycle the endometrium is currently in, for example proliferative (follicular phase) or secretory (luteal phase)[266].

1.6.2 Epidemiology

Endometrial cancer, cancer of the *corpus uteri*, is the sixth most common malignancy in women worldwide[97]. In 2020, 417 367 new cases were diagnosed and 97 370 women succumbed to endometrial cancer worldwide[97]. Incidence of endometrial cancer has increased over recent decades, thought to be due to a rise in obesity and shifts in female reproductive patterns[270,271].

In Norway, 817 endometrial cancer cases were newly diagnosed in 2022[96]. Five-year relative survival (2018-2022) is 86.2%, markedly improved since 1970 which reported 70.7%[96,106]. Although incidence has been increasing since 1965, it has plateaued in the last 10-15 years (Figure 22).

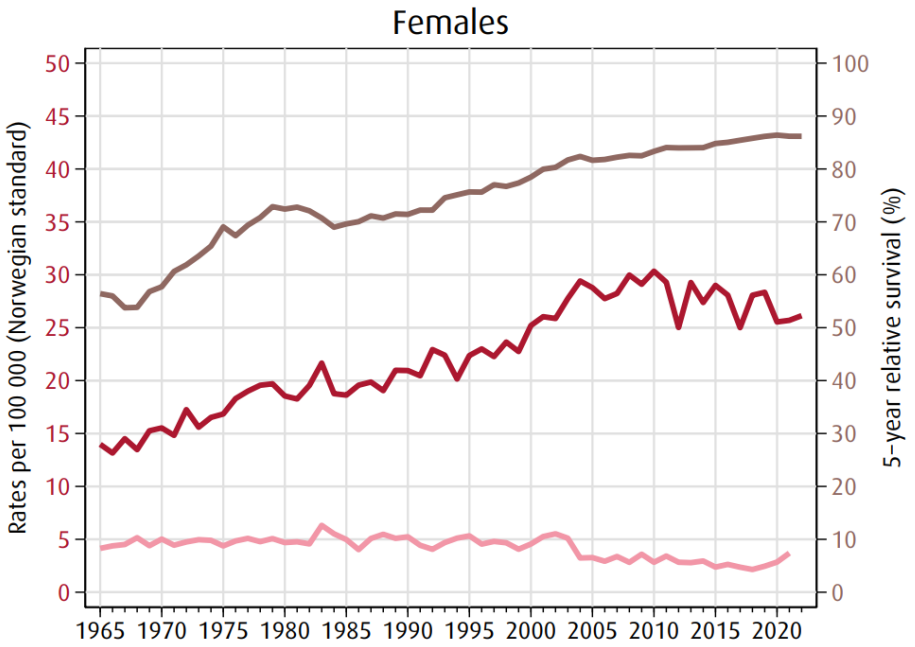


Figure 22: **Cancer of the corpus uteri**. Trends in incidence, mortality and 5-year survival relative survival proportions for cancer of the corpus uteri in Norway[96]. The top line (brown) represents **survival**, the middle line (red) represents **incidence**, and the bottom line (pink) represents **mortality**. Figure reproduced with permission from the Cancer Registry in Norway.

Risk factors for development of endometrial cancer are listed in Table 7. Exposure to unopposed oestrogen stimulation is cited as one of the major causes of endometrial cancer development[272-275]. The primary source of oestrogen in the body is the ovary and adrenal glands. Another source of oestrogen, primarily in post-menopausal women, is adipose tissue[276]. Various studies have demonstrated the association of obesity and weight gain with endometrial cancer risk[277-280]. A study by Stevens *et al.* (2014) reported that a high BMI at age 18, and weight gain in adulthood was strongly associated with the risk of endometrial cancer[281]. Administration of tamoxifen, prescribed in the treatment of HR+ breast cancer, has been associated with increased risk for

endometrial cancer[282-284]. Patients taking tamoxifen should be made aware of the symptoms of endometrial cancer whilst they are being treated[275,285].

Table 7: **Risk factors for endometrial cancer**

Associated risk factors for endometrial cancer
Increased age[286]
Long-term, unopposed oestrogen stimulation[272-274]
Early menarche, late menopause[287]
Low parity[287]
Obesity[286,288]
Diabetes[289,290]
PCOS[291-293]
Lynch syndrome[294]
Cowden syndrome[295]
Tamoxifen treatment in breast cancer[282-284]
Family history of endometrial, ovarian, breast, or colon cancer

1.6.3 Endometrial Hyperplasia

1.6.3.1 Definition

Endometrial hyperplasia is characterised by the excessive proliferation of endometrial glands as a result of prolonged exposure to unopposed oestrogen stimulation[296]. It is described as a lesion with a greater gland-to-stroma ratio in comparison to normal surrounding endometrium[297]. It usually precedes endometrial endometrioid carcinoma.

1.6.3.2 Background

In 1985, Kurman and colleagues published a paper on the long-term study of 170 patients with endometrial hyperplasia[298]. They classified endometrial hyperplasia into two categories: those that did not display

cytological atypia and those that did. Further subgrouping of these categories, by glandular complexity, resulted in four classifications: simple hyperplasia (SH), complex hyperplasia (CH), simple atypical hyperplasia (SAH) and complex atypical hyperplasia (CAH). The simple categories displayed less glandular crowding than the complex categories. Additionally, progression to carcinoma differed from 1% (SH), 3% (CH), 8% (SAH), to 29% (CAH). The trend observed was not statistically significant. In 1994, this classification was adopted by the World Health Organisation (WHO) guidelines for diagnosis of endometrial hyperplasia[299]. It would form the basis of the classification of endometrial hyperplasia and prognostic stratification for treatment decisions for nearly 20 years.

The WHO94 scheme would unify the terminology and classification of endometrial hyperplasia in a time when previous terminology was confusing and a source of disagreement between experts[299]. However, the classification was still based on a visual assessment of morphological features of which cytological atypia is particularly poorly reproducible[299-303]. Furthermore, classification of four groups did not adequately reflect the treatment options: monitoring, progestin treatment or hysterectomy[299]. New classifications were proposed to improve the WHO94 scheme[302,304]. In 2003, the WHO recognised the EIN scheme proposed by the Endometrial Collaborative Group, fronted by George Mutter,[304,305] and in 2014 this was adopted (WHO14)[306]. Improved reproducibility of diagnosis and clinical applicability has been reported following transition from a 4-tiered classification system to a two-tiered system[307-310]. The latest classification WHO20 is an updated version of the WHO14 scheme[275].

1.6.3.3 The WHO20 Classification

The WHO20 terminology classifies two diagnostic groups:

1. Endometrial hyperplasia without atypia (EHwA)
2. Endometrial atypical hyperplasia (EAH) or Endometrial Intraepithelial Neoplasia (EIN)

Patients with endometrial hyperplasia may present with postmenopausal or irregular bleeding[275]. A histological sample is required for diagnosis, to assess morphological features of the endometrium. EHwA presents as a proliferation of glands of irregular size and is often lacking cytological atypia[275]. On the other hand, EAH presents as an increased gland to stroma ratio (glandular crowding) and cytological atypia, which is distinct from surrounding normal endometrium[275]. EHwA has a low risk of progression to endometrial cancer (1-5%) in comparison to EAH (8-40%) which also has a high-risk of concurrent carcinoma[275,311-318].

1.6.3.4 D-Score

The D-score was developed in the 1980's. Designed as a prognostic assistance tool, it utilised objective, morphometric measurements of morphological features to assign a risk of progression score (Table 8)[319]. There are three components to the D-score algorithm: the volume percentage stroma (VPS), outer surface density (OSD) and standard deviation of the shortest nuclear axis (SDS). This measures the architectural (VPS and OSD) and cytological (SDS) features of endometrial hyperplasia. Several studies discuss its prognostic value, and it has been validated in the USA, the Netherlands and Norway[320-327]. D-score can be used to assist with the classification of non-EIN and EIN (Table 8)[299,328]. A meta-analysis by Raffone *et al.* (2019) reported that classification of endometrial hyperplasia by the EIN system in

conjunction with D-score better predicted the risk of cancer than the WHO systems[329]. Several studies report similar accuracy between the EIN classification and the WHO schemes although sensitivity and specificity may differ[308,330-332].

Table 8: Risk of progression and EIN classification prediction according to the D-score.

Score	**Risk**	**EIN prediction**
D-score ≥ 1	Low	Benign hyperplasia
D-score 0-1	Slightly higher	EIN
D-score <0	High	

1.6.3.5 Endometrial Hyperplasia in Norway

There is no national registry for endometrial hyperplasia, in Norway. However, with more than 750 new cases of endometrial cancer reported each year, the Norwegian Department of Health, estimates that the number of cases diagnosed with endometrial hyperplasia is between 3,000 and 4,000 per year[328]. Classification of endometrial hyperplasia using the latest WHO scheme and D-score is recommended in Norway[328]. Patients considered low-risk, with a D-score ≥0 or EHwA diagnosis, are recommended for conservative treatment (progesterone treatment)[328]. Patients considered high-risk, with a D-score <0 or EAH/EIN diagnosis, are recommend for surgical treatment, where hysterectomy is the primary recommendation[328]. However, pre-menopausal women wishing to conserve fertility may receive conservative treatment with regular monitoring[328].

1.6.3.6 From Endometrial Hyperplasia to Cancer

The relationship between endometrial hyperplasia and cancer was first recognised in 1932 by Howard Taylor[333]. Approximately 80% of endometrial cancers are preceded by endometrial hyperplasia[298]. It is believed that the pathogenesis of the two begins with oestrogen[†]. It is well-established that endometrial hyperplasia occurs in a background of unopposed oestrogen stimulation[272-275,334]. As discussed in section 1.5.4.4, oestrogen is involved in regulating biological pathways related to cell proliferation, differentiation, and survival[146]. The term "unopposed" in conjunction with oestrogen stimulation refers to insufficient counterbalance by progesterone. Increased proliferation is observed during the follicular phase of the menstrual cycle, when progestin levels are low[148] (Figure 21). As a result of unopposed stimulation, prolonged proliferation gives rise to disordered endometrium[335]. Over time, this disorder increases and tissue transitions into early-stage EHwA[335] (Figure 23). Distinguishing disordered proliferative endometrium from EHwA is possible, as the former lacks a prominent increase in the overall ratio of glands to stroma[335]. Increased endometrial cancer risk is observed in women administered with exogenous oestrogens in the absence of progestins[146,273]. This observation is negated when progesterone is administered in combination[274,336,337]. One study suggested that unopposed oestrogen as a direct cause of endometrial carcinogenesis, is just as likely as progesterone deficiency[336].

[†] Endometrial cancer in this context refers to type I endometrioid cancer.

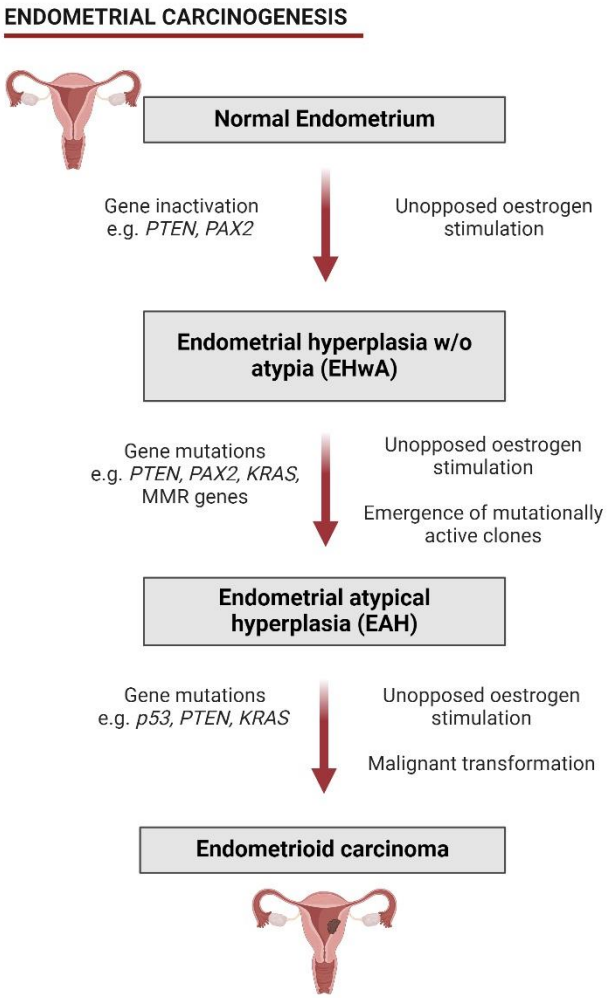


Figure 23: **Endometrial carcinogenesis (simplified)**. *Created with BioRender.com*

Whilst the average age of women diagnosed with EAH is 50-55 years, this is 5-10 years younger than for endometrial carcinoma, suggesting that progression of EAH to cancer may take several years[275,307,323]. In fact the average interval has been reported as 4 years, in the absence of coexisting cancer[323]. The emergence of EIN/EAH as localised monoclonal outgrowths of endometrial cells with altered cytology and

architecture is well-established[275,326,338,339]. Monoclonal growths acquire a growth advantage in comparison to normal surrounding tissue, that is polyclonal[326]. The emergence of mutationally active clones may eventually lead to an aggressive phenotype and lead to cancer (Figure 23). This emergence has been described as a separate event to oestrogenic stimulation, however, oestrogen may act as a positive selector of mutant cells[335,340,341].

1.6.4 Biomarkers in Endometrial Hyperplasia

To combat the interobserver variability of visual assessment of endometrial hyperplasia, a "panel approach" may improve accuracy of diagnosis. This could include assessment of protein biomarkers by IHC, genetic profiling, or quantitative measurement of morphological features. Several biomarkers have been investigated for their promise as candidates for such a panel. The WHO20 recognises that MSI (microsatellite instability), *PAX2* inactivation, *PTEN*, *KRAS* and *CTNNB1* (encodes β-catenin) mutation alterations are observed in EAH[275] (Figure 23). Alterations have also been reported in *FGFR2*, *ARID1A*, *p53*, and *MYC*[296]. No single gene alone has been sufficiently indicative of risk of progression, although assessment of mutational burden may have prognostic value[296].

1.6.4.1 PTEN

The tumour suppressor gene *PTEN* (phosphatase and tensin homologue) encodes a phosphatase protein of the same name. In 1997, it was mapped to chromosome 10 (10q23) and identified as frequently disrupted in multiple sporadic tumour types[342,343]. PTEN is an inhibitor of the PI3K/AKT signalling pathway[344]. Loss of PTEN function results in upregulation of this signalling pathway and promotion of cell survival and proliferation[345]. Elevated PI3K/AKT signalling is associated with

carcinogenesis[346]. PTEN expression fluctuates during the menstrual cycle[347]. Literature suggests that PTEN may be regulated by steroid hormones and microRNAs in the endometrium[348]. Oestrogen has been demonstrated to downregulate PTEN through increased phosphorylation, whilst progesterone upregulates PTEN by decreasing its phosphorylation[348-350]. Furthermore, *PTEN* germline mutations were discovered in Cowden disease, a disease conferring increased risk of cancer such as endometrial cancer[351].

Inactivation of *PTEN* has been described as an early event in endometrial carcinogenesis[352-355] (Figure 23). One study reported that *PTEN* mutations were more common in EAH progression cases than non-progression cases (60% vs. 35%)[296]. Furthermore, progressive PTEN protein loss is associated with cancer progression and concurrent cancer[334,355-358]. However, it is not considered a good independent predictor of progression or response to conservative treatment[334,355,359-361]. Approximately equal levels of PTEN-normal and PTEN-loss has been observed in women with proliferative endometrium and endometrial hyperplasia[358,359]. Additionally, presence of inactivated PTEN glands are reported in histologically normal endometrium[362]. Thus, PTEN may be recommended as an informative biomarker, but its role as an independent predictive and prognostic marker remains controversial.

1.6.4.2 PAX2

The *PAX2* gene encodes the transcription factor PAX2 (paired-box protein 2), which is active in embryogenesis, resistance to apoptosis and promotion of proliferation[363,364]. It is suggested that PAX2 is indirectly regulated by oestrogen and its receptors but its molecular mechanism in endometrial carcinogenesis is not yet thoroughly understood[365,366]. Additionally, PAX2 may regulate p53 through binding to its regulatory

region, inhibiting protein production[367]. As with PTEN, progressive loss of expression of PAX2 has been observed from normal endometrium to hyperplasia to cancer[358,366,368-370], however, one study observed an increase in PAX2 expression[371] (Figure 23). The role of PAX2 as a prognostic marker for progression is debated[372]. Some studies report no association between PAX2 expression and progestin therapy resistance or relapse following conservative treatment patients[352,373]. Although the usefulness of PAX2 as an independent diagnostic marker is debated, it may be useful in combination with histological evaluation as an adjunct marker for detection of EAH/EIN[374,375].

1.6.5 Endometrial Cancer

Endometrial carcinoma (EC) has been traditionally classified into two histopathological types: type I (endometrioid) tumours and type II (non-endometrioid). Type I tumours are often low-grade and associated with oestrogen stimulation[275]. Type II tumours are associated with more aggressive tumours and are unrelated to oestrogen stimulation[275]. Further stratification may be performed by staging (FIGO stage) and histological grading. Type I tumours represent 80-90% of all EC and have a better prognosis than type II[275]. The focus of this thesis is type I.

The stratification of EC into type I and type II is criticized for its poor inter-observer variability particularly in high-grade EC[275,376-378]. In 2013, the Cancer Genome Atlas (TCGA) described four distinct molecular subtypes of EC and suggested reclassification of EC[379]. The four subtypes were based on identification of different molecular markers (Table 9). These new subtypes demonstrated prognostic potential for more objective classification of EC, providing an opportunity for improvement on current clinical management of patients[380,381]. The POLE-mutated subgroup have distinctly higher 5-year relative survival in comparison to the other subgroups, in particular the p53-mutant

subgroup[275]. It is not yet extensively implemented in clinical practice due to high costs and requirement of fresh or frozen tumour tissue[328,379,382,383]. Therefore, a surrogate solution to molecular classification of endometrial cancer has been proposed using a combination of targeted POLE sequencing and immunohistochemistry for MMR (mismatch repair) and p53 proteins[275]

Table 9: Molecular classification of endometrial carcinoma[275]

Subtype	Molecular status	Prognosis
POLEmut	Pathogenic POLE variants identified.	Excellent
MMRd	Mismatch repair deficient *POLE wt, non-pathogenic*	Intermediate
NSMP	p53 wt *POLE wt, non-pathogenic* *MMR-proficient*	Intermediate-excellent
p53mut	p53 mutant *POLE wt, non-pathogenic* *MMR-proficient*	Poor

1.6.6 Challenges in the Diagnosis of Endometrial Hyperplasia

> *"The trouble is that overdiagnosis and underdiagnosis are often intrinsically conjoined, locked perpetually on two ends of a seesaw."*[2] - Siddhartha Mukherjee, 2010

Despite improvements of the guidelines for the classification of endometrial hyperplasia, diagnosis is still based on a visual assessment. Therefore, it suffers from its inherent subjectivity. Although, D-score may improve prognostic assessment by objective, quantitative scoring, it is not widely available, and analysis is time-consuming. Accurate

prognosis is necessary to reduce the risk of over- or under-treatment of patients. The primary curative treatment remains hysterectomy for EAH, however, varying rates of malignant transformation should be factored into treatment decisions[384]. The endometrium itself is dynamic and undergoes cyclic shedding. This may result in a natural "removal" of abnormal lesions, where some patients will spontaneously regress[385,386]. Additionally, there is no *distinct* threshold that separates EHwA from EAH morphologically, and even EAH from EC; it is a continuous spectrum. Therefore, even for the trained eye, borderline cases are difficult to distinguish as one or the other. With the rise in obesity in the population and rising incidence of endometrial cancer, it is more important than ever to improve guidelines for treatment decisions, particularly in patients wishing to preserve fertility.

1.6.7 AI in Endometrial Carcinogenesis

A review article published in 2023, identified 30 studies that have utilised machine-learning in endometrial cancer[387]. Exploring the role of AI in refining the diagnosis and prognosis of endometrial carcinogenesis has covered several topics (listed below), in the areas of Radiology[388-393], Pathology[394-400], and pre-diagnostic screening[401-404].

- Detection and Diagnosis
- Predictive modelling and prognosis
- Risk assessment
- Decision-making support

The most frequently utilised algorithms are neural networks and SVMs[387]. Collaborations between experts in AI, oncology, and gynaecology will be crucial to advancing research in these areas and translating AI innovations into improved outcomes for individuals affected by endometrial cancer.

Exploitation of AI has potential in the development of diagnostic and prognostic support tools for classification of normal, EAH/EIN and EC[394-397]. Downing and colleagues developed a feature extraction tool and a random forest model to classify normal, precancerous (EIN) or malignant (type I EC) endometrium[394]. The most optimal model contained 75 variables. The top ten predictor variables consisted of fragment-level, gland-level, and lumen-level features[394]. The top predictor was percentage stromal surface area per fragment[394].

One study used automated feature extraction on HE stained WSIs, and a cox regression risk model for assignment of low- and high-risk scores for progression in endometrial cancer[399]. Two features were used in their final model: proportion of medium elongated stromal nuclei (ratio_bin5) and proportion of nuclei with medium crowded neighbours (distMean_bin6)[399]. In addition to providing accurate predictions of EC survival, the model contributed additional prognostic value to factors like tumour stage and grade[399]. Furthermore, the study reported association with features and mutation status in *TP53* and *TTN*. Alternatively, Makris et al. (2017) present an ANN model to discriminate between benign and malignant endometrial nuclei in cytological specimens[400]. Use of cytology represents a less invasive alternative for early detection of malignant disease.

Introduction of AI to a clinical workflow must present with notable advantages. The AI method must be an improvement on or equal to the current method, specifically with regards to patient diagnosis and treatment. If the model fulfils this requirement other benefits will aid the likelihood of adoption, for example a reduction in hands-on time or elimination of fatigue and recall bias. Although research shows AI as having the potential to improve patient diagnosis and treatment, it must still fit into the context of the clinic and not overcomplicate or make diagnosis inefficient for the clinician.

2 Aims

The overall objective of this thesis is to develop and evaluate biomarkers and AI-tools for use in breast cancer and endometrial carcinogenesis. Both diagnosis and prognosis of breast cancer and endometrial hyperplasia have potential for improvement. Subjective assessments due to visual evaluation of morphological criteria may be reduced through objective quantification of morphological features and biomarkers. This may lead to improvements in treatment allocation through improved accuracy of diagnostic classification and prognostic predictions. The works presented in this thesis endeavour to evaluate and validate biomarkers for the prognostic assessment of endometrial hyperplasia and breast cancer. Additionally, to develop and evaluate AI-tools for automated feature extraction and progression risk prediction in endometrial hyperplasia and quantification of proliferation biomarkers in breast cancer.

AIMS PAPER I

To explore the prognostic value of PAX2 and PTEN in predicting progression in endometrial hyperplasia and cancer. We aim to assess PAX2 and PTEN expression in normal proliferative, hyperplastic, and cancerous endometrial tissue and if manual scoring of these protein biomarkers is associated with progression.

AIMS PAPER II

To develop an automated feature extraction tool and risk assessment model for prediction of progression in endometrial hyperplasia.

AIMS PAPER III

To evaluate and develop automated quantification tools for Ki67 scoring in HR+/HER2-/LN- breast cancer. We aim to compare manual hotspot and global methods in addition to automated in-house tools using a commercial platform, a commercial CE-IVD tool and an in-house tool developed using open-source software.

AIMS PAPER IV

To validate the prognostic value of an automated mitosis detection tool, using deep learning, in several cancer types. To assess if mitotic counting can be generalised for use in many cancers, both where clinical usability is known and unknown.

3 Methodology

3.1 Ethical Considerations

All patient material used in the investigations conducted in this thesis complied with national ethical requirements. For **paper I, II and III** ethical approval was requested and granted by the Regional Committees for Medical and Health Research Ethics (REK), Western Health Region (REK vest). For **paper I** and **II**, the ethical approval was covered by the REK 2010/2464. For **paper III**, ethical approval was covered by REK 2010/1241. For **paper IV,** ethical approval was granted for each patient cohort and is outlined in their respective publications[405-414]. For **paper IV**, the Stavanger breast cohorts were covered by REK 2010/1241.

3.2 Patient Cohort

For **papers I-III**, archival FFPE material tissue from Stavanger University Hospital (SUH) was used. For **paper IV**, archival material from SUH for the breast cancer cohort was used for the prognostic validation dataset. Additionally archival material from 11 cohorts were used as described[414].

3.2.1 Endometrial Database

A retrospective patient cohort, consisting of 2671 patients with a diagnosis of proliferative endometrium, endometrial hyperplasia or endometrial cancer was available. For **paper I**, tissue with a primary diagnosis of proliferative endometrium, EIN and endometrioid carcinoma were eligible for inclusion (Figure 24). Inclusion criteria were as follows: available FFPE material (endometrioid carcinoma cases were available as TMA blocks) and at least one clinical follow-up which for

non-progression cases was required to be more than 6 months after primary diagnosis[415]. For **paper II**, tissue with a primary diagnosis of endometrial hyperplasia were eligible for inclusion (Figure 24). The following inclusion criteria were used: original diagnosis of endometrial hyperplasia, at least one follow-up sample (non-progression cases required to be more than 6 months after primary diagnosis), FFPE tissue block available, and a minimum area of 4.0mm^2 endometrial tissue available in the tissue section[398]. These cases (N=467) made up the development dataset, further details are outlined in **paper II**.

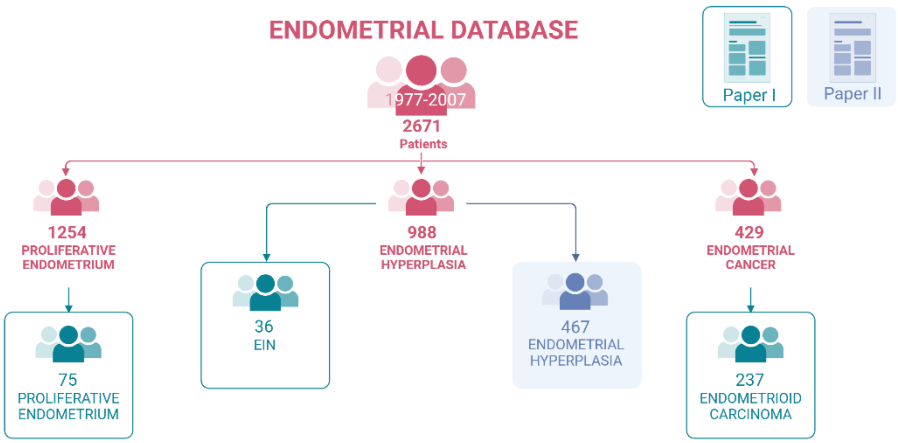


Figure 24: **Overview of the Stavanger Endometrial Database**, consisting of retrospective material. *Created using BioRender.com.*

3.2.2 Breast Cancer Database

For **paper III and IV**, patients who received a primary diagnosis of breast cancer in the years 1990-1998 and 2000-2004, at SUH, were eligible for inclusion in the study (N=687). As the purpose of **paper III** was to investigate Ki67 quantification and its prognostic value, the following inclusion criteria were used: <71 years of age, LN- status, HR+/HER2- status, and at least one clinical follow-up which for non-progression cases was required to be more than 6 months after primary diagnosis

(Figure 25). These cases (N=367) made up the development dataset, further details are outlined in **paper III**. For **paper IV**, inclusion criteria for entry to the study included if patients had FFPE material which allowed for creation of new HE stained tissue sections (Figure 25). These new sections, alongside the pseudonymised database, were sent to the Institute for Cancer Genetics and Informatics (ICGI), located at the Norwegian Radium Hospital in Oslo, Norway.

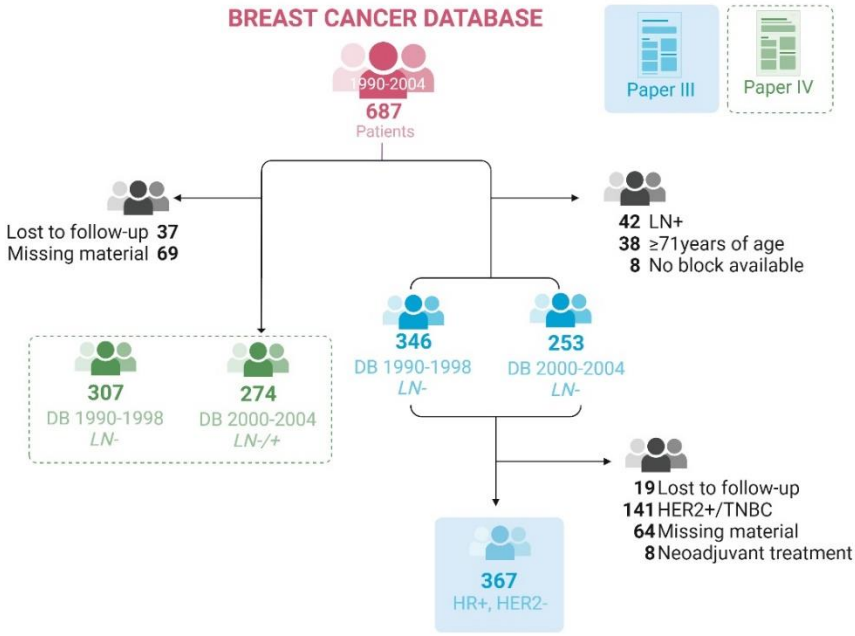


Figure 25: **Overview of the Breast Cancer Database**, used in this thesis. *Created using BioRender.com.*

3.3 Immunohistochemistry

Immunohistochemistry (IHC) is essential to and extensively used in pathology, particularly cancer[416]. The technique itself utilises antibodies to localise and bind cellular antigens in FFPE tissue[417]. De Matos *et al.* (2010)[418] outlines six scenarios where IHC is used in pathology:

1. Identification of cell types, cell secretions, organisms, and structures.
2. Discrimination of benign versus malignant cells
3. Histopathological diagnosis of tumours
4. Subtyping of tumours
5. Characterisation of primary site of a tumour
6. Prognostic assessment for therapeutic indication

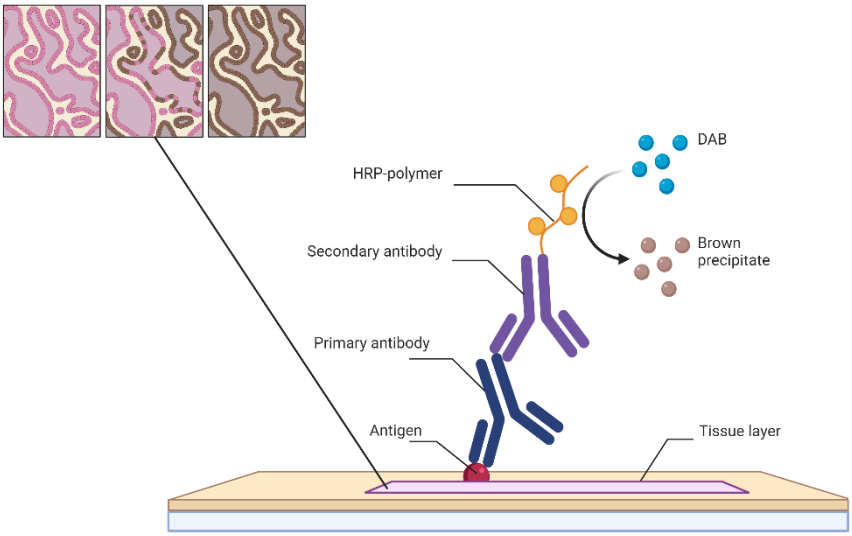


Figure 26: **Immunohistochemistry** staining using horseradish peroxidase and DAB. *Created using BioRender.com.*

The advantage of IHC is that the antigen-antibody complex is visualised as a colour signal[417]. Tissue sections are also counterstained with haematoxylin (blue stain) for clearer visualisation of tissue morphology[417]. Following fixation, embedding, and sectioning, the IHC staining process consists of several steps as outlined in Table 10 and visualised in Figure 26. The process of IHC is now mostly automated. Proper fixation is required for a good IHC stain; under- (<12hrs) or over-fixation (>48hrs) may lead to false negative results/weak stains[419]. Quality control is essential to ensure adequate IHC staining. This includes optimal antibody dilution and optimised protocols. Progressive

loss of antigenicity over time may occur in FFPE tissue, particularly if the tissue blocks are stored incorrectly such as in ambient or higher temperatures[418-420]. This is particularly prevalent for some markers and should be considered when using retrospective material. For most IHC stains, sections should be freshly cut, and the stained tissue should be quality controlled wherever possible.

Table 10: The IHC process.

Step	Description
	Deparaffinisation and Rehydration
1	**Antigen retrieval** This is used to "unmask" epitopes for antibody binding[417]. It involves the breaking of protein crosslinks that were introduced during fixation[419]. Common antigen retrieval techniques include heating sections in water or a buffered solution such as citrate or EDTA buffer[417].
	Blocking *A procedure used to block endogenous peroxidase*
2	**Primary antibody incubation** The tissue section is incubated with either a monoclonal or polyclonal antibody. The former targets a single epitope and is more specific, whilst a polyclonal antibody can bind to multiple epitopes and is more sensitive[419].
3	**Visualisation** Most modern methods utilise highly sensitive labelled polymer/multimer methods, which amplifies the signal. The tissue section is incubated with a polymer/multimer labelled with both the secondary antibody and enzyme molecules such as horseradish peroxidase (HRP).[421]
	Chromogen The tissue is incubated with a chromogen substrate, diaminobenzidine (DAB), these enzymes will produce a coloured product (brown)[417,419].
5	**Counterstain** The section is counterstained with haematoxylin (a blue nuclear stain) to visualise negative nuclei and morphology.

3.4 Conventional quantification of biomarkers

3.4.1 Mitotic Count in Breast Cancer

As described in *section 1.5.4.6*, the mitotic count is reported independently in the pathology report. Counting of mitoses is performed in areas with the highest mitotic activity, often at the periphery of the tumour[89]. These regions should be free of necrosis and inflammation, if possible. Only active mitotic cells should be counted. A score is then generated based on the number of mitoses counted in the field area (*see appendix 2*). Mitotic counting, at SUH, is now performed digitally on WSI using the local PACS. Counting is still performed manually but the observer can now use manual annotations, and the total field area can also be measured. The benefit of this method is that all annotations can be saved and reviewed later if necessary.

Prior to the establishment of the mitotic count, the mitotic activity index (MAI) was developed as a more standardised method for counting mitoses. Previously mitoses were counted in 10 high power fields (HPF) regardless of the diameter of the field of view (FOV), of which it has been criticised[422]. For MAI assessment, the number of mitoses is counted in the highest proliferative area in the periphery of the tumour, in *n* high power fields (400x) with a total area of 1.59mm^2. Areas with benign tissue, necrotic tissue or high inflammation are avoided. Only well-defined mitotic figures are counted[192,193]. The MAI was used in **paper III** and **IV** and counted with a microscope.

3.4.2 Ki67 Score in Breast Cancer

The Ki67 score is expressed as the percentage of positively stained invasive tumours cells in a defined region. Several methods exist for this quantification.

3.4.2.1 Hotspot Score

The hotspot score requires visual evaluation of the entire invasive region of the tumour section at low power for selection of the most proliferative region, the hotspot, for scoring. Once the most proliferative region is identified, both positive and negative nuclei are counted in a single FOV at high power (40X). A recommendation of 500 total tumour nuclei should be counted[118]. Usually one FOV is used, however, a consecutive FOV may be used if <500 nuclei are counted. This scoring method was used in **paper III**.

$$Ki67\ Score\ (\%Ki67) = \frac{\#\ of\ positive\ Ki67\ tumour\ nuclei}{total\ \#\ of\ pos.and\ neg.tumour\ nuclei} \times 100$$

3.4.2.2 Global Score

The global score method, developed by the International Ki67 in Breast Cancer Working Group (IKWG)[423], requires identification of the relative percentage of negative, low, medium, or high Ki67 expression in the invasive tumour region. The number of positive and negative tumour nuclei are then counted in a typewriter fashion, in up to four representative fields (up to 100 nuclei per field). These fields are assigned based on the relative percentage present in the tumour area. For example, if the entire tumour region is negative – all fields chosen should be negative. If 75% of the tumour is low and 25% is high, then 3 FOVs should be placed in a representative low region and 1 FOV in a representative high region. The unweighted and weighted global scores are then calculated. The IKWG provide a protocol and visual scoring tool on their website[423]. This scoring method was used in **paper III**.

3.4.3 PTEN Scoring in the Endometrium

There is currently no standardised method for the interpretation of PTEN IHC. A variety of methods are reported in the literature[424]. Such methods score PTEN according to complete loss of PTEN in the area of interest, presence of any PTEN-null gland, the percentage of PTEN-positive cells or the intensity of the PTEN staining[424]. Some methods result in a dichotomous classification whilst others in three or more groups[424]. Travaglino and colleagues investigated optimal scoring of PTEN and reported varying accuracy between methods, and concluded further study is required[424]. In **paper I**, for scoring of PTEN, we utilised the "presence of any PTEN-null gland" method, in accordance to a previous study by our group (Steinbakk *et al.* 2009)[425]. A lesion was scored as PTEN negative if one or more PTEN-null gland(s) (absence of nuclear or cytoplasmic PTEN staining) was observed in the lesion of interest. Stromal PTEN acted as a positive control[415].

3.4.4 PAX2 Scoring in the Endometrium

Like PTEN, there is currently no standardised method for PAX2 IHC scoring. In **paper I**, the H-score was used to quantify PAX2 protein expression. The H-score takes into account both intensity of the stain and the percentage of cells at each staining intensity. The percentage of glandular nuclei that display negative (0), weak positive (1), positive (2), strong positive (3) intensity is measured. This percentage is then multiplied with the respective staining score to generate a value between 0 and 300. To reduce subjective influence, two independent observers may score the lesion of interest, and if the H-score differs by 50, a consensus discussion is held. Alternatively, the mean score can be calculated. The former strategy was used in **paper I**[415].

3.5 Digital Image Analysis

3.5.1 Visiopharm

Visiopharm® was founded in 2002 by CEO Michael Grunkin and CTO Johan Doré Hansen in Denmark[426]. The company focuses on the design, development, manufacturing, installation, and service of AI-driven diagnostic pathology software[426]. The company provides both research solutions and IVDR cleared diagnostic algorithms.

The Visiopharm® image analysis platform (Visiopharm®, Hørsholm, Denmark) allows for the user to develop their own in-house algorithms using the APP author. The user can design their own classification algorithms, measurements and scoring of images. The APP author consists of several components:

1) Training labels (image classes for training),
2) Magnification and region of interest specifications
3) Classifier method (AI - Deep Learning, Bayesian, K-means, Decision Forest, etc.), colour channel (RGB-R, RGB-G etc.), and filters (mean, median etc.),
4) Post-processing (morphological operators such as dilate, erode, etc.)
5) Output variables (measurement of segmented features and simple calculations)

An in-house application (APP) can be designed according to the desired endpoint and multiple APPs can be developed and run in tandem. For example, one APP could demarcate the tissue region, which would exclude background and background noise from later analysis. Another could perform a segmentation task within the demarcated tissue, and a final APP could perform measurements of these segments. This design is a trademark feature of Visiopharm®. Its advantage lies in the ability

to run or rerun parts of the APP individually, to identify sources of error or to correct for any mistakes without having to rerun the entire process, which can be time-consuming and require significant computer processing power. A basic illustration of our Visiopharm® workflow for APP development is shown in Figure 27.

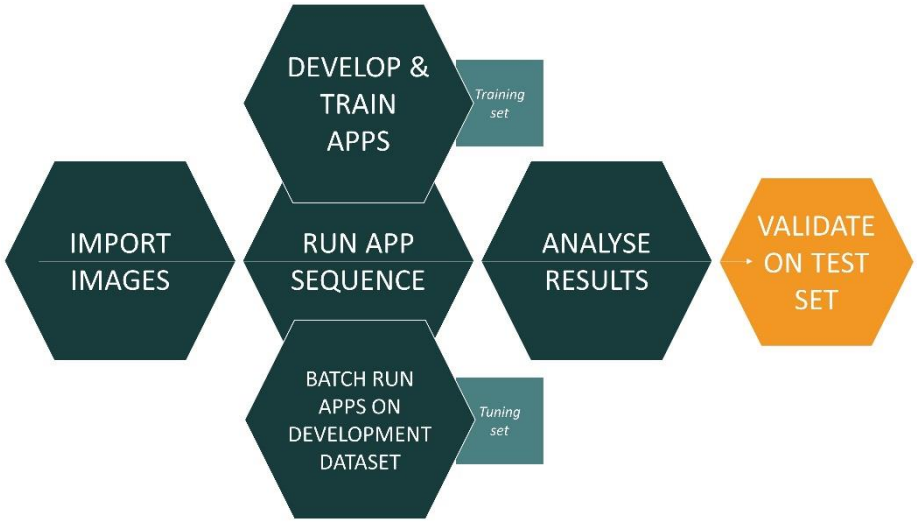


Figure 27: **The workflow used for in-house APPs developed in Visiopharm®.**

3.5.1.1 The ENDOAPP

The ENDOAPP sequence was designed, developed, and trained in-house using the Visiopharm® system (Version 2020.08 Visiopharm A/S, Hørsholm, Denmark). It consists of several independent APPs that are designed to be run in tandem. A description of each APP in sequence is provided below and in Figure 28:

- 01 Tissue detection
- 02 Gland, stroma, and lumen segmentation
- 03 Hotspot detection
- 04 Architectural feature extraction
- 05 Nuclei segmentation pre-processing

- 06 Epithelial nuclei segmentation
- 07 Nuclei feature extraction.

Table 11: Overview of training criteria, type of classifier, post-processing requirements and if the APP included a calculation step for each APP in the ENDOAPP.

APP	OUTPUT	TRAINING / ANNOTATIONS	CLASSIFIER	POST-PROCESSING	CALCULATIONS
01	Tissue ROI	✓ *Tissue + background*	K-means clustering	✓	×
02	Gland, stroma and lumen segmentation	✓ *Gland, lumen, and stroma*	K-means clustering	✓	×
03	Heatmap and hotspot ROI placement	×	×	✓	×
04	Measurement of % stroma	×	×	×	✓
05	Gland ROI	×	×	✓	×
06	Epithelial nuclei segmentation	✓ *Nuclei*	K-means clustering	✓	×
07	Measurement of shortest diameter of epithelial nuclei	×	×	×	✓

For each APP that required training labels (01, 02, 06), manual annotations were made on representative structures (Table 11). The APPs were trained using these annotations from the training dataset. Annotations were added iteratively, and the APP retrained, until the result was considered satisfactory. A satisfactory result was decided by the operator and based on sufficient segmentation of features. For example, for APP 02 glands were required to be correctly labelled and clearly distinguishable from surrounding stroma. Additionally post-processing steps were added to further enhance segmentation. Following training, the APPs were run on the tuning set and performance evaluated.

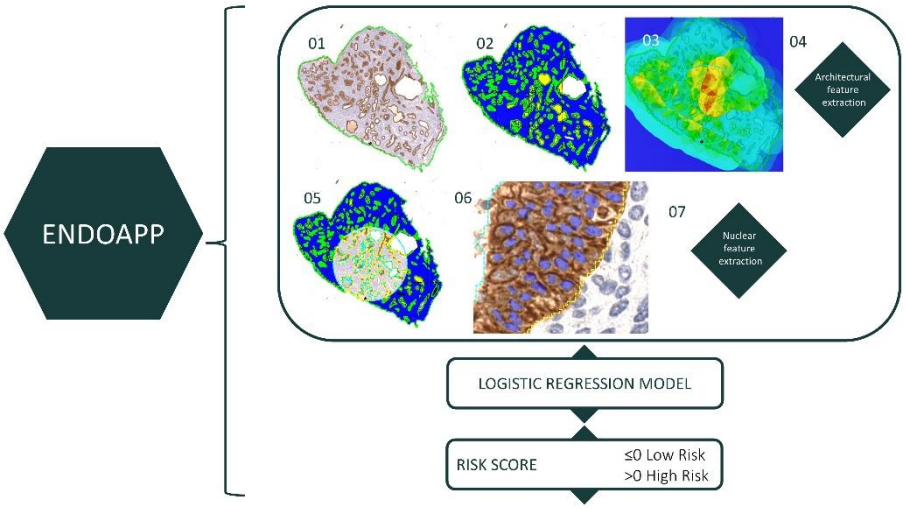


Figure 28: **The ENDOAPP**. Seven APPs were developed in Visiopharm® for segmentation and feature extraction. These features were input into a logistic regression model to generate a risk score for progression.

Using output data from the development dataset, a logistic regression model was used to generate a risk score, and a threshold for risk of progression (low or high) was determined using ROC curve analysis. The logistic regression model was separate to the Visiopharm® APP sequence. The output file from Visiopharm®, containing feature extraction data (the area of all labels in the hotspot ROI, % stroma, and the lesser diameter of each epithelial nuclei label), was entered into the model. Further information on the model can be found in **paper II**.

3.5.1.2 The In-House Ki67 APP (VIS1-HS)

A Ki67 quantification APP sequence was designed, developed, and trained in-house using the Visiopharm® system (Version 2020.08). Like the ENDOAPP it consists of several independent APPs that are designed to be run in tandem. A description of each APP in sequence in Table 12 and in Figure 29:

- 01 Tissue detection
- 02 Positive nuclei detection
- 03 Hotspot detection
- 04 Tumour segmentation
- 05 Nuclei segmentation
- 06 Hotspot quantification

For each APP that required training labels (01, 02, 04, 05), manual annotations were made on representative structures (Table 12). Annotations were added iteratively, and the APP retrained, until the result was considered satisfactory. A satisfactory result was decided by the operator and based on sufficient segmentation of features. For example, for APP 05 positive and negative nuclei needed to be correctly labelled and clearly distinguishable. Additionally post-processing steps were added to further enhance segmentation. Following training, the APPs were run on the tuning set and performance evaluated.

Table 12: Overview of training criteria, type of classifier, post-processing requirements and if the APP included a calculation step, for the in-house Visiopharm® APP sequence (VIS1-HS).

APP	OUTPUT	TRAINING	CLASSIFIER	POST-PROCESSING	CALCULATIONS
01	Tissue ROI	✓ *Tissue and background annotations*	K-means clustering	✓	×
02	Positive nuclei segmentation	✓ *Positive nuclei annotations*	K-means clustering	✓	×
03	Heatmap and hotspot ROI placement	×	×	✓	×
04	Tumour segmentation	✓ *Tumour and non-tumour area annotations*	K-means clustering	✓	×
05	Nuclei segmentation	✓ *Positive and negative nuclei annotations and background*	K-means-clustering	✓	×
06	Hotspot Ki67 Score	×	×	×	✓

For positive cell detection (APP 02) negative nuclei (tumour) were not segmented as this extended analysis time 10-20-fold. As a result, only negative and positive nuclei were segmented in the hotspot ROI. Therefore, a global score was not generated for the in-house APP as the APP took too long to run and often crashed. The VIS1 APP runtime was between 30s to 3 minutes.

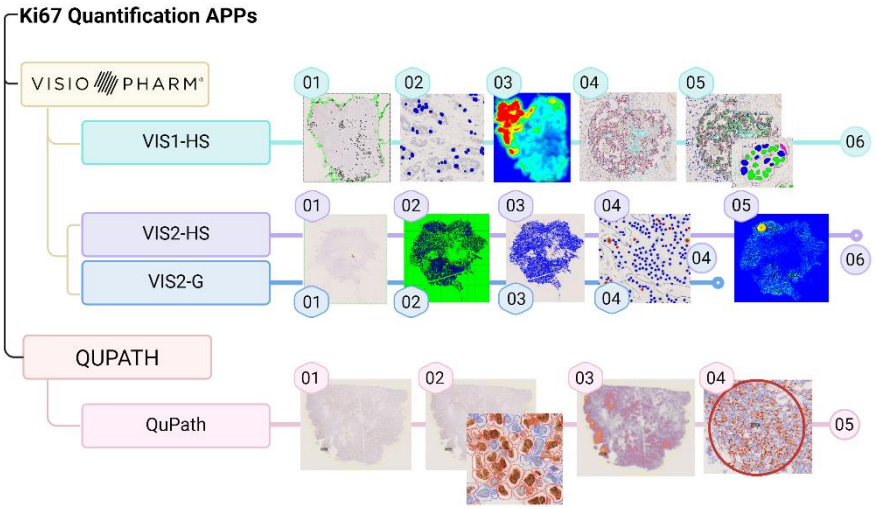


Figure 29: **The Ki67 applications**. Visual overview of the Visiopharm® APPs and QuPath workflows for Ki67 quantification.

3.5.1.3 The Visiopharm® Ki67 APP (VIS2-HS/G)

The Visiopharm Ki67 APP is CE-IVD approved, and consisted of several independent APPs that are designed to be run in tandem. A description of each APP in sequence is provided below and in Figure 29. The sequence for global scoring (VIS2-G) utilised APPs 01-04, whilst the sequence for hotspot scoring (VIS2-HS) utilised APPS 01-06.

- 01 #10182 – IHC Tissue Detection, AI
- 02 #10180 – Invasive Tumor Detection, AI
- 03 #10180 – Invasive Tumor Postprocessing
- 04 #10173 – Ki-67 Nuclei APP, Breast Cancer, AI
- 05 #10114 - Hot Spot Detection
- 06 #10114 - Hot Spot Quantification

The APPs were provided free-of-charge for research purposes, as outlined in **paper III**. The APPs were pre-trained/validated and approved for Ki67 scoring in breast cancer.

Table 13: Overview of the type of classifier and if the APP included a calculation step, for the commercial Visiopharm® APP sequence (VIS2-HS/G).

APP	OUTPUT	CLASSIFIER	CALCULATIONS
01	Tissue ROI	DeepLabv3	✖
02	Invasive tumour ROI	DeepLabv3	✖
03	Invasive tumour postprocessing	✖	✖
04	Nuclei segmentation. Global Ki67 score	U-Net	✓
05	Hotspot detection	✖	✖
06	Hotspot Ki67 Score	✖	✓

3.5.2 QuPath

QuPath is an open-source, biomedical image analysis platform, designed for digital pathology[249]. It's development aimed to meet the growing need of the community for a free, user-friendly image analysis software that could handle complex texts, without the need of an IT background[249]. A PubMed search, "QuPath", revealed 204 publications and the pioneer publication from 2017 has been cited over 4000 times, according to Google Scholar (as of January 2024). The software offers both unsupervised ML-based cell detection and supervised classification of WSIs[427].

3.5.2.1 The QuPath Ki67 Tool (QuPath)

Several studies have published on the use of QuPath for quantitation of Ki67 in breast cancer[428-432]. The IKWG also link to a QuPath protocol for Ki67 evaluation on core biopsy WSIs on their website[423]. For this thesis and as described in **paper III**, a cell classifier script was created using the cell detection and cell classifier tools in QuPath (*Appendix 3*). Annotations of positive and negative tumour nuclei from the breast training dataset were used to train the classifier. It was possible to save the code/script of the most optimal model. The script was run on all

cases in the development dataset. The workflow is outlined below and illustrated in Figure 29.

- 01 Manual delineation of the invasive tumour region
- 02 Nuclei cell classifier script.
- 03 Density heat map applied for automated placement of hotspot ROI.
- 04 The hotspot ROI is dynamic – it can be moved and the Ki67 hotspot score is updated automatically.
- 05 Ki67 score was generated automatically.

3.5.3 The Mitosis Detection Tool

A mitosis detection model was validated in **paper IV**. Deep learning models (Mask R-CNN neural networks) were previously trained and tuned using annotated mitotic figures from the TUPAC16 dataset[235]. This dataset contains 73 WSIs of HE stained breast cancer tissue sections, scanned by the Aperio ScanScope XT and Leica SCN400 scanners (Leica Biosystems, IL, USA). Ground truth annotations were created by consensus agreement of at least two pathologists. The Mask R-CNN (regional-based CNN) is an object segmentation network. Essentially the network will take an image and detect objects (objects are marked with a bounding box) of which it will assign a probability that the object is a mitotic figure. A detailed description of how the network was trained, tuned, and validated can be found in the supplementary material for **paper IV**. The models were validated on WSIs from a total population of uterine sarcomas[433]. The best performing model was selected for prognostic validation (**paper IV**). For prognostic validation of the model, 13 cohorts representing different cancer types (breast, bladder, prostate, colorectal, lung, liver metastases and endometrial) comprised the dataset. The tumour region was annotated manually prior to model prediction. The number of mitotic figures detected by the model, in the entire tumour tissue,

was divided by the total area of the tumour to calculate the number of mitoses per mm^2.

3.6 Statistical Analysis

All statistical analyses, for **paper I-III**, were performed using SPSS for windows (version 26.0.0; IBM SPSS Statistics) and R Studio (2022.02.0 Build 443). For **paper IV**, statistical analyses were performed using R (version 3.5.2). Assumptions for each statistical test were considered and verified. A p-value of <0.05 was considered as significant.

3.6.1 Agreement

The intraclass correlation coefficient (ICC) was used to measure agreement. In **paper II**, the ICC was used to assess agreement between manual operators of the ENDOAPP, between scanners, and between classification methods and the ENDOAPP. In **paper III**, the ICC was used to assess agreement between manual and DIA scoring methods. The guidelines proposed by Koo and Li (2016)[434] for interpretation of the ICC were used, as depicted in Table 14. Cohens κ was used to assess the agreement of the D-score against all other methods, using categorical variables, in **paper II**.

Table 14: Interpretation of the Intraclass Correlation Coefficient (ICC) as suggested by Koo and Li (2016)

ICC	Interpretation
<0.50	Poor
0.05 – 0.75	Moderate
0.75 – 0.90	Good
>0.90	Excellent

3.6.2 ROC Curve

A receiving operating characteristic (ROC) curve was created to assess the discriminative ability of the ENDOAPP in **paper II**, and each Ki67 scoring method in **paper III**. The area under the curve (AUC) ranges from 0.5 – 1.0 and indicates the model's ability to discriminate between individuals with or without the outcome of interest[76]. The ROC curve was used define cut points for binary classification of individuals. For **paper II,** the intention was to assign a cut-off for patients with low or high-risk for progression to endometrioid carcinoma. For **paper III**, the intention was to evaluate cut-offs for low- and high-risk Ki67. Different cut-offs will also affect the trade-off between sensitivity and specificity[76]. A general guideline for interpretation of the AUC has been described by Hosmer and Lemeshow (2013)[76] (Table 15).

Table 15: Guidelines for interpretation of the AUC statistic in ROC curve analysis, defined by Hosmer and Lemeshow (2013)[76].

AUC	Interpretation
=0.50	No discrimination
0.05< AUC < 0.7	Poor discrimination
0.70 ≤ AUC < 0.8	Acceptable discrimination
0.80 ≤ AUC < 0.9	Excellent discrimination
≥0.9	Outstanding discrimination

3.6.3 Performance Metrics

Performance metrics were calculated to assess the prognostic value and predictive capacity of classification methods in **paper II** and **III** and are described in Table 16. For a more detailed description of each metric see Table 2.

Table 16: Performance Metrics

Metric	Acronym	Calculation
True positive	TP	
True negative	TN	
False positive	FP	
False negative	FN	
Positive predictive value (Precision)	PPV	$= \frac{TP}{TP + FP}$
Negative predictive value	NPV	$= \frac{TN}{FN + TN}$
Sensitivity (Recall)	Se.	$= \frac{TP}{TP + FN}$
Specificity	Sp.	$= \frac{TN}{FP + TN}$
Accuracy		$= \frac{TP + TN}{N}$
		$= \frac{prev. \times se.}{(prev. \times se.) + ((1 - prev.) * sp.)}$

3.6.4 Survival and Multivariate Analysis

For **paper I-IV**, Kaplan Meier analysis was used to determine association between variables and progression-free survival. Significant differences were analysed by the log-rank test. The endpoint for **paper I and II** was diagnosis with progression (see Table 17) in the follow-up (event) or censored to the last known follow-date or date of hysterectomy. Patients without progression in the follow-up were required to have a minimum of six months follow-up. The endpoint for **paper III and IV**, was the first diagnosis of a distant metastasis in the follow-up, or as otherwise specified in **paper IV** for the non-breast cancer types.

Patients with no distant metastasis in the follow-up were censored to last known follow-up and were required to have at least six months between this date and primary diagnosis date. Hazard ratios were generated using cox regression survival analysis for **paper III and IV**.

Table 17: Kaplan Meier survival analysis endpoint (event) for endometrial cases in paper I and II.

Primary diagnosis	Event
Proliferative endometrium	Endometrial hyperplasia or endometrioid carcinoma
Endometrial hyperplasia	Endometrioid carcinoma
Endometrioid carcinoma	Distant metastasis or death due to disease.

3.6.5 Univariate and Multivariate Analysis

Logistic regression analysis was used to assess the relationship between one predictor variable (univariate) or several predictor variables (multivariate) and a binary outcome (progression or no progression) in **paper II**. Furthermore, the formula for the most optimal model, determined by the backward wald method, was used to establish the risk score formula for the ENDOAPP (Figure 28).

Cox regression survival analysis was not only used to report hazard ratios but also to assess the relationship between one predictor variable (univariate) or several predictor variables (multivariate) and a binary outcome (progression or no progression) in **paper II-IV.** One of the main differences between logistic regression and cox regression, is that the latter factors time to progression into the analysis.

4 Summary of the Papers

4.1 Paper I: Assessing the prognostic value of PAX2 and PTEN in endometrial carcinogenesis.

Current methods for diagnosis of endometrial hyperplasia (EH) are based on subjective visual assessments. The EIN classification, incorporated into the WHO guidelines in 2014[306], has demonstrated improved reproducibility and prognostic value of diagnosis in comparison to previous classifications[304,307,323,435]. However, the WHO20 classification, is still based on a qualitative assessment. Therefore, use of prognostic biomarkers may improve the stratification of low- and high-risk patients. Loss of PTEN and PAX2 expression has been demonstrated in endometrial carcinogenesis and may be an interesting target for prognostic assessment[355,358,368,436].

In **paper I** we investigated if the quantification of PTEN and PAX2 can predict progression free survival in endometrial intraepithelial neoplasia (EIN) and endometrial endometrioid carcinoma (EEC). The patient cohort (N=348) consisted of endometrial cases diagnosed between 1977-2004 at SUH, with proliferative endometrium (N=75), EIN (N=36) and EEC (N=237). Long-term clinical follow-up was available (6-310 months, median: 126). No progression to EEC was observed for proliferative cases, whilst 28% of EIN cases had progression (N=10). We evaluated PTEN and PAX2 expression detected by IHC. A case was classified as PTEN null when absence of staining was observed in one or more glands. PAX2 expression was quantified by H-score. Absence of PTEN expression was more common in EEC than EIN and proliferative endometrium (64% vs. 11%, respectively). A progressive decrease in PAX2 expression was observed from proliferative to EIN to EEC. We investigated if binary categorisation of PAX2 in EIN cases could predict

progression-free survival. Notably, patients with a high-risk score (H-score ≤ 75) had a significantly lower progression-free survival than those with a low-risk score. PAX2 expression was not prognostic for EEC progression. Nor was PTEN expression prognostic in EIN or EEC.

To conclude, we were unable to validate the prognostic value of PTEN. The literature reports conflicting results, some studies associated PTEN positive status with favourable prognosis[437,438], whilst others suggest PTEN negative status is indicative[439,440]. On the other hand, we reported that quantification of PAX2 expression in EIN had prognostic value for predicting progression-free survival. However, the H-score method is still based on a subjective visual assessment. Therefore, we considered this investigation as a pilot study and planned to verify our findings in a larger cohort and investigate the use of DIA for evaluation.

4.2 Paper II: Automated Prognostic Assessment of Endometrial Hyperplasia for Progression Risk Evaluation using Artificial Intelligence

Quantitative measurement of the morphological features used in the diagnosis of endometrial hyperplasia (EH) has demonstrated prognostic value[324,441].

In **paper II** we developed a tool (ENDOAPP) using AI for the automated measurement of morphologic and cytological features of endometrial tissue and quantification of progression risk score in EH. The development dataset consisted of whole slide images of PAN-CK+ stained tissue sections, from 388 patients diagnosed with endometrial hyperplasia between 1980 and 2007. Follow-up data was available for all patients (mean: 140 months). The ENDOAPP was developed using Visiopharm® and was designed to detect the region with the highest density of glands, and then to measure percentage stroma and

epithelial gland nuclei variation in this region. A logistic regression model was developed using the recommended predictor variables to generate a progression risk score. The ENDOAPP had acceptable discriminative power (AUC: 0.765). Furthermore, it had the highest accuracy alongside the semi-automatic method D-score (88-91% vs. 91%, respectively), and marginally outperformed the WHO94, WHO20 and EIN clinical classification schemes. Additionally, we observed that inter-observer agreement was strong to moderate for manual operators of the ENDOAPP (ICC: 0.828), and moderate between scanners (ICC: 0.791).

In summary, we observed comparable performance metrics for the ENDOAPP method compared to the D-score[299,319-321,327]. Notably, we reported comparable and improved performance metrics in comparison to traditional visual assessments[327,442]. Interobserver variability and inter-scanner variability was comparable to interobserver variability of traditional classification schemes[305,308,435]. We therefore propose that AI-based tools demonstrate potential for more objective, prognostic evaluation of endometrial hyperplasia. Furthermore, we need to validate our findings on a test set, preferably temporally and geographically distinct from the development set.

4.3 Paper III: The Ki67 Dilemma: Investigating Prognostic Cut-Offs and Inter-Platform Reproducibility for Automated Ki67 Scoring in Breast Cancer

New technological advances and the deployment of digital pathology allows for the development and implementation of AI-based scoring tools in the clinic. Quantification of Ki67 score in breast cancer has documented prognostic and predictive value[197-201]. Therefore, it is a primary target for automated scoring on a digital platform.

In **paper III**, we evaluated the agreement between manual (hotspot and global) and respective DIA scoring methods. Furthermore, we explored cut-offs for stratification of low and high Ki67 score. We aimed to address the concern whether cut-offs that are prognostic using manual methods remain prognostic using automated, DIA methods. From a cohort of 367 HR+/HER2-/LN- breast cancer patients, 294 were eligible for analysis. The 73 cases removed from the study cited issues relating to poor staining quality. Manual scoring of invasive tumour regions in a hotspot or by global unweighted/weighted scoring was performed. Additionally, a commercial and an in-house algorithm, developed on the commercial platform Visiopharm®, were used for DIA quantification of Ki67. An open-source (QuPath) algorithm was also developed. Reproducibility analysis revealed good agreement between manual methods and their respective DIA methods (ICC>0.8). The DIA methods had better discriminative ability than the manual methods. A range of cut-offs were prognostic in the study cohort. However, the use of a single cut-off for all scoring methods demonstrated differences in distributions of prediction outcomes (e.g., false negative and positives).

In conclusion, we reported strong agreement between platforms and manual scoring methods. Automated scoring methods outperformed manual methods, although all were prognostic. These observations were comparable to other studies[206,219,220,429,430,443]. However, we report the need for caution regarding adoption of a single cut-off for different methods as this can affect the number of over and under-treated patients. As current guidelines also recommend molecular profiling for treatment decisions, we would need to verify our results alongside these risk scores. Particularly if automated, digital Ki67 scoring is useful in the intermediate risk category. Furthermore, the in-house tools require validation on a test set.

4.4 Paper IV: Applicability of mitotic figure counting by deep learning: a development and pan-cancer validation study.

Quantification of mitotic figures is used to measure proliferation in cancer and is incorporated in the grading system for breast cancer. Mitotic figures have been assessed manually but deep learning algorithms are being investigated for the automation of this task.

In **paper IV**, a deep learning method (Mask R-CNN), for automated mitosis detection on WSIs, was validated for its prognostic impact in several cancer types. The network was trained on images from the TUPAC16 breast cancer database[235]. The model was validated on a test set consisting of 372 uterine sarcomas, and the best performing model was selected for further validation. To assess prognostic impact and generalisability, the deep learning model was validated on 14 571 patients, from 13 cohorts, with cancer samples from the prostate, breast, lung, endometrium, colon, rectum, and colorectal liver metastases. The prognostic value of the mitotic count (number of mitotic figures per mm^2) was evaluated using cox regression analysis. The mitotic count was prognostic in breast, prostate, endometrial, lung cancer and liver metastases from colorectal cancer. It was not prognostic in colorectal cancer.

In summary, the use of automated mitotic figure counting using a deep learning model, has potential to be used for the prognostic evaluation of mitotic count in several cancers. The method was prognostic in several cancer types which currently do not routinely perform mitotic counting including prostate and bladder cancer. To conclude, we present an automated method that has potential as an alternative to manual counting.

5 Discussion and Future Perspectives

> *"Medical breakthroughs take time… If we start today, and seize this moment, and the focus and the energy and the resources that it demands, there is no telling how many lives we could change. And every single one of those lives matter."*[88] - Former President of the USA, Barack Obama, 2015

5.1 The Future of PTEN and PAX2 as a Biomarker

Research many times over, has demonstrated the prognostic relevance of PTEN and PAX2 in endometrial carcinogenesis[296,334,352-358,366,368-370]. However, their expression varies in normal, hyperplastic, and neoplastic endometrium, to the extent that their value as a quantifiable biomarker is controversial[334,352,355,359-361,372,373]. It seems pinpointing the exact moment in which their expression changes and accurately associating it with one of two outcomes – progression or no progression, remains a challenge.

In **paper I**, PTEN scoring was not significant for prediction of progression of endometrial hyperplasia and endometrioid cancer. This may be attributed to the choice of scoring method. Despite this we still observed decreasing expression from normal and hyperplastic endometrium to neoplasia. Travaglino and colleagues discuss differences in prognostic accuracy of PTEN scoring methods[424]. The authors determined that the scoring method we used in **paper I**, presence of any null gland, had the lowest accuracy of the methods investigated. It was reported to have a low sensitivity (0.57), low specificity (0.55), and low overall accuracy (AUC=0.63)[424]. Therefore, one might presume that perhaps another method would yield a more accurate quantification. In comparison, the only methods Travaglino and colleagues report, with moderate diagnostic accuracy (AUC=0.78),

were intensity scoring methods which classified weak-to-null expression[424]. One might postulate that a different scoring method, such as this, would yield a different result for the evaluation of PTEN in the study cohort. However, literature suggests that current scoring methods for PTEN are still not sufficient to suggest its value as an independent predictor of progression or response to conservative treatment[334,355,359-361].

Unlike PTEN, in **paper I**, we observed a significant difference between progression and non-progression EIN patients for PAX2 expression quantified by H-score. These results are promising and indicate its potential as a biomarker in EAH. However, the study cohort used in **paper I** was small, consisting of only 36 EIN patients. It would be worthwhile to validate these findings in a larger cohort. Small cohorts run the risk of producing false positive or false negative results due to bias or overfitting[444]. Furthermore, although valuable for exploration or pilot studies, such as ours, they cannot completely reflect trends in a population. Therefore, validation studies in larger and new cohorts are essential for biomarker investigation. Unfortunately, when we investigated PAX2 expression, by H-score, in a larger cohort, we found no association with progression-free survival in EIN (*unpublished*).

Although, PTEN and PAX2 play a role in endometrial carcinogenesis, their role as biomarkers for improved diagnosis and stratification of patients with endometrial hyperplasia appears limited. However, perhaps evaluation of PTEN and PAX2 by AI could reveal new insight. Whilst manual scoring methods have been limited to human capability, AI has the power to overcome flaws such as recall bias and fatigue. It can also enhance analysis by combining the assessment of morphological features on HE with IHC-quantification of biomarkers. This may reveal new avenues not yet investigated. If could also be combined with molecular analysis or be used to predict molecular

subtypes on HE. To conclude, although a standstill may have been reached for manual evaluation of PAX2 and PTEN, new technology might yet uncover patterns unseen to us at present. For example, it may be more prudent to use molecular data to assess these biomarkers or even use AI to predict biomarker expression patterns directly on HE, avoiding IHC altogether. In turn we may find new applications for the use of PTEN and PAX2 as clinical biomarkers, perhaps we have merely lacked the tools to sufficiently evaluate them.

5.2 Validation of AI as a Diagnostic Assistance Tool

5.2.1 Validation and Verification

Validation and verification is a continuous process, and should be conducted prior to implementation and throughout the lifecycle of a system, product, software, or tool. The purpose of this is to identify any discrepancies or deviations from a product's intended use or specifications. Validation involves testing a product to see whether it performs as intended, complies with predefined requirements, and aligns with the parameters established within the medical context. Verification requires confirmation that a product behaves as expected and fulfils its objectives in the environment it was intended for. The validation process is usually more extensive than verification. For example, a newly developed algorithm must be validated on a test set, whereas a CE-IVD approved algorithm must be verified when first deployed at a new site. Validation or verification may be required following changes to the model's design, such as a software update, or when new variables are introduced, such as a new scanner.

Validation of an algorithm requires testing that the model can be generalised on relevant populations, outside of the development dataset, from different locations and using different equipment such as

scanners or staining/fixation procedures[445]. A rigorous performance evaluation is recommended to assess if the algorithm still functions under a variety of conditions, that may not be covered by a test set. This is often performed in an external validation, often in a multi-centre trial. This requires a broad validation of multiple cohorts that represent the intended target population with respect to age, sex, ethnicity, prevalence and additionally samples taken from a range of laboratories also their own technical equipment and protocols[445]. Prospective trials are also a way to assess the generalisability of a tool, by assessing real-world data in real-time. Validation studies are vital to uncover any bias or overfitting in the model. This also applies to biomarker validation studies.

In **paper II**, we have developed the ENDOAPP and reported a promising performance for the prognostic evaluation of endometrial hyperplasia. However, the APP was not validated on a test set. Therefore, its clinical potential and performance is limited to the development dataset and no substantial conclusions can be made. Many factors will need to be considered and a multi-stage validation study will be required before considering clinical implementation. Firstly, only two scanners were evaluated, whereas many exist on the market. We reported some variation between scanners, which although not significant, still affected the proportion of false negative and false positive cases. The scanners used were produced by the same manufacturer, and therefore use the same image file format: ndpi. It would be prudent to test the ENDOAPP on other WSIs produced by other scanners. Furthermore, to test different staining protocols and antibodies. In paper II, we only used one antibody and one staining protocol. Staining variation is one of the most important factors to keep in mind when evaluating model performance. We are currently in the early stages of a validation study, which will utilise the staining and scanning protocols of four hospitals in the western health region of Norway. Additionally, an external dataset

from Colombia is currently being collated by our collaboration partner and we are also collecting new samples locally for a test set for validation of the ENDOAPP on more recent samples. Another avenue to be explored for future work would be to investigate a combination of the ENDOAPP with molecular analysis of biomarkers/expression profiles. Developing an AI model for prediction of progression or diagnostic classification on HE, using a deep learning model, would also be a worthwhile pursuit.

For **paper III**, we have compared in-house developed and commercial CE-IVD tools. We demonstrate good agreement between the methods, but ideally would need to further validate the tools on a larger cohort representing different scanners and staining protocols. Commercial algorithms for the quantification of Ki67 in breast cancer have been validated and certified for this purpose. Most of these tools appear to use pathologist annotations or score as ground truth, the level of agreement between the algorithm and annotations/pathologist score is assessed. How this validation data is reported varies and it is not always easy to access. To note, the cut-offs used have not been validated, and it is this that affects diagnostic and treatment decisions. These still require considerable validation and standardisation and this was the focus of **paper III**. The study cohort used in **paper III**, is however, not sufficient to consider a cutoff suitably validated. Further research in prospective trials and multi-centre validation studies are required to elucidate if cut-offs currently in use for manual methods are transferable to automated methods, and if they are transferable between different platforms.

Planning a validation study, one should also consider the desired clinical context: is the model designed to be used on a local, regional, national, or international scale? In-house (local) validation could require fewer test variables. If the goal is to implement the APP locally, then local

factors such as the scanner and staining protocol used at the local site should be sufficient for validation. In comparison, if the APP is intended to be used on a larger scale, national approval, CE-IVD, or FDA certification would be required. This is a much more extensive and time-consuming approach. To choose in-house validation vs. multi-centre validation is also unfortunately dependent on the resources available. Regardless, choice of in-house or multi-centre validation *must be ethically permissible*.

Regulations set by the FDA (USA) and EU (European Union) define a diagnostic AI-assistance tool as a medical device also termed a "medical device software", "health software" or "software as a medical device". A medical device software is "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device" as defined by the International Medical Device Regulators Forum (IMDRF)[446]. Strict protocols outline the approval and use of software medical devices. This is favourable for ensuring good clinical practice and patient safety; however, there may be negative impacts, for example the cost of approving a device for the company or public institutions (funded by taxpayers), delays in implementation and so on[447]. However, stringent regulations are required, as the quality of evidence surrounding some devices have been criticized. This includes lack of external validation, multi-centre evaluation, majority are based on retrospective studies which run the risk of biased data, heterogeneity in performance metrics used and lack of availability of datasets allowing for replication and/or extensive validation[447]. Furthermore, the in-house exception rule requires certain requirements to be met and no equivalent CE-IVD certified product must be available on the market.

Regarding the certification of AI tools for clinical use, a study by Wu and colleagues aimed to shed light on how these tools were being assessed

and evaluated prior to approval by the FDA (2015-2020)[448]. The authors reported that 72% (93 of 130) did not publicly report multi-site evaluation and for 55% (71 of 130) of devices, the median evaluation sample size was 300. Although the FDA has access to the number of sites evaluated and sample size, this is not always publicly available which makes it difficult for a technician or clinician to critically evaluate how the AI tool has been trained and validated[448].

The outcome of an AI tool will also influence the extent to which the tool will require validation. For example, tools where the outcome will decide treatment decisions might require more extensive validation than a tumour vs. non-tumour segmentation tool. Ultimately, patient safety is most important, and therefore how an AI tool is developed and validated *must* be sufficient, transparent, and comply with good clinical practice and regulatory guidelines.

5.2.2 Retrospective vs. Prospective Datasets

Retrospective datasets are valuable in that the data is readily available, in comparison to prospective datasets. However, established retrospective databases are more likely to have been collated for a different purpose than to train AI. Therefore, there is a higher risk of systemic bias in these datasets. The U.S. Food and Drug Administration (FDA) in a 2021 action plan, recognised that AI/ML algorithms are vulnerable to the bias already present in historical datasets and healthcare systems[449]. Furthermore, they address the need to identify and eliminate such bias in order to ensure robustness and generalisability of these algorithms over time[449]. Interestingly, in a study investigating FDA-approved AI devices, approved between 2015-2020, almost all the devices were evaluated in retrospective studies (126 of 130)[448]. In Norway, many AI tools in radiology, prior to implementation, are being validated in prospective clinical trials.

5.2.3 Considerations on Technology

AI diagnostic assistance tools in pathology are intended to analyse either WSIs of HE stained tissue or IHC stained tissue. In **paper II**, the ENDOAPP was developed to quantify morphology in IHC stained tissue (PAN-CK+). The use of IHC, instead of HE, was proposed due to the clearer distinction of glands and stroma. Although performance was satisfactory, with acceptable discrimination, it did not come without its limitations. Firstly, it was sensitive to staining variation, in cases where gland DAB visualisation was weak, segmentation was poor and the ENDOAPP would overestimate the %stroma, and underestimate %gland. This increased the risk of a false negative result. This is a weakness of study design, as the model was highly dependent on colour. This was also notable in the epithelial nuclei segmentation step, if the nuclei staining was weak then nuclei segmentation was poor and thus the measurement of nuclei variation became more inaccurate. The ENDOAPP is entirely dependent on good staining i.e., distinct separation of blue and brown visualisation, to obtain accurate results and predictions. In hindsight, with the technologies now available, a deep learning method, trained on HE slides, may be a more robust choice. Nonetheless, the ENDOAPP demonstrates that the use of AI to quantify the morphological features in endometrial hyperplasia, can be prognostic, as previously encountered with D-score.

5.3 The Future of Proliferation in Breast Cancer

It is universally acknowledged that scoring of Ki67 should be standardised. However, with a plethora of methods available and many expert opinions to be regarded, deployment of a single, optimised method is unlikely. Perhaps a single method is not the answer. It is more likely that many methods can be standardised according to the same

specifications. To achieve this requires international cooperation and inter-disciplinary collaboration, of which the IKWG has a head start.

The future of Ki67 in breast cancer, I believe lies in validation. For example, as demonstrated in **paper III**, and other studies[198,206,253-258], Ki67 scoring is prognostic under a variety of conditions, may it be type of digital platform or the cut-off. Importantly, automated digital scoring is equivalent or better than manual methods. Yet, adoption of DIA methods is not as simple as we would be led to believe. As we investigated in **paper III**, as did another[443], the total number of tumour cells contributes to reproducibility and prognostic accuracy of Ki67 score. Manual global score, despite its terminology, estimates a relative global score and scores 300-400 tumour cells. Manual hotspot score usually scores ≥500 tumour cells. With DIA, there is no technological limit to the number of tumour cells that can be counted in a WSI. It gives us the opportunity to identify new thresholds which were limited by what was efficient and physically possible by pathologists and technicians. However, there is currently no standard tumour cell count specification in commercial DIA scoring APPs available. It is therefore up to the institution. Furthermore, commercial scoring APPs tend to mimic a pathologist rather than be designed to optimise prediction of prognosis or treatment response by evaluation of a cut-off. Therefore, there is still a need for improvement of Ki67 scoring. In the future, if we wish to standardise Ki67 DIA scoring, further investigation should be performed into the consequences of unstandardised and standardised tumour cell counts.

Future work should include further comparisons with molecular profiling methods. Additionally, better definition of cut-offs and how they should be used: if Ki67 should be used independently of molecular profiles or as a pre-screening tool. This needs to be validated in large multi-centre studies but should also be verified locally. In addition, one

may expect that with improvements in AI technology, perhaps other methods will replace Ki67. Already, studies have demonstrated the potential of AI to predict molecular subtypes on HE[450-452]. As demonstrated in **paper IV**, it is possible to develop algorithms to detect objects in HE slides, with prognostic success. In the future it is likely that the tasks taken on by AI will be more complex, combining object detection and molecular findings. A panel approach might even be covered by a single AI method. Perhaps idealistic, considering the same was hypothesised nearly 50 years ago, yet not impossible knowing what we do now.

As Geoffrey Keynes stated all those years ago, "Standardization must not, however, be allowed to create a fixed belief that no further improvement is possible and that any suggested change is necessarily to be regarded with disapproval"[112]. Yet, change takes time, and the technology available is still evolving. It is doubtful that the Ki67 scoring dilemma will be solved any time soon, but I do believe it will be improved. Furthermore, we are creatures of habit, we find comfort in what is familiar, so to ignore the prognostic value of Ki67 in favour of other methods would be to slight it too soon. However, to believe that Ki67 will be just as important in the future as it is today, would be too unimaginative.

The future of mitotic counting in pathology might be a little clearer than for Ki67. It's role in TNM staging of breast cancer is indisputable and many laboratories are transitioning from mitotic counting using a microscope to digital WSIs. The next step is likely to be the implementation of AI-based object detection tools, like the model evaluated in **paper IV**. Furthermore, validation in multiple cancer types not only increases the robustness of such a tool, due to the variable visual presentation of mitotic figures in different cancer types, but also its generalisability. As observed in **paper IV**, mitotic counting may be

prognostic in more than just breast cancer and this is worth pursuing in further studies. Additionally, it will be important to pursue investigations focusing on prognostic thresholds in other cancer types, as well as in breast cancer. In the development of AI, it is important to look beyond the confines of what is relevant today to access the full potential of these tools.

5.4 Implementation of AI for Clinical Practice

Institutions across continents are at varying stages of deployment of digital pathology and several have now begun the process of implementing AI in their clinics. Despite the promise of AI in digital pathology, implementation of such tools is not yet widespread and the number of in-depth implementation case studies for AI tools is lacking. Potential reasons for this may be attributed to scepticism surrounding the use of such tools due to ethical and safety concerns but also due to lack of understanding amongst various stakeholders (e.g. health care professionals, patients, IT technicians, legal entities, etc.)[453]. Furthermore, concerns have arisen regarding the cost of such tools, with the business case requiring the reduction of other costs such as IHC as a direct result of AI implementation[246]. An in-house developed algorithm may save on the cost of purchasing a commercial solution, but the resources for integration, monitoring and updating will also need to be supplied from within. Unfortunately, in many countries public institutions will be unable to allocate the time and personnel required due to budget constraints. Furthermore, we need to consider which APP or vendor to choose, and if APPs are selected from multiple vendors a new problem arises regarding usability – training pathologists to familiarise themselves with different user interfaces (UIs) in addition to the systems they already use. Also, whether DICOM or proprietary image file formats should be used, with considerations made for if they affect APP performance. With time, vendors and users alike will most

likely adapt as we already begin to see with some vendors working together to provide their APP solutions via each other's systems[454,455].

Protocols for implementation of digital pathology have been described by Williams and colleagues at Leeds Teaching Hospitals in Leeds, UK[14,456,457]. Furthermore, checklists for the critical evaluation of AI tools prior to clinical implementation have been published[35,58]. These provide useful guidelines to consider when implementing AI locally. This is particularly important with the recent EU-IVDR regulations (EU 2017/746). Additionally, development, validation and implementation of AI are not the only important steps to consider, but what happens to the AI after deployment. Quality control and monitoring should be performed, just as it is for other medical techniques such as IHC. For AI, monitoring over time will be critical to ensure patient safety. Quality control should be considering and where necessary performed regularly and following changes to the pipeline, e.g. software update, a new antibody, a new scanner, following scanner calibration, and so on. Not only monitoring of the hardware or algorithm performance is necessary, but we need to acknowledge how the pathologist will be influenced by the deployment of AI-assistance tools. We still need to address how to handle scenarios where the pathologist disagrees with the tool, yet the literature shows the AI-tool as being more prognostic than the former. Where does the ethical responsibility and liability lie? With the pathologist or with AI? Such scenarios must be considered to ensure ethical use of diagnostic assistance tools in the clinic and that they are being managed in the best way possible. There is also the concern of reliance, whereby a pathologist or technologist may become less critical of AI over time and therefore, tools of lower quality or with critical, yet subtle errors may slip through into the clinic.

Designing an AI tool is just the first step in a much larger process. The patient and intended use, should always be at the focus during

development, validation, implementation, and monitoring. The use of AI has the potential to revolutionise patient care in medicine, with the clinician at the helm. This will only be achievable through the collaborative efforts of healthcare professionals, computer scientists and policymakers. Furthermore, international collaboration is essential to ensuring equal and fair access to healthcare. AI has the potential to achieve this, to mitigate bias in a landscape littered with it, but only through extensive testing and validation. Only by research can we improve upon technology, contribute to medical breakthroughs, uncover, and eliminate bias, to relieve the burden upon healthcare professionals and improve patient health.

6 Concluding Remarks

Progress is inevitable. Paradigm shifts spurred by new technology may yield a greater understanding of the role of biomarkers in disease. The progress we have made today has been dependent on the discoveries of the past. At the core of this thesis is the objective to explore methods for their potential to improve diagnosis and prognosis by reducing the risk of over- and under-treatment. Deployment of digital pathology and AI creates new avenues for exploring biomarkers, tissue, and molecular data alike. The findings of this thesis demonstrate the need for validation of this new technology. Furthermore, as research continues to prove, the complexity of cell biology and pathology is immense. To narrow down a disease such as cancer, to a single biomarker, is too simplistic. However, to consider every aspect of each cancer and the individual it resides in is dauntingly complex, particularly in the face of the ever-increasing molecular data influx. Here, AI has the potential to aid us in tackling this problem.

Where science fiction has encouraged us to fear AI, it is unsurprising that society is apprehensive to its extensive adoption and integration. However, as with all change, it is important to meet it, albeit with caution. In days where news headlines are inundated with mention of AI, following the explosive popularity of AI solutions such as ChatGPT, the lack of knowledge regarding the consequences surrounding their unregulated use is a real concern. In medicine, digital pathology and AI is designed for the simple purpose of aiding a clinician; to relieve the burden placed on too few to help the many. Furthermore, to expand healthcare access beyond physical borders. The consensus in the fields of radiology and pathology has been that AI will not replace medical

personnel but will augment them often citing that those who do not use AI will be the ones who are replaced[35,49,458,459].

Humans are notoriously resilient and demonstrate rapid adaptability, and as humans are the driving force behind AI, we expect it to do the same. Yet, we stand now, as the physicians of the past did before us, with new technology that present us with opportunity. At present it is yet unknown how AI will shape healthcare and society or to what extent it should, but we must utilise the resources at hand and adapt as we always have, one day at a time.

"Science is an endurance sport. To produce that single illuminating experiment, a thousand nonilluminating experiments have to be sent into the trash; it is battle between nature and nerve."[460]
– Siddhartha Mukherjee, 2016

Appendices

Appendix 1 – References

1. Collins English Dictionary. Definition: Pathology. 2024. Accessed January 23, 2024. https://www.collinsdictionary.com/dictionary/english/pathology
2. Mukherjee S. The Emperor of All Maladies A Biography of Cancer. Fourth Estate; 2011.
3. Krzyszczyk P, Acevedo A, Davidoff EJ, et al. The growing role of precision and personalized medicine for cancer treatment. Technology (Singap World Sci). 2018;6(3-4):79-100. doi: 10.1142/S2339547818300020
4. Erlich P. Partial Cell Functions. Nobel Lecture. December 11, 1908. Accessed January 25, 2024. https://www.nobelprize.org/uploads/2018/06/ehrlich-lecture.pdf
5. Heffner S, Colgan O, Doolan C. Leica Biosystems. Digital Pathology. Accessed January 13, 2023. https://www.leicabiosystems.com/knowledge-pathway/digital-pathology/
6. Metter DM, Colgan TJ, Leung ST, Timmons CF, Park JY. Trends in the US and Canadian pathologist workforces from 2007 to 2017. JAMA Netw Open. 2019;2(5):e194337. doi:10.1001/jamanetworkopen.2019.4337.
7. Robboy SJ, Weintraub S, Horvath AE, et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. Arch Pathol Lab Med. 2013;137(12):1723-1732. doi: 10.5858/arpa.2013-0200-OA
8. Steiner DF, Chen P-HC, Mermel CH. Closing the translation gap: AI applications in digital pathology. Biochim Biophys Acta Rev Cancer. 2020;1875(1):188452. doi: 10.1016/j.bbcan.2020.188452
9. Zhao F, Dong D, Du H, et al. Diagnosis of endometrium hyperplasia and screening of endometrial intraepithelial neoplasia in histopathological images using a global-to-local multi-scale convolutional neural network. Comput Methods Programs Biomed. 2022:221:106906. doi: 10.1016/j.cmpb.2022.106906
10. Mudenda V, Malyangu E, Sayed S, Fleming K. Addressing the shortage of pathologists in Africa: Creation of a MMed Programme in Pathology in Zambia. Afr J Lab Med. 2020;9(1):974. doi:10.4102/ajlm.v9i1.974

11. Kumar P. Providing the providers—remedying Africa's shortage of health care workers. N Eng J Med. 2007;356(25):2564-2567. doi: 10.1056/NEJMp078091
12. Hitchcock CL. The future of telepathology for the developing world. Arch Pathol Lab Med. 2011;135(2):211-214. doi: 10.5858/135.2.211
13. Mosquera-Zamudio A G-SM, Sprockel J, Riano-Moreno JC, Janssen EAM, Pantanowitz L, Medina RP. Gloria: Globalization of a Telepathology Network with Artificial Intelligence Applications in Colombia: The GLORIA Program Study Protocol. Submitted to J Pathol 2024.
14. Williams BJ, Jayewardene D, Treanor DJH. Digital immunohistochemistry implementation, training and validation: experience and technical notes from a large clinical laboratory. J Clin Pathol. 2019;72(5):373-378. doi:10.1136/jclinpath-2018-205628
15. Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor DJH. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. Histopathology. 2018;72(4):662-671. doi: 10.1111/his.13403
16. Atallah NM, Toss MS, Verrill C, Salto-Tellez M, Snead D, Rakha EA. Potential quality pitfalls of digitalized whole slide image of breast pathology in routine practice. Mod Pathol. 2022;35(7)903-910. doi:10.1038/s41379-021-01000-8
17. The Medical Imaging Technology Association (MITA). DICOM. Accessed January 17, 2024. https://www.dicomstandard.org/
18. Rålund M. A DICOM Breakthrough in digital pathology. SECTRA; 2022. Accessed January 23, 2024. https://medical.sectra.com/resources/a-dicom-breakthrough-in-digital-pathology/
19. Cross S, Furness P, Igali L, Snead D, Treanor D. Best practice recommendations for implementing digital pathology January 2018. The Royal College of Pathologists. Accessed February 9, 2024. https://www.rcpath.org/static/f465d1b3-797b-4297-b7fedc00b4d77e51/Best-practice-recommendations-for-implementing-digital-pathology.pdf
20. Saini T, Bansal B, Dey P. Digital cytology: Current status and future prospects. Diagn Cytopathol. 2023;51(3):211-218. doi: 10.1002/dc.25099

21. Biology & Biochemistry Imaging Core (BBIC). Leica Z Stacks. University of Houston. Accessed February 9, 2024. https://bbic.nsm.uh.edu/protocols/leica-z-stacks
22. Stathonikos N, Nguyen TQ, Spoto CP, Verdaasdonk MA, van Diest PJ. Being fully digital: perspective of a Dutch academic pathology laboratory. Histopathology. 2019;75(5):621-635. doi: 10.1111/his.13953
23. Tabata K, Uraoka N, Benhamida J, et al. Validation of mitotic cell quantification via microscopy and multiple whole-slide scanners. Diagn Pathol. 2019;14(1):65. doi: 10.1186/s13000-019-0839-8
24. Sturm B, Creytens D, Cook MG, et al. Validation of whole-slide digitally imaged melanocytic lesions: Does z-stack scanning improve diagnostic accuracy? J Pathol Inform. 2019;10(1):6. doi: 10.4103/jpi.jpi_46_18
25. Kim D, Burkhardt R, Alperstein SA, et al. Evaluating the role of Z-stack to improve the morphologic evaluation of urine cytology whole slide images for high-grade urothelial carcinoma: results and review of a pilot study. Cancer Cytopathol. 2022;130(8):630-639. doi: 10.1002/cncy.22595
26. Aeffner F, Adissu HA, Boyle MC, et al. Digital microscopy, image analysis, and virtual slide repository. ILAR J. 2018;59(1):66-79. doi: 10.1093/ilar/ily007
27. Merriam-Webster. Definition: "Image". 2024. Accessed: January 17, 2024. https://www.merriam-webster.com/dictionary/image
28. Lyra M, Ploussi A, Georgantzoglou A. Matlab as a tool in nuclear medicine image processing. In: Ionescu P. MATLAB-A Ubiquitous tool for the practical engineer. 2011: chap 23. Accessed February 9, 2024. https://www.intechopen.com/chapters/21962
29. Rewcastle E. Assessing Ki67 in Breast Cancer using Image Processing and Artificial Intelligence. ELE922 Biomedical Data Analysis. University of Stavanger; 2022.
30. Merriam-Webster. Definition: "Machine-Learning". 2024. Accessed: January 17, 2024. https://www.merriam-webster.com/dictionary/machine%20learning
31. Holzinger A. Introduction to MAchine Learning & Knowledge Extraction (MAKE). Mach Learn Knowl Extr. 2019;1(1):1-20. https://doi.org/10.3390/make1010001
32. Artificial Intelligence Medical Devices (AIMD) Working Group. Machine Learning-enabled Medical Devices: Key Terms and Definitions. 2022.

Accessed: May 6, 2022. https://www.imdrf.org/sites/default/files/2022-05/IMDRF%20AIMD%20WG%20Final%20Document%20N67.pdf

33. Bewtra A. The Ultimate Guide to Semi-Supervised Learning. V7 Labs. July 1, 2022. Accessed 27 January 2024. https://www.v7labs.com/blog/semi-supervised-learning-guide
34. Baheti P. Supervised and Unsupervised Leaning [Differences & Examples]. V7 Labs. October 1, 2021. Accessed 27 January 2024. https://www.v7labs.com/blog/supervised-vs-unsupervised-learning
35. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform. 2021;28(1):e100251. doi: 10.1136/bmjhci-2020-100251
36. Wetteland R. Automated Grading of Bladder Cancer using Deep Learning. Dissertation. University of Stavanger; 2022. Accessed February 9, 2024. https://uis.brage.unit.no/uis-xmlui/handle/11250/2977487
37. Education Ecosystem (LEDU). Understanding K-means Clustering in Machine Learning. Towards Data Science. September 12, 2018. Accessed July 4, 2023. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
38. Gupta N. Artificial neural network. Network and Complex Systems. 2013;3(1):24-28.
39. IBM. What are convolutional neural networks? Accessed February 18, 2024. https://www.ibm.com/topics/convolutional-neural-networks
40. Dongare A, Kharde R, Kachare AD. Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT). 2012;2(1):189-194.
41. Shah D. The Essential Guide to Pytorch Loss Functions. V7 Labs. July 13, 2022. Accessed January 27, 2024. https://www.v7labs.com/blog/pytorch-loss-functions
42. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-536.
43. Donges N. Gradient Descent in Machine Learning: A Basic Introduction. Built In. March 27, 2023. Accessed April 28, 2023. https://builtin.com/data-science/gradient-descent
44. Pramoditha R. How to Choose the Optimal Learning Rate for Neural Networks. Towards Data Science. September 21, 2022. Accessed April 28, 2023. https://towardsdatascience.com/how-to-choose-the-optimal-learning-rate-for-neural-networks-362111c5c783

45. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. 2020;92(4):807-812. doi: 10.1016/j.gie.2020.06.040.
46. Galanis N-I, Vafiadis P, Mirzaev K-G, Papakostas GA. Convolutional Neural Networks: A Roundup and Benchmark of Their Pooling Layer Variants. Algorithms. 2022;15(11):391.
47. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16(11):703-715. doi: 10.1038/s41571-019-0252-y
48. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023;388(13):1201-1208. doi: 10.1056/NEJMra2302038
49. Ahmad Z, Rahim S, Zubair M, Abdul-Ghafar J. Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: Present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. Diagn Pathol. 2021;16(1):24. doi: 10.1186/s13000-021-01085-4
50. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-945. doi: 10.1038/nature03001
51. Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience. 2018;7(6):giy065. doi: 10.1093/gigascience/giy065
52. The Cancer Genome Atlas Program (TCGA). Genomic Data Commons Data Portal (GDC). Genomic Data Commons Data Portal (GDC). Accessed July 6, 2023. https://portal.gdc.cancer.gov/projects/TCGA-BRCA
53. Tafavvoghi M, Bongo LA, Shvetsov N, Busund L-TR, Møllersen K. Publicly available datasets of breast histopathology H&E whole-slide images: A scoping review. J Pathol Inform. 2024:100363. https://doi.org/10.1016/j.jpi.2024.100363
54. Artificial Intelligence in BreastScreen Norway - a randomized controlled trial. ClinicalTrials.gov identifier: NCT06032390. Updated January 12, 2024. Accessed February 9, 2024. https://www.kreftregisteret.no/en/Research/Projects/aims/
55. Flach RN, Stathonikos N, Nguyen TQ, Ter Hoeve ND, van Diest PJ, van Dooijeweert C. CONFIDENT-trial protocol: a pragmatic template for clinical

implementation of artificial intelligence assistance in pathology. BMJ Open. 2023;13(6):e067437. doi: 10.1136/bmjopen-2022-067437.
56. Joshi I, Cushnan D. A Buyer's Guide to AI in Health and Care. NHS Transformation Directorate; September 8, 2020.
57. Collins English Dictionary. Definition: "Ground truth". 2024. Accessed: January 17, 2024.
58. Colling R, Pitman H, Oien K, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. J Pathol. 2019;249(2):143-150. doi: 10.1002/path.5310
59. 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(18):1806-1816. doi: 10.1001/jama.2019.16489
60. Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. December 18, 2017. Accessed July 4, 2023. https://laurenoakdenrayner.com/2017/12/18/the-chestxray14-dataset-problems/
61. Homeyer A, Geißler C, Schwen LO, et al. Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. Mod Pathol. 2022;35(12):1759-1769.
62. Oxford English Dictionary. Definition: "Bias". 2024. Accessed: January 17, 2024.
63. Rizzoli A. 40+ Data Science Interview Questions and Answers. V7 Labs. June 11, 2021. Accessed: January 27, 2024. https://www.v7labs.com/blog/data-science-interview-questions-and-answers
64. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. Future Healthc J. 2021;8(1):40-48. doi: 10.7861/fhj.2020-0233
65. Perez CC. Invisible women: Data bias in a world designed for men. 1st ed. Chatto & Windus; 2019.
66. Bierer BE, Meloney LG, Ahmed HR, White SA. Advancing the inclusion of underrepresented women in clinical research. Cell Rep Med. 2022;3(4):100553. doi: 10.1016/j.xcrm.2022.100553
67. Fultinavičiūtė U. Sex and science: underrepresentation of women in early-stage clinical trials. Clinical Trials Arena. October 17, 2022. Accessed May 25, 2023. https://www.clinicaltrialsarena.com/features/underrepresentation-women-early-stage-clinical-trials/

68. NIH Central Resource of Grants and Funding Information. NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research. October 9, 2001. Accessed: February 9, 2024. https://grants.nih.gov/policy/inclusion/women-and-minorities/guidelines.htm
69. Smith, J, Noble H. Bias in Research. Evidence-Based Nursing. 2014;17:100-101.
70. Leuzzi C, Sangiorgi GM, Modena MG. Gender-specific aspects in the clinical presentation of cardiovascular disease. Fundam Clin Pharmacol. 2010;24(6):711-717. doi: 10.1111/j.1472-8206.2010.00873.x.
71. Norris CM, Yip CY, Nerenberg KA, et al. State of the science in women's cardiovascular disease: a Canadian perspective on the influence of sex and gender. J Am Heart Assoc. 2020;9(4):e015634. doi: 10.1161/JAHA.119.015634
72. Georgiev D, Hamberg K, Hariz M, Forsgren L, Hariz GM. Gender differences in Parkinson's disease: a clinical perspective. Acta Neurol Scand. 2017;136(6):570-584. doi: 10.1111/ane.12796
73. Kim H-I, Lim H, Moon A. Sex differences in cancer: epidemiology, genetics and therapy. Biomol Ther (Seoul). 2018;26(4):335-342. doi: 10.4062/biomolther.2018.103
74. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154(11):1247-1248. doi: 10.1001/jamadermatol.2018.2348
75. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med. 2020;3(1):81. doi: 10.1038/s41746-020-0288-5
76. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. John Wiley & Sons; 2013.
77. Larner A. The 2x2 matrix: contingency, confusion and the metrics of binary classification. 2nd ed. Springer Nature; 2024.
78. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health. 2017;5:307. doi: 10.3389/fpubh.2017.00307
79. Rewcastle E. The Influence of Scientific Discovery in Breast Cancer Treatment; the Evolution of Personalised Medicine. TN900 Theory of Science and Ethics. University of Stavanger; 2021.
80. Weinberg RA. The Biology of Cancer. 1st ed. Garland Science; 2006.

81. National Cancer Institute (NCI). What is Cancer? October 11, 2021. Accessed: January 31, 2024. https://www.cancer.gov/about-cancer/understanding/what-is-cancer
82. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-674. doi: 10.1016/j.cell.2011.02.013
83. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70. doi: 10.1016/s0092-8674(00)81683-9
84. Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov. 2022;12(1):31-46. doi: 10.1158/2159-8290.CD-21-1059.
85. Sherr CJ, McCormick F. The RB and p53 pathways in cancer. Cancer Cell. 2002;2(2):103-112. doi: 10.1016/s1535-6108(02)00102-2.
86. National Cancer Institute Dictionary. Definition: "Biomarker". 2024. Accessed: February 9, 2024. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker
87. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US). 2016. Updated November 29, 2021. Accessed August 3, 2023. https://www.ncbi.nlm.nih.gov/books/NBK326791/
88. Obama B. Remarks by the President on Precision Medicine. Office of the Press Secretary; 2015. Accessed: January 20, 2024. https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/remarks-president-precision-medicine
89. WHO Classification of Tumours Editorial Board. Breast Tumours. 5th ed. vol 2. World Health Organisation Classification of Tumours. International Agency for Research on Cancer; 2019.
90. Sproston NR, Ashworth JJ. Role of C-reactive protein at sites of inflammation and infection. Front Immunol. 2018;9:754. doi: 10.3389/fimmu.2018.00754
91. Melroe NH, Stawarz KE, Simpson J, Henry WK. HIV RNA quantitation: Marker of HIV infection. J Assoc Nurses AIDS Care. 1997;8(5):31-38. doi: 10.1016/S1055-3290(97)80027-1
92. Norwegian Breast Cancer Group. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft. 16th ed. Norwegian Department of Health; 2021. Accessed February 9, 2021. https://nbcgblog.files.wordpress.com/2021/03/nasjonalt-handlingsprogram-for-pasienter-med-brystkreft-01.03.2021-16-utgave.pdf

93. James PA, Oparil S, Carter BL, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). JAMA. 2014;311(5):507-520. doi: 10.1001/jama.2013.284427
94. Rizzo JD, Brouwers M, Hurley P, Seidenfeld J, Somerfield MR, Temin S. American Society of Clinical Oncology/American Society of Hematology clinical practice guideline update on the use of epoetin and darbepoetin in adult patients with cancer. J Oncol Pract. 2010;6(6):317-320. doi: 10.1200/JOP.2010.000132
95. Eldridge L. Overview of Neutropenia During Chemotherapy. Verywellhealth. Updated October 6, 2022. Accessed August 3, 2023. https://www.verywellhealth.com/neutropenia-and-chemotherapy-2249337
96. Cancer Registry of Norway. Cancer in Norway 2022 - Cancer incidence, mortality, survival and prevalence in Norway. 2023. https://www.kreftregisteret.no/globalassets/cancer-in-norway/2022/cin_report-2022.pdf
97. Sung H, Ferlay J, Siegel RL, 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(3):209-249. doi: 10.3322/caac.21660
98. Chen X, Wang Q, Zhang Y, Xie Q, Tan X. Physical activity and risk of breast cancer: a meta-analysis of 38 cohort studies in 45 study reports. Value Health. 2019;22(1):104-128. doi: 10.1016/j.jval.2018.06.020
99. Kolb R, Zhang W. Obesity and breast cancer: a case of inflamed adipose tissue. Cancers (Basel). 2020;12(6):1686. doi: 10.3390/cancers12061686
100. Zeinomar N, Knight JA, Genkinger JM, et al. Alcohol consumption, cigarette smoking, and familial breast cancer risk: findings from the Prospective Family Study Cohort (ProF-SC). Breast Cancer Res. 2019;21(1):128. doi: 10.1186/s13058-019-1213-1
101. Ursin G, Bernstein L, Lord S, et al. Reproductive factors and subtypes of breast cancer defined by hormone receptor and histology. Br J Cancer. 2005;93(3):364-371. doi: 10.1038/sj.bjc.6602712.
102. Albrektsen G, Heuch I, Hansen S, Kvåle G. Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects. Br J Cancer. 2005;92(1):167-175. doi: 10.1038/sj.bjc.6602302
103. Titus-Ernstoff L, Longnecker MP, Newcomb PA, et al. Menstrual factors in relation to breast cancer risk. Cancer Epidemiol Biomarkers Prev. 1998;7(9):783-789.

104. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. Lancet. 2001;358(9291):1389-1399. doi: 10.1016/S0140-6736(01)06524-2
105. National Cancer Institute. BRCA Gene Mutations: Cancer Risk and Genetic Testing. National Cancer Institute. November 19, 2020. Accessed: April 22, 2021. https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet
106. Cancer Registry of Norway. Cancer in Norway 2020 - Cancer incidence, mortality, survival and prevalence in Norway. 2021.
107. Memorial Sloan Kettering Cancer Centre. Anatomy of the Breast. Accessed: February 9, 2024. https://www.mskcc.org/cancer-care/types/breast/anatomy-breast
108. Seer Training Modules. Breast Anatomy. National Cancer Institute. Accessed: February 9, 2024. https://training.seer.cancer.gov/breast/anatomy/
109. Lerner BH. The Breast Cancer Wars. Oxford University Press; 2001.
110. Dodge DG. Surgical Treatment of Breast Cancer. The Journal of Lancaster General Hospital. 2007;2(1):27-31.
111. Halsted WS. The Results of Operations for the Cure of Cancer of the Breast Performed at the John Hopkins Hospital from June, 1889, to January, 1894. Ann Surg. 1894;20(5):497-55. Doi: 10.1097/00000658-189407000-00075.
112. Keynes G. Conservative treatment of cancer of the breast. Br Med J. 1937;2(4004):643.
113. Simoni RD, Hill RL, Vaughan M. The discovery of estrone, estriol, and estradiol and the biochemical study of reproduction. The work of Edward Adelbert Doisy. J Biol Chem. 2002;277(28):e1-e2. https://doi.org/10.1016/S0021-9258(19)66427-6
114. Moore DD. A conversation with elwood jensen. Annu Rev Physiol. 2012;74:1-11. doi: 10.1146/annurev-physiol-020911-153327
115. Jensen VE, Jacobson HI. Fate of Steroid Estrogens in Target Tissues. In: Biological Activities of Steroids in Relation to Cancer. Elsevier; 1960. https://doi.org/10.1016/C2013-0-12113-7
116. Quirke VM. Tamoxifen from failed contraceptive pill to best-selling breast cancer medicine: a case-study in pharmaceutical innovation. Front Pharmacol. 2017;8:620. doi: 10.3389/fphar.2017.00620

117. Merriam-Webster. Definition: "Clinicopathologic". 2024. Accessed: January 5, 2024.
118. Norwegian Breast Cancer Group. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft. 2023. https://nbcgblog.files.wordpress.com/2023/02/11.01.2023-nasjonalt-handlingsprogram-for-brystkreft-19.-utgave-publisert-11.01.23.pdf
119. Bagaria SP, Ray PS, Sim M-S, et al. Personalizing breast cancer staging by the inclusion of ER, PR, and HER2. JAMA Surg. 2014;149(2):125-129. doi: 10.1001/jamasurg.2013.3181
120. Elkin EB, Hudis C, Begg CB, Schrag D. The effect of changes in tumor size on breast carcinoma survival in the US: 1975–1999. Cancer. 2005;104(6):1149-1157. doi: 10.1002/cncr.21285
121. Rezo A, Dahlstrom J, Shadbolt B, Rodins K, Zhang Y, Davis AJ. Tumor size and survival in multicentric and multifocal breast cancer. Breast. 2011;20(3):259-263. doi: 10.1016/j.breast.2011.01.005
122. Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MM. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. BMJ. 2015;351:h4901 doi: 10.1136/bmj.h4901
123. Soerjomataram I, Louwman MW, Ribot JG, Roukema JA, Coebergh JWW. An overview of prognostic factors for long-term survivors of breast cancer. Breast Cancer Res Treat. 2008;107(3):309-330. doi: 10.1007/s10549-007-9556-1
124. Elston CW, Ellis IO, Pinder SE. Pathological prognostic factors in breast cancer. Crit Rev Oncol Hematol. 1999;31(3):209-223. doi: 10.1016/s1040-8428(99)00034-7
125. Rosen PP, Groshen S. Factors influencing survival and prognosis in early breast carcinoma (T1N0M0–T1N1M0): assessment of 644 patients with median follow-up of 18 years. Surg Clin North Am. 1990;70(4):937-962. doi: 10.1016/s0039-6109(16)45190-x
126. Chia SK, Speers CH, Bryce CJ, Hayes MM, Olivotto IA. Ten-year outcomes in a population-based cohort of node-negative, lymphatic, and vascular invasion–negative early breast cancers without adjuvant systemic therapies. J Clin Oncol. 2004;22(9):1630-1637. doi: 10.1200/JCO.2004.09.070

127. Carter CL, Allen C, Henson DE. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer. 1989;63(1):181-187. doi: 10.1002/1097-0142(19890101)63:1<181::aid-cncr2820630129>3.0.co;2-h
128. de Boer M, van Dijck JA, Bult P, Borm GF, Tjan-Heijnen VC. Breast cancer prognosis and occult lymph node metastases, isolated tumor cells, and micrometastases. J Natl Cancer Inst. 2010;102(6):410-425. doi: 10.1093/jnci/djq008
129. Veronesi U, Galimberti V, Zurrida S, Merson M, Greco M, Luini A. Prognostic significance of number and level of axillary node metastases in breast cancer. Breast. 1993;2(4):224-228. https://doi.org/10.1016/0960-9776(93)90004-Y
130. Jatoi I, Hilsenbeck SG, Clark GM, Osborne CK. Significance of axillary lymph node metastasis in primary breast cancer. J Clin Oncol. 1999;17(8):2334-2334. doi: 10.1200/JCO.1999.17.8.2334
131. D'eredita G, Giardina C, Martellotta M, Natale T, Ferrarese F. Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution. Eur J Cancer. 2001;37(5):591-596. doi: 10.1016/s0959-8049(00)00435-4
132. Dai X, Li T, Bai Z, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res. 2015;5(10):2929-43.
133. Bloom H, Richardson W. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11(3):359-377. doi: 10.1038/bjc.1957.43
134. Chang JM, McCullough AE, Dueck AC, et al. Back to Basics: Traditional Nottingham Grade Mitotic Counts Alone are Significant in Predicting Survival in Invasive Breast Carcinoma. Ann Surg Oncol. 2015;22:509-515. doi: 10.1245/s10434-015-4616-y
135. National Disease Registration Service. NPI Score. National Health Service (NHS). Updated December 7, 2022. Accessed: February 9, 2024. https://digital.nhs.uk/ndrs/data/data-sets/cosd/cosd-user-guide/site-specific---breast
136. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat. 1992;22(3):207-219. doi: 10.1007/BF01840834
137. Balslev I, Axelsson CK, Zedeler K, Rasmussen BB, Carstensen B, Mouridsen HT. The Nottingham prognostic index applied to 9,149 patients from the

studies of the Danish Breast Cancer Cooperative Group (DBCG). Breast Cancer Res Treat. 1994;32(3):281-290. doi: 10.1007/BF00666005

138. Rakha EA, Agarwal D, Green AR, et al. Prognostic stratification of oestrogen receptor-positive HER 2-negative lymph node-negative class of breast cancer. Histopathology. 2017;70(4):622-631. https://doi.org/10.1111/his.13108
139. Ellis IO, Rakha, Emad. Nottingham Breast Pathology Research Group. University of Nottingham. Accessed February 22, 2024. https://www.nottingham.ac.uk/research/groups/pathology/nottingham-breast-pathology.aspx
140. Santen RJ, Simpson E. History of estrogen: its purification, structure, synthesis, biologic actions, and clinical implications. Endocrinology. 2019;160(3):605-625. doi: 10.1210/en.2018-00529
141. Huggins C, Bergenstal DM. Inhibition of human mammary and prostatic cancers by adrenalectomy. Cancer Res. 1952;12(2):134-141.
142. Kennedy B. Hormone therapy for advanced breast cancer. Cancer. 1965;18(12):1551-1557. doi: 10.1002/1097-0142(196512)18:12<1551::aid-cncr2820181206>3.0.co;2-1
143. Block GE, Ellis RS, DeSombre E, Jensen E. Correlation of estrophilin content of primary mammary cancer to eventual endocrine treatment. Ann Surg. 1978;188(3):372. doi: 10.1097/00000658-197809000-00012.
144. Fuentes N, Silveyra P. Estrogen receptor signaling mechanisms. Adv Protein Chem Struct Biol. 2019;116:135-170. doi: 10.1016/bs.apcsb.2019.01.001
145. Ozyurt R, Ozpolat B. Molecular Mechanisms of Anti-Estrogen Therapy Resistance and Novel Targeted Therapies. Cancers (Basel). 2022;14(21):5206. doi: 10.3390/cancers14215206
146. Kaaks R, Lukanova A, Kurzer MS. Obesity, endogenous hormones, and endometrial cancer risk: a synthetic review. Cancer Epidemiol Biomarkers Prev. 2002;11(12):1531-1543.
147. Corner GW, Allen WM. Physiology of the corpus luteum: II. Production of a special uterine reaction (progestational proliferation) by extracts of the corpus luteum. American Journal of Physiology-Legacy Content. 1929;88(2):326-339.
148. Ali S BK, O'Malley B. 90 Years of progesterone: Ninety years of progesterone: the 'other' ovarian hormone. J Mol Endocrinol. 2020;65(1):E1-E4. doi: 10.1530/JME-20-0145

149. O'Malley BW. 90 YEARS OF PROGESTERONE: Reminiscing on the origins of the field of progesterone and estrogen receptor action. J Mol Endocrinol. 2020;65(1):C1-C4. doi: 10.1530/JME-20-0042
150. Purdie C, Quinlan P, Jordan L, et al. Progesterone receptor expression is an independent prognostic variable in early breast cancer: a population-based study. Br J Cancer. 2014;110(3):565-572. doi: 10.1038/bjc.2013.756
151. National Cancer Institute. Hormone Therapy for Breast Cancer. Updated July 12, 2022. Accessed January 31, 2024. https://www.cancer.gov/types/breast/breast-hormone-therapy-fact-sheet
152. Early Breast Cancer Trialist's Collaborative Group (EBCTCG). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet. 2011;378(9793):771-784. doi: 10.1016/S0140-6736(11)60993-8
153. Klein DJ, Thorn CF, Desta Z, Flockhart DA, Altman RB, Klein TE. PharmGKB summary: tamoxifen pathway, pharmacokinetics. Pharmacogenet Genomics. 2013;23(11):643-647.
154. Helland T, Henne N, Bifulco E, et al. Serum concentrations of active tamoxifen metabolites predict long-term survival in adjuvantly treated breast cancer patients. Breast Cancer Res. 2017;19(1):125. doi: 10.1186/s13058-017-0916-4.
155. Osborne CK. Tamoxifen in the treatment of breast cancer. N Engl J Med. 1998;339(22):1609-1618. doi: 10.1056/NEJM199811263392207
156. Gielen SC, Santegoets LA, Hanifi-Moghaddam P, Burger CW, Blok LJ. Signaling by estrogens and tamoxifen in the human endometrium. J Steroid Biochem Mol Biol. 2008;109(3-5):219-223. doi: 10.1016/j.jsbmb.2008.03.021
157. Hu R, Hilakivi-Clarke L, Clarke R. Molecular mechanisms of tamoxifen-associated endometrial cancer. Oncol Lett. 2015;9(4):1495-1501. doi: 10.3892/ol.2015.2962
158. Horwitz KB, Sartorius CA. 90 years of progesterone: progesterone and progesterone receptors in breast cancer: past, present, future. J Mol Endocrinol. 2020;65(1):T49-T63. doi: 10.1530/JME-20-0104
159. Bardou V-J, Arpino G, Elledge RM, Osborne CK, Clark GM. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. J Clin Oncol. 2003;21(10):1973-1979. doi: 10.1200/JCO.2003.09.099

160. McGuire WL, Clark GM, Dressler LG, Owens MA. Role of steroid hormone receptors as prognostic factors in primary breast cancer. NCI Monogr. 1986;1:19-23.
161. Cui X, Schiff R, Arpino G, Osborne CK, Lee AV. Biology of progesterone receptor loss in breast cancer and its implications for endocrine therapy. J Clin Oncol. 2005;23(30):7721-7735. doi: 10.1200/JCO.2005.09.004
162. Beltjens F, Molly D, Bertaut A, et al. ER−/PR+ breast cancer: A distinct entity, which is morphologically and molecularly close to triple-negative breast cancer. Int J Cancer. 2021;149(1):200-213. doi: 10.1002/ijc.33539
163. Setiawan VW, Monroe KR, Wilkens LR, Kolonel LN, Pike MC, Henderson BE. Breast cancer risk factors defined by estrogen and progesterone receptor status: the multiethnic cohort study. Am J Epidemiol. 2009;169(10):1251-1259. doi: 10.1093/aje/kwp036
164. Bae SY, Kim S, Lee JH, et al. Poor prognosis of single hormone receptor-positive breast cancer: similar outcome as triple-negative breast cancer. BMC Cancer. 2015;15(1):1-9. doi: 10.1186/s12885-015-1121-4
165. Li Y, Yang D, Yin X, et al. Clinicopathological characteristics and breast cancer–specific survival of patients with single hormone receptor–positive breast cancer. JAMA Netw Open. 2020;3(1):e1918160-e1918160. doi: 10.1001/jamanetworkopen.2019.18160
166. Rakha EA, El-Sayed ME, Green AR, et al. Biologic and clinical characteristics of breast cancer with single hormone receptor–positive phenotype. J Clin Oncol. 2007;25(30):4772-4778. doi: 10.1200/JCO.2007.12.2747
167. Ahmed SS, Thike AA, Zhang K, Lim JCT, Tan PH. Clinicopathological characteristics of oestrogen receptor negative, progesterone receptor positive breast cancers: re-evaluating subsets within this group. J Clin Pathol. 2017;70(4):320-326. doi: 10.1136/jclinpath-2016-203847
168. Dowsett M, Houghton J, Iden C, et al. Benefit from adjuvant tamoxifen therapy in primary breast cancer patients according oestrogen receptor, progesterone receptor, EGF receptor and HER2 status. Ann Oncol. 2006;17(5):818-826. doi: 10.1093/annonc/mdl016
169. Anderson WF, Chu KC, Chatterjee N, Brawley O, Brinton LA. Tumor variants by hormone receptor expression in white patients with node-negative breast cancer from the surveillance, epidemiology, and end results database. J Clin Oncol. 2001;19(1):18-27. doi: 10.1200/JCO.2001.19.1.18

170. Hsu JL, Hung M-C. The role of HER2, EGFR, and other receptor tyrosine kinases in breast cancer. Cancer Metastasis Rev. 2016;35:575-588. doi: 10.1007/s10555-016-9649-6
171. Gerbin CS. Activation of ERBB Receptors. Nature Education. 2010;3(9):35.
172. Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Subtypes of breast cancer. In: Mayrovitz HN. Breast Cancer [Internet]. Exon Publications; 2022: chap 3. Accessed: January 11, 2024.
173. Seshadri R, Firgaira F, Horsfall D, McCaul K, Setlur V, Kitchen P. Clinical significance of HER-2/neu oncogene amplification in primary breast cancer. The South Australian Breast Cancer Study Group. J Clin Oncol. 1993;11(10):1936-1942. doi: 10.1200/JCO.1993.11.10.1936
174. Tandon AK, Clark GM, Chamness GC, Ullrich A, McGuire W. HER-2/neu oncogene protein and prognosis in breast cancer. J Clin Oncol. 1989;7(8):1120-1128. doi: 10.1200/JCO.1989.7.8.1120
175. Hartmann LC, Ingle JN, Wold LE, et al. Prognostic value of c-erbB2 overexpression in axillary lymph node positive breast cancer. Results from a randomized adjuvant treatment protocol. Cancer. 1994;74(11):2956-2963. doi: 10.1002/1097-0142(19941201)74:11<2956::aid-cncr2820741111>3.0.co;2-v
176. Krishnamurti U, Silverman JF. HER2 in breast cancer: a review and update. Adv Anat Pathol. 2014;21(2):100-107. doi: 10.1097/PAP.0000000000000015
177. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344(11):783-792. doi: 10.1056/NEJM200103153441101
178. Baselga J, Perez EA, Pienkowski T, Bell R. Adjuvant trastuzumab: a milestone in the treatment of HER-2-positive early breast cancer. Oncologist. 2006;11(S1):4-12. doi: 10.1634/theoncologist.11-90001-4
179. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med. 2005;353(16):1659-1672. doi: 10.1056/NEJMoa052306
180. Gianni L, Eiermann W, Semiglazov V, et al. Neoadjuvant chemotherapy with trastuzumab followed by adjuvant trastuzumab versus neoadjuvant chemotherapy alone, in patients with HER2-positive locally advanced breast cancer (the NOAH trial): a randomised controlled superiority trial with a

parallel HER2-negative cohort. Lancet. 2010;375(9712):377-384. doi: 10.1016/S0140-6736(09)61964-4

181. Di Leo A, Gomez HL, Aziz Z, et al. Phase III, double-blind, randomized study comparing lapatinib plus paclitaxel with placebo plus paclitaxel as first-line treatment for metastatic breast cancer. J Clin Oncol. 2008;26(34):5544. doi: 10.1200/JCO.2008.16.2578
182. Baselga J, Bradbury I, Eidtmann H, et al. Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): a randomised, open-label, multicentre, phase 3 trial. Lancet. 2012;379(9816):633-640. doi: 10.1016/S0140-6736(11)61847-3
183. Geyer CE, Forster J, Lindquist D, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. N Engl J Med. 2006;355(26):2733-2743. doi: 10.1056/NEJMoa06432
184. OWise. What is HER2-low Breast Cancer? 2024. Accessed February 18, 2024. https://owise.uk/what-is-her2-low-breast-cancer/
185. Rakha EA, Pinder SE, Bartlett JM, et al. Updated UK Recommendations for HER2 assessment in breast cancer. J Clin Pathol. 2015;68(2):93-92. doi: 10.1136/jclinpath-2014-202571
186. European Medical Agency. Enhertu. Updated November 29, 2023. Accessed February 22, 2024. https://www.ema.europa.eu/en/medicines/human/EPAR/enhertu
187. Rakha EA, Pinder SE, Bartlett JM, et al. Updated UK Recommendations for HER2 assessment in breast cancer. J Clin Pathol. 2014;68(2)93-99. doi: 10.1136/jclinpath-2014-202571
188. The American Society of Clinical Oncology/College of American Pathologists. ASCO Guidelines for HER2 Testing in Breast Cancer. 2018. Accessed February 22, 2024. https://old-prod.asco.org/sites/new-www.asco.org/files/content-files/practice-and-guidelines/documents/2018-her2-testing-algorithms.pdf
189. Collins K, Jacks T, Pavletich NP. The cell cycle and cancer. Proc Natl Acad Sci. 1997;94(7):2776-2778. https://doi.org/10.1073/pnas.94.7.2776
190. Suryadinata R, Sadowski M, Sarcevic B. Control of cell cycle progression by phosphorylation of cyclin-dependent kinase (CDK) substrates. Biosci Rep. 2010;30(4):243-255. doi: 10.1042/BSR20090171
191. Encyclopaedia Britannica. Mitosis. 2024. Accessed February 10, 2024. https://www.britannica.com/science/mitosis

192. Baak JP, Gudlaugsson E, Skaland I, et al. Proliferation is the strongest prognosticator in node-negative breast cancer: significance, error sources, alternatives and comparison with molecular prognostic markers. Breast Cancer Res Treat. 2009;115(2):241-254. doi: 10.1007/s10549-008-0126-y
193. Baak JP, van Diest PJ, Voorhorst FJ, et al. Prospective multicenter validation of the independent prognostic value of the mitotic activity index in lymph node–negative breast cancer patients younger than 55 years. J Clin Oncol. 2005;23(25):5993-6001. doi: 10.1200/JCO.2005.05.511
194. Aziz S, Wik E, Knutsvik G, et al. Evaluation of tumor cell proliferation by Ki-67 expression and mitotic count in lymph node metastases from breast cancer. PloS One. 2016;11(3):e0150979. doi: 10.1371/journal.pone.0150979
195. Baak JP, Gudlaugsson E, Skaland I, et al. Proliferation is the strongest prognosticator in node-negative breast cancer: significance, error sources, alternatives and comparison with molecular prognostic markers. Breast Cancer Res Treat. 2009;115:241-254. doi: 10.1007/s10549-008-0126-y
196. Stuart-Harris R, Caldas C, Pinder S, Pharoah P. Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. Breast. 2008;17(4):323-334. doi: 10.1016/j.breast.2008.02.002
197. Smith I, Robertson J, Kilburn L, et al. Long-term outcome and prognostic value of Ki67 after perioperative endocrine therapy in postmenopausal women with hormone-sensitive early breast cancer (POETIC): an open-label, multicentre, parallel-group, randomised, phase 3 trial. Lancet Oncol. 2020;21(11):1443-1454. doi: 10.1016/S1470-2045(20)30458-7
198. Rimm DL, Leung SC, McShane LM, et al. An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer. Mod Pathol. 2019;32(1):59-69. doi: 10.1038/s41379-018-0109-4
199. Cheang MC, Chia SK, Voduc D, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst. 2009;101(10):736-750. doi: 10.1093/jnci/djp082.
200. De Azambuja E, Cardoso F, de Castro G, et al. Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12 155 patients. Br J Cancer. 2007;96(10):1504-1513. doi: 10.1038/sj.bjc.6603756
201. Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. Lancet Oncol. 2010;11(2):174-183. doi: 10.1016/S1470-2045(09)70262-1

202. Penault-Llorca F, André F, Sagan C, et al. Ki67 expression and docetaxel efficacy in patients with estrogen receptor-positive breast cancer. J Clin Oncol. 2009;27(17):2809-2815. doi: 10.1200/JCO.2008.18.2808
203. Thomssen C, Balic M, Harbeck N, Gnant M. St. Gallen/Vienna 2021: a brief summary of the consensus discussion on customizing therapies for women with early breast cancer. Breast Care (Basel). 2021;16(2):135-143. doi: 10.1159/000516114.
204. Polley M-YC, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. J Natl Cancer Inst. 2013;105(24):1897-1906. doi: 10.1093/jnci/djt306.
205. Polley M-YC, Leung SC, Gao D, et al. An international study to increase concordance in Ki67 scoring. Mod Pathol. 2015;28(6):778-786. doi: 10.1038/modpathol.2015.38
206. Røge R, Riber-Hansen R, Nielsen S, Vyberg M. Proliferation assessment in breast carcinomas using digital image analysis based on virtual Ki67/cytokeratin double staining. Breast Cancer Res Treat. 2016;158(1):11-19. doi: 10.1007/s10549-016-3852-6. Epub 2016 Jun 9
207. KVAST-gruppen för bröstpatologi. Bröstcancer vårdprogram. Regionala Cancercentrum i Samverkan. V.4.3. March 28, 2023. Accessed August 15, 2023. https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardprogram/kvalitetsdokument-for--patologi/#_ENREF_19
208. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-752. doi: 10.1038/35021093
209. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA. 2001;98(19):10869-10874. doi: 10.1073/pnas.191367098.
210. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817-2826. doi: 10.1056/NEJMoa041588 doi: 10.1056/NEJMoa041588
211. Sørlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci. 2003;100(14):8418-8423. doi: 10.1073/pnas.0932692100
212. Vieira AF, Schmitt F. An update on breast cancer multigene prognostic tests—emergent clinical biomarkers. Front Med (Lausanne). 2018;5:248. doi: 10.3389/fmed.2018.00248

213. Habel LA, Shak S, Jacobs MK, et al. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. Breast Cancer Res. 2006;8(3):1-15. doi: 10.1186/bcr1412
214. Buyse M, Loi S, Van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst. 2006;98(17):1183-1192. doi: 10.1093/jnci/djj329
215. Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the MINDACT trial. J Clin Oncol. 2008;26(5):729-735. doi: 10.1200/JCO.2007.14.3222
216. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160. doi: 10.1200/JCO.2008.18.1370.
217. Nielsen TO, Parker JS, Leung S, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. Clin Cancer Res. 2010;16(21):5222-5232. doi: 10.1158/1078-0432.CCR-10-1282
218. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol. 2011;5(1):5-23. doi: 10.1016/j.molonc.2010.11.003
219. Pons L, Hernández-León L, Altaleb A, et al. Conventional and digital Ki67 evaluation and their correlation with molecular prognosis and morphological parameters in luminal breast cancer. Sci Rep. 2022;12(1):1-9. doi: 10.1038/s41598-022-11411-5.
220. Thakur SS, Li H, Chan AM, et al. The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. PLoS One. 2018;13(1):e0188983 doi: 10.1371/journal.pone.0188983
221. Paik S, Kwon Y, Lee MH, et al. Systematic evaluation of scoring methods for Ki67 as a surrogate for 21-gene recurrence score. NPJ Breast Cancer. 2021;7(1):1-8. doi: 10.1038/s41523-021-00221-z
222. Pascual T, Perrone G, Morales S, et al. Limitations in predicting PAM50 intrinsic subtype and risk of relapse score with Ki67 in estrogen receptor-positive HER2-negative breast cancer. 2017;8(13):21930. doi: 10.18632/oncotarget.15748
223. Baskota SU, Dabbs DJ, Clark BZ, Bhargava R. Prosigna® breast cancer assay: histopathologic correlation, development, and assessment of size, nodal status, Ki-67 (SiNK™) index for breast cancer prognosis. Mod Pathol. 2021;34(1):70-76. doi: 10.1038/s41379-020-0643-8

224. Noske A, Anders S-I, Ettl J, et al. Risk stratification in luminal-type breast cancer: Comparison of Ki-67 with EndoPredict test results. Breast. 2020;49:101-107. doi: 10.1016/j.breast.2019.11.004
225. Sahebjam S, Aloyz R, Pilavdzic D, et al. Ki 67 is a major, but not the sole determinant of Oncotype Dx recurrence score. Br J Cancer. 2011;105(9):1342-1345. doi:10.1038/bjc.2011.402
226. Nielsen TO, Leung SCY, Rimm DL, et al. Assessment of Ki67 in breast cancer: updated recommendations from the international Ki67 in breast cancer working group. J Natl Cancer Inst. 2021;113(7):808-819. doi: 10.1093/jnci/djaa201
227. White R. Some Remarks on the Nature and Treatment of Cancers. Lond Med J. 1784;5(1):70-75.
228. Shah SM, Khan RA, Arif S, Sajid U. Artificial intelligence for breast cancer analysis: Trends & directions. Comput Biol Med. 2022;142:105221. doi: 10.1016/j.compbiomed.2022.105221
229. Wetstein SC, de Jong VM, Stathonikos N, et al. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. Sci Rep. 2022;12(1):15102. doi: 10.1038/s41598-022-19112-9
230. Jaroensri R, Wulczyn E, Hegde N, et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. NPJ Breast Cancer. 2022;8(1):113. doi: 10.1038/s41523-022-00478-y
231. Polónia A, Campelos S, Ribeiro A, et al. Artificial intelligence improves the accuracy in histologic classification of breast lesions. Am J Clin Pathol. 2021;155(4):527-536. doi: 10.1093/ajcp/aqaa151
232. MItosis DOmain Generlization Challenge 2022. MIDOG 2022. Grand Challenge. Accessed: February 11, 2024. https://midog2022.grand-challenge.org/
233. Aubreville M, Stathonikos N, Bertram CA, et al. Mitosis domain generalization in histopathology images—the MIDOG challenge. Med Image Anal. 2023;84:102699. doi: 10.1016/j.media.2022.102699
234. Veta M, Van Diest PJ, Willems SM, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med Image Anal. 2015;20(1):237-248. doi: 10.1016/j.media.2014.11.010
235. Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. Med Image Anal. 2019;54:111-121. doi: 10.1016/j.media.2019.02.012

236. Mahmood T, Arsalan M, Owais M, Lee MB, Park KR. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. J Clin Med. 2020;9(3):749. doi: 10.3390/jcm9030749
237. Paeng K, Hwang S, Park S, Kim M. A unified framework for tumor proliferation score prediction in breast histopathology. In Cardoso M, et al. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Vol. 10553. Springer, Cham. 2017:231-239. https://doi.org/10.1007/978-3-319-67558-9_27
238. Chen H, Dou Q, Wang X, Qin J, Heng P. Mitosis detection in breast cancer histology images via deep cascaded networks. Proceedings of the AAAI Conference on Artificial Intelligence. 2016;30(1) https://doi.org/10.1609/aaai.v30i1.10140
239. Jiménez G, Racoceanu D. Deep learning for semantic segmentation vs. classification in computational pathology: application to mitosis analysis in breast cancer grading. Front Bioeng Biotechnol. 2019;7:145. doi: 10.3389/fbioe.2019.00145
240. Balkenhol MC, Tellez D, Vreuls W, et al. Deep learning assisted mitotic counting for breast cancer. Lab Invest. 2019;99(11):1596-1606. doi: 10.1038/s41374-019-0275-0
241. Tellez D, Balkenhol M, Otte-Höller I, et al. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. IEEE Trans Med Imaging. 2018;37(9):2126-2136. doi: 10.1109/TMI.2018.2820199.
242. Fernandez-Martín C, Kiraz U, Silva-Rodríguez J, Morales S, Janssen EA, Naranjo V. Challenging mitosis detection algorithms: Global labels allow centroid localization. Springer; 2022:482-490.
243. Fernandez-Martín C, Silva-Rodriguez J, Kiraz U, Morales S, Janssen EA, Naranjo V. Uninformed Teacher-Student for hard-samples distillation in weakly supervised mitosis localization. Computerized Medical Imaging and Graphics. 2024;112:102328.
244. Pantanowitz L, Hartman D, Qi Y, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. Diagn Pathol. 2020;15(1):1-10. doi: 10.1186/s13000-020-00995-z
245. Bertram CA, Aubreville M, Donovan TA, et al. Computer-assisted mitotic count using a deep learning–based algorithm improves interobserver reproducibility and accuracy. Vet Pathol. 2022;59(2):211-226. doi: 10.1177/03009858211067478

246. Flach RN, Fransen NL, Sonnen AF, et al. Implementation of Artificial Intelligence in Diagnostic Practice as a Next Step after Going Digital: The UMC Utrecht Perspective. Diagnostics (Basel). 2022;12(5):1042. doi: 10.3390/diagnostics12051042
247. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol. 2018;42(12):1636. doi: 10.1097/PAS.0000000000001151
248. Challa B, Tahir M, Hu Y, et al. Artificial Intelligence–Aided Diagnosis of Breast Cancer Lymph Node Metastasis on Histologic Slides in a Digital Workflow. Mod Pathol. 2023;36(8):100216. doi: 10.1016/j.modpat.2023.100216
249. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep. 2017;7(1):16878. doi: 10.1038/s41598-017-17204-5
250. Andersen NL, Brügmann A, Lelkaitis G, Nielsen S, Lippert MF, Vyberg M. Virtual double staining: a digital approach to immunohistochemical quantification of estrogen receptor protein in breast carcinoma specimens. Appl Immunohistochem Mol Morphol. 2018;26(9):620-626. doi: 10.1097/PAI.0000000000000502
251. Couture HD, Williams LA, Geradts J, 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):30. doi: 10.1038/s41523-018-0079-1
252. 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(7):e197700-e197700. doi: 10.1001/jamanetworkopen.2019.7700
253. Klauschen F, Wienert S, Schmitt WD, et al. Standardized Ki67 diagnostics using automated scoring—clinical validation in the GeparTrio breast cancer study. Clin Cancer Res. 2015;21(16):3651-3657. doi: 10.1158/1078-0432.CCR-14-1283
254. Zhong F, Bi R, Yu B, Yang F, Yang W, Shui R. A comparison of visual assessment and automated digital image analysis of Ki67 labeling index in breast cancer. 2016;11(2):e0150505. doi: 10.1371/journal.pone.0150505
255. 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(1):33-42. doi: 10.1007/s10549-018-4669-2

256. Dy A, Nguyen N-NJ, Meyer J, et al. AI improves accuracy, agreement and efficiency of pathologists for Ki67 assessments in breast cancer. Sci Rep. 2024;14(1):1283. doi: 10.1038/s41598-024-51723-2
257. Stålhammar G, Martinez NF, Lippert M, et al. Digital image analysis outperforms manual biomarker assessment in breast cancer. Mod Pathol. 2016;29(4):318-329. doi: 10.1038/modpathol.2016.34
258. Stålhammar G, Robertson S, Wedlund L, et al. Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. Histopathology. 2018;72(6):974-989. doi: 10.1111/his.13452
259. Holten-Rossing H, Møller Talman M-L, Kristensson M, Vainer B. Optimizing HER2 assessment in breast cancer: application of automated image analysis. Breast Cancer Res Treat. 2015;152:367-375. doi: 10.1007/s10549-015-3475-3
260. Saltz J, Gupta R, Hou L, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 2018;23(1):181-193.e7. doi: 10.1016/j.celrep.2018.03.086
261. Heindl A, Sestak I, Naidoo K, Cuzick J, Dowsett M, Yuan Y. Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. J Natl Cancer Inst. 2018;110(2):166-175. doi: 10.1093/jnci/djx137
262. Thagaard J, Stovgaard ES, Vognsen LG, et al. Automated quantification of sTIL density with H&E-based digital image analysis has prognostic potential in triple-negative breast cancers. Cancers (Basel). 2021;13(12):3050. doi: 10.3390/cancers13123050
263. Bai Y, Cole K, Martinez-Morilla S, et al. An open-source, automated tumor-infiltrating lymphocyte algorithm for prognosis in triple-negative breast cancer. Clin Cancer Res. 2021;27(20):5557-5565. doi: 10.1158/1078-0432.CCR-21-0325
264. Sun P, He J, Chao X, et al. A computational tumor-infiltrating lymphocyte assessment method comparable with visual reporting guidelines for triple-negative breast cancer. EBioMedicine. 2021;70:103492 doi: 10.1016/j.ebiom.2021.103492
265. SEER Training Modules. Female Reproductive System. National Cancer Institute. Accessed 20.01.2024, https://training.seer.cancer.gov/anatomy/reproductive/female/

266. Reed BG, Carr BR. The normal menstrual cycle and the control of ovulation. [Updated August 5, 2018]. In: Fiengold KR, Anwalt B, Blackman MR, et al. Endotext [Internet]. South Dartmouth (MA). MDText.com; 2000. Accessed February 12, 2024. https://www.ncbi.nlm.nih.gov/books/NBK279054/
267. Yu K, Huang Z-Y, Xu X-L, Li J, Fu X-W, Deng S-L. Estrogen receptor function: Impact on the human endometrium. Front Endocrinol (Lausanne). 2022;13:827724. doi: 10.3389/fendo.2022.827724
268. Taraborrelli S. Physiology, production and action of progesterone. Acta Obstet Gynecol Scand. 2015;94:8-16. doi: 10.1111/aogs.12771
269. Orlowski M, Sarao MS. Physiology, follicle stimulating hormone. [Updated May 1, 2023], In: StatPearls [Internet]. StatPearls Publishing. Accessed: February 12, 2024. https://www.ncbi.nlm.nih.gov/books/NBK535442/
270. Lortet-Tieulent J, Ferlay J, Bray F, Jemal AJ. International patterns and trends in endometrial cancer incidence, 1978–2013. 2018;110(4):354-361. doi: 10.1093/jnci/djx214
271. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E. Endometrial cancer. 2016;387(10023):1094-1108. doi: 10.1016/S0140-6736(15)00130-0
272. Wiegratz I, Kuhl H. Endometrial cancer and hormone-replacement therapy. Lancet. 2005;366(9481):201-202. doi: 10.1016/S0140-6736(05)66902-4
273. Grady D, Gebretsadik T, Kerlikowske K, Ernster V, Petitti D. Hormone replacement therapy and endometrial cancer risk: a meta-analysis. Obstet Gynecol. 1995;85(2):304-313. doi: 10.1016/0029-7844(94)00383-O
274. Furness S, Roberts H, Marjoribanks J, Lethaby A, Hickey M, Farquhar C. Hormone therapy in postmenopausal women and risk of endometrial hyperplasia. Cochrane Database Syst Rev. 2009;15(2):CD000402. doi: 10.1002/14651858.CD000402.pub4
275. World Health Organisation of Tumours Editorial Board. Female Genital Tumours. 5th ed. vol 4. World Health Organisation Classification of Tumours. International Agency for Research on Cancer; 2020.
276. Guerre-Millo M. Adipose tissue hormones. J Endocrinol Invest. 2002;25:855-861. https://doi.org/10.1007/BF03344048
277. Zhang S, Gong T-T, Liu F-H, et al. Global, Regional, and National Burden of Endometrial Cancer, 1990–2017: Results From the Global Burden of Disease Study, 2017. Front Oncol. 2019;9:1440. doi: 10.3389/fonc.2019.01440
278. Schouten LJ, Goldbohm RA, Van Den Brandt PA. Anthropometry, physical activity, and endometrial cancer risk: results from the Netherlands Cohort Study. J Natl Cancer Inst. 2004;96(21):1635-1638. doi: 10.1093/jnci/djh291

279. Trentham-Dietz A, Nichols H, Hampton J, Newcomb P. Weight change and risk of endometrial cancer. Int J Epidemiol. 2006;35(1):151-158. doi: 10.1093/ije/dyi226
280. Hosono S, Matsuo K, Hirose K, et al. Weight gain during adulthood and body weight at age 20 are associated with the risk of endometrial cancer in Japanese women. J Epidemiol. 2011;21(6):466-473. doi: 10.2188/jea.je20110020
281. Stevens VL, Jacobs EJ, Patel AV, Sun J, Gapstur SM, McCullough ML. Body weight in early adulthood, adult weight gain, and risk of endometrial cancer in women not using postmenopausal hormones. Cancer Causes Control. 2014;25(3):321-328. doi: 10.1007/s10552-013-0333-7
282. Cohen I. Endometrial pathologies associated with postmenopausal tamoxifen treatment. Gynecol Oncol. 2004;94(2):256-266. doi: 10.1016/j.ygyno.2004.03.048
283. Fisher B, Costantino JP, Redmond CK, et al. Endometrial cancer in tamoxifen-treated breast cancer patients: findings from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14. J Natl Cancer Inst. 1994;86(7):527-537. doi: 10.1093/jnci/86.7.527
284. Neri F, Maggino T. Surveillance of endometrial pathologies, especially for endometrial cancer, of breast cancer patients under tamoxifen treatment. Eur J Gynaec Oncol. 2009;30(4):357-360.
285. Helsedirektoratet. Nasjonalt handlingsprogram med retningslinjer for gynekologisk kreft. [National Guidelines for Gynecological Cancer]. Updated June 28, 2021. Accessed February 12, 2024. https://www.helsedirektoratet.no/retningslinjer/gynekologisk-kreft--handlingsprogram?tidligere-versjoner
286. Zhao J, Hu Y, Zhao Y, Chen D, Fang T, Ding M. Risk factors of endometrial cancer in patients with endometrial hyperplasia: implication for clinical treatments. BMC Womens Health. 2021;21(1):312. doi: 10.1186/s12905-021-01452-9
287. Dossus L, Allen N, Kaaks R, et al. Reproductive risk factors and endometrial cancer: the European Prospective Investigation into Cancer and Nutrition. Int J Cancer. 2010;127(2):442-451. doi: 10.1002/ijc.25050
288. Kokts-Porietis RL, Elmrayed S, Brenner DR, Friedenreich CM. Obesity and mortality among endometrial cancer survivors: a systematic review and meta-analysis. Obes Rev. 2021;22(12):e13337. doi: 10.1111/obr.13337

289. Friberg E, Orsini N, Mantzoros C, Wolk A. Diabetes mellitus and risk of endometrial cancer: a meta-analysis. Diabetologia. 2007;50(7):1365-1374. doi: 10.1007/s00125-007-0681-5
290. Saed L, Varse F, Baradaran HR, et al. The effect of diabetes on the risk of endometrial Cancer: an updated a systematic review and meta-analysis. BMC Cancer. 2019;19(1):527. doi: 10.1186/s12885-019-5748-4
291. Amiri M, Bidhendi-Yarandi R, Fallahzadeh A, Marzban Z, Tehrani FR. Risk of endometrial, ovarian, and breast cancers in women with polycystic ovary syndrome: A systematic review and meta-analysis. Int J Reprod Biomed. 2022;20(11):893-914. doi: 10.18502/ijrm.v20i11.12357
292. Li Z, Wang Y-H, Wang L-L, et al. Polycystic ovary syndrome and the risk of endometrial, ovarian and breast cancer: An updated meta-analysis. Scott Med J. 2022;67(3):109-120. doi: 10.1177/00369330221107099
293. Haoula Z, Salman M, Atiomo W. Evaluating the association between endometrial cancer and polycystic ovary syndrome. Human Reprod. 2012;27(5):1327-1331. doi: 10.1093/humrep/des042
294. Buchanan DD, Rosty C, Clendenning M, Spurdle AB, Win AK. Clinical problems of colorectal cancer and endometrial cancer cases with unknown cause of tumor mismatch repair deficiency (suspected Lynch syndrome). Appl Clin Genet. 2014;7:183-193. doi: 10.2147/TACG.S48625
295. Tan M-H, Mester JL, Ngeow J, Rybicki LA, Orloff MS, Eng C. Lifetime cancer risks in individuals with germline PTEN mutations. Clin Cancer Res. 2012;18(2):400-407. doi: 10.1158/1078-0432.CCR-11-228
296. Russo M, Newell JM, Budurlean L, et al. Mutational profile of endometrial hyperplasia and risk of progression to endometrioid adenocarcinoma. Cancer. 2020;126(12):2775-2783. doi: 10.1002/cncr.32822
297. Nees LK, Heublein S, Steinmacher S, et al. Endometrial hyperplasia as a risk factor of endometrial cancer. Arch Gynecol Obstet. 2022;306(2):407-421. doi: 10.1007/s00404-021-06380-5
298. Kurman RJ, Kaminski PF, Norris HJ. The behavior of endometrial hyperplasia. A long-term study of “untreated” hyperplasia in 170 patients. Cancer. 1985;56(2):403-412. doi: 10.1002/1097-0142(19850715)56:2<403::aid-cncr2820560233>3.0.co;2-x
299. Baak J, Mutter G. EIN and WHO94. J Clin Pathol. 2005;58(1):1-6. doi: 10.1136/jcp.2004.021071

300. Allison KH, Reed SD, Voigt LF, Jordan CD, Newton KM, Garcia RL. Diagnosing endometrial hyperplasia: why is it so difficult to agree? Am J Surg Pathol. 2008;32(5):691-8. doi: 10.1097/PAS.0b013e318159a2a0
301. Skov B, Broholm H, Engel U, et al. Comparison of the reproducibility of the WHO classifications of 1975 and 1994 of endometrial hyperplasia. Int J Gynecol Pathol. 1997;16(1):33-37. doi: 10.1097/00004347-199701000-00006
302. Bergeron C, Nogales FF, Masseroli M, et al. A multicentric European study testing the reproducibility of the WHO classification of endometrial hyperplasia with a proposal of a simplified working classification for biopsy and curettage specimens. Am J Surg Pathol. 1999;23(9):1102-1108. doi: 10.1097/00000478-199909000-00014
303. Kendall BS, Ronnett BM, Isacson C, et al. Reproducibility of the diagnosis of endometrial hyperplasia, atypical hyperplasia, and well-differentiated carcinoma. Am J Surg Pathol. 1998;22(8):1012-1019. doi: 10.1097/00000478-199808000-00012.
304. Mutter GL. Endometrial intraepithelial neoplasia (EIN): will it bring order to chaos? Gynecologic oncology. 2000;76(3):287-290. doi: 10.1006/gyno.1999.5580
305. Tavassoli F DP, editors. Tumours of the Breast and Female Genital Organs. World Health Organisation Classification of Tumours. IARC Press; 2003.
306. Kurman RJ, Carcangiu ML, Herrington CS, Young RH, editors. WHO Classification of Tumours of Female Reproductive Organs. 4th ed. International Agency for Research on Cancer; 2014.
307. Hecht JL, Ince TA, Baak JP, Baker HE, Ogden MW, Mutter GL. Prediction of endometrial carcinoma by subjective endometrial intraepithelial neoplasia diagnosis. Mod Pathol. 2005;18(3):324-330. doi: 10.1038/modpathol.3800328
308. Ordi J, Bergeron C, Hardisson D, et al. Reproducibility of current classifications of endometrial endometrioid glandular proliferations: further evidence supporting a simplified classification. Histopathology. 2014;64(2):284-292. doi: 10.1111/his.12249.
309. Sherman ME, Ronnett BM, Ioffe OB, et al. Reproducibility of biopsy diagnoses of endometrial hyperplasia: evidence supporting a simplified classification. Int J Gynecol Pathol. 2008;27(3):318-325. doi: 10.1097/PGP.0b013e3181659167

310. Moore E, Shafi M. Endometrial hyperplasia. Obstetrics, Gynaecology & Reproductive Medicine. 2013;23(3):88-93. https://doi.org/10.1016/j.ogrm.2013.01.002
311. Gallos ID AM, Clark T, Faraj R, Rosenthal A, Smith P GJ. Green-top Guideline: Management of Endometrial Hyperplasia. No.67. February 2016. Royal College of Obstetricians & Gynaecologists. Accessed Feburary 2, 2024. https://www.rcog.org.uk/globalassets/documents/guidelines/green-top-guidelines/gtg_67_endometrial_hyperplasia.pdf
312. Lacey Jr JV, Sherman ME, Rush BB, et al. Absolute risk of endometrial carcinoma during 20-year follow-up among women with endometrial hyperplasia. J Clin Oncol. 2010;28(5):788. doi: 10.1200/JCO.2009.24.1315
313. Trimble CL, Kauderer J, Zaino R, et al. Concurrent endometrial carcinoma in women with a biopsy diagnosis of atypical endometrial hyperplasia. Cancer. 2006;106(4):812-819. doi: 10.1002/cncr.21650
314. Reed SD, Newton KM, Garcia RL, et al. Complex hyperplasia with and without atypia: clinical outcomes and implications of progestin therapy. Obstet Gynecol. 2010;116(2 Pt 1):365-373. doi: 10.1097/AOG.0b013e3181e93330
315. Creasman WT, Mannel RS, Mutch DG, Tewari K. Disaia and Creasman Clinical Gynecologic Oncology, E- Book. 10th ed. Elsevier Health Sciences; 2022.
316. Urban RR, Reed SD. Endometrial hyperplasia: Management and prognosis. UpToDate. Updated: January 25, 2024. Accessed: February 12, 2024. https://www.uptodate.com/contents/endometrial-hyperplasia-management-and-prognosis
317. Chen Y-L, Cheng W-F, Lin M-C, Huang C-Y, Hsieh C-Y, Chen C-A. Concurrent endometrial carcinoma in patients with a curettage diagnosis of endometrial hyperplasia. J Formos Med Assoc. 2009;108(6):502-507. doi: 10.1016/S0929-6646(09)60098-X
318. Doherty MT, Sanni OB, Coleman HG, et al. Concurrent and future risk of endometrial cancer in women with endometrial hyperplasia: A systematic review and meta-analysis. 2020;15(4):e0232231. doi: 10.1371/journal.pone.0232231
319. Baak JP, Nauta J, Wisse-Brekelmans E, Bezemer P. Architectural and nuclear morphometrical features together are more important prognosticators in endometrial hyperplasias than nuclear morphometrical features alone. J Pathol. 1988;154(4):335-341. doi: 10.1002/path.1711540409

320. Baak J, Kuik D, Bezemer P. The additional prognostic value of morphometric nuclear arrangement and DNA-ploidy to other morphometric and stereologic features in endometrial hyperplasias. Int J Gynecol Cancer. 1994;4(5):289-297. doi: 10.1046/j.1525-1438.1994.04050289.x
321. Baak JP, Snijders W, van Dierman B, van Diest PJ, Diepenhorst FW, Benraadt J. Prospective multicenter validation confirms the pronostic superiority of the Endometrial Carcinoma Prognostic Index in International Federation of Gynecology and Obstetrics Stage 1 and 2 Endometrial Carcinoma. J Clin Oncol. 2003;21(22):4214-4221. doi: 10.1200/JCO.2003.02.087
322. Baak J, Wisse-Brekelmans E, Fleege J, Van Der Putten H, Bezemer P. Assessment of the risk on endometrial cancer in hyperplasia, by means of morphological and morphometrical features. Pathol Res Pract. 1992;188(7):856-859. doi: 10.1016/S0344-0338(11)80244-X
323. Baak JP, Mutter GL, Robboy S, et al. The Molecular Genetics and Morphometry-Based Endometrial Intraepithelial Neoplasia Classification System Predicts Disease Progression in Endometrial Hyperplasia More Accurately than the 1994 World Health Organization Classification System. Cancer. 2005;103(11):2304-2312. doi: 10.1002/cncr.21058
324. Baak JPA, Ørbo A, Van Diest PJ, et al. Prospective multicenter evaluation of the mophometric D-score for Prediction of the Outcome of Endometrial Hyperplasias. Am J Surg Pathol. 2001;25(7):930-935. doi: 10.1097/00000478-200107000-00012
325. Dunton CJ, Baak JP, Palazzo JP, van Diest PJ, McHugh M, Widra EA. Use of computerized morphometric analyses of endometrial hyperplasias in the prediction of coexistent cancer. Am J Obstet Gynecol. 1996;174(5):1518-1521. doi: 10.1016/s0002-9378(96)70599-9
326. Mutter GL, Baak J, Crum CP, Richart RM, Ferenczy A, Faquin WC. Endometrial precancer diagnosis by histopathology, clonal analysis, and computerized morphometry. J Pathology. 2000;190(4):462-469. doi: 10.1002/(SICI)1096-9896(200003)190:4<462::AID-PATH590>3.0.CO;2-D
327. Ørbo A, Baak JP, Kleivan I, et al. Computerised morphometrical analysis in endometrial hyperplasia for the prediction of cancer development. A long term retrospective study from northern Norway. J Clin Pathol. 2000;53(9):697-703. doi: 10.1136/jcp.53.9.697
328. Helsedirektoratet. Nasjonalt handlingsprogram med retningslinjer for gynekologisk kreft. [National Guidelines for Gynecological Cancer]. Updated June 30, 2023. Accessed January 16, 2024.

https://www.helsedirektoratet.no/retningslinjer/gynekologisk-kreft--handlingsprogram?tidligere-versjoner

329. Raffone A, Travaglino A, Saccone G, et al. Endometrial hyperplasia and progression to cancer: which classification system stratifies the risk better? A systematic review and meta-analysis. Arch Gynecol Obstet. 2019;299(5):1233-1242. doi: 10.1007/s00404-019-05103-1
330. Lacey Jr JV, Mutter GL, Nucci MR, et al. Risk of subsequent endometrial carcinoma associated with endometrial intraepithelial neoplasia classification of endometrial biopsies. 2008;113(8):2073-2081. doi: 10.1002/cncr.23808
331. Salman MC, Usubutun A, Boynukalin K, Yuce K. Comparison of WHO and endometrial intraepithelial neoplasia classifications in predicting the presence of coexistent malignancy in endometrial hyperplasia. J Gynecol Oncol. 2010;21(2):97-101. doi: 10.3802/jgo.2010.21.2.97
332. Travaglino A, Raffone A, Saccone G, et al. Endometrial hyperplasia and the risk of coexistent cancer: WHO versus EIN criteria. 2019;74(5):676-687. doi: 10.1111/his.13776
333. Taylor Jr HC. Endometrial hyperplasia and carcinoma of the body of the uterus. American Journal of Obstetrics and Gynecology. 1932;23(3):309-332. https://doi.org/10.1016/S0002-9378(32)90820-5
334. Sanderson PA, Critchley HO, Williams AR, Arends MJ, Saunders PT. New concepts for an old problem: the diagnosis of endometrial hyperplasia. Hum Reprod Update. 2017;23(2):232-254. doi: 10.1093/humupd/dmw042
335. Mutter GL, Zaino RJ, Baak JP, Bentley RC, Robboy SJ. Benign endometrial hyperplasia sequence and endometrial intraepithelial neoplasia. Int J Gynecol Pathol. 2007;26(2):103-114. doi: 10.1097/PGP.0b013e31802e4696
336. Key T, Pike M. The dose-effect relationship between'unopposed'oestrogens and endometrial mitotic rate: its central role in explaining and predicting endometrial cancer risk. Br J Cancer. 1988;57(2):205-212. doi: 10.1038/bjc.1988.44
337. Allen N, Yang T, Olsson H. Endometrial cancer and oral contraceptives: an individual participant meta-analysis of 27 276 women with endometrial cancer from 36 epidemiological studies. Lancet Oncol. 2015;16(9):1061-1070. doi: 10.1016/S1470-2045(15)00212-0
338. Hecht JL, Ince TA, Baak JP, Baker HE, Ogden MW, Mutter G. Prediction of endometrial carcinoma by subjective endometrial intraepithelial neoplasia diagnosis. 2005;18(3):324-330. doi: 10.1038/modpathol.3800328.

339. Jovanovic AS, Boynton KA, Mutter GL. Uteri of women with endometrial carcinoma contain a histopathological spectrum of monoclonal putative precancers, some with microsatellite instability. Cancer Res. 1996;56(8):1917-1921.
340. Sanderson PA, Critchley HO, Williams AR, Arends MJ, Saunders PT. New concepts for an old problem: the diagnosis of endometrial hyperplasia. Human reproduction update. 2017;23(2):232-254.
341. Owings RA, Quick CM. Endometrial intraepithelial neoplasia. Arch Pathol Lab Med. 2014;138(4):484-491. doi: 10.5858/arpa.2012-0709-RA
342. Li J, Yen C, Liaw D, et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science. 1997;275(5308):1943-1947. doi: 10.1126/science.275.5308.1943
343. Steck PA, Pershouse MA, Jasser SA, et al. Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23. 3 that is mutated in multiple advanced cancers. Nature Genet. 1997;15(4):356-362. doi: 10.1038/ng0497-356
344. Kurose K, Zhou X-P, Araki T, Cannistra SA, Maher ER, Eng C. Frequent loss of PTEN expression is linked to elevated phosphorylated Akt levels, but not associated with p27 and cyclin D1 expression, in primary epithelial ovarian carcinomas. Am J Pathol. 2001;158(6):2097-2106. doi: 10.1016/S0002-9440(10)64681-0
345. Song MS, Salmena L, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor. Nat Rev Mol Cell Biol. 2012;13(5):283-296. doi: 10.1038/nrm3330.
346. Fruman DA, Chiu H, Hopkins BD, Bagrodia S, Cantley LC, Abraham RT. The PI3K pathway in human disease. Cell. 2017;170(4):605-635. doi: 10.1016/j.cell.2017.07.029.
347. Mutter GL, Lin M-C, Fitzgerald JT, Kum JB, Eng C. Changes in endometrial PTEN expression throughout the human menstrual cycle. J Clin Endocrinol Metab. 2000;85(6):2334-2338. doi: 10.1210/jcem.85.6.6652.
348. Scully MM, Palacios-Helgeson LK, Wah LS, Jackson TA. Rapid estrogen signaling negatively regulates PTEN activity through phosphorylation in endometrial cancer cells. Horm Cancer. 2014;5(4):218-231. doi: 10.1007/s12672-014-0184-z
349. Guzeloglu-Kayisli O, Kayisli UA, Al-Rejjal R, Zheng W, Luleci G, Arici A. Regulation of PTEN (phosphatase and tensin homolog deleted on chromosome 10) expression by estradiol and progesterone in human

endometrium. J Clin Endocrinol Metab. 2003;88(10):5017-5026. doi: 10.1210/jc.2003-030414.

350. Chen R, Zhang M, Liu W, et al. Estrogen affects the negative feedback loop of PTENP1-miR200c to inhibit PTEN expression in the development of endometrioid endometrial carcinoma. Cell Death Dis. 2018;10(1):4. doi: 10.1038/s41419-018-1207-4
351. Liaw D, Marsh DJ, Li J, et al. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet. 1997;16(1):64-67. doi: 10.1038/ng0597-64
352. Sletten ET, Arnes M, Lysa LM, Moe BT, Straume B, Orbo A. Prediction of Relapse After Therapy Withdrawal in Women with Endometrial Hyperplasia: A Long-term Follow-up Study. Anticancer Res. 2017;37(5):2529-2536. doi: 10.21873/anticanres.11595
353. Baak JPA, van Diermen B, Steinbakk A, et al. Lack of PTEN expression in endometrial intraepithelial neoplasia is correlated with cancer progression. Hum Pathol. 2005;36(5):555-561. doi: 10.1016/j.humpath.2005.02.018
354. Ørbo A, Nilsen M, Arnes M, Pettersen I, Larsen K. Loss of expression of MLH1, MSH2, MSH6, and PTEN related to endometrial cancer in 68 patients with endometrial hyperplasia. Int J Gynecol Pathol. 2003;22(2):141-148. doi: 10.1097/00004347-200304000-00005
355. Pavlakis K, Messini I, Vrekoussis T, et al. PTEN-loss and nuclear atypia of EIN in endometrial biopsies can predict the existence of a concurrent endometrial carcinoma. Gynecol Oncol. 2010;119(3):516-519. doi: 10.1016/j.ygyno.2010.08.023
356. Yang HP, Meeker A, Guido R, et al. PTEN expression in benign human endometrial tissue and cancer in relation to endometrial cancer risk factors. Cancer Causes Control. 2015;26(12):1729-1736. doi: 10.1007/s10552-015-0666-5
357. Lee H, Choi HJ, Kang CS, Lee HJ, Lee WS, Park CS. Expression of miRNAs and PTEN in endometrial specimens ranging from histologically normal to hyperplasia and endometrial adenocarcinoma. Mod Pathol. 2012;25(11):1508-1515. doi: 10.1038/modpathol.2012.111
358. Monte NM, Webster KA, Neuberg D, Dressler GR, Mutter GL. Joint loss of PAX2 and PTEN expression in endometrial precancers and cancer. Cancer Res. 2010;70(15):6225-6232. doi: 10.1158/0008-5472.CAN-10-0149

359. Lacey JV, Mutter GL, Ronnett BM, et al. PTEN expression in endometrial biopsies as a marker of progression to endometrial carcinoma. Cancer Res. 2008;68(14):6014-6020. doi: 10.1158/0008-5472.CAN-08-1154
360. Raffone A, Travaglino A, Saccone G, et al. PTEN expression in endometrial hyperplasia and risk of cancer: a systematic review and meta-analysis. Arch Gynecol Obstet. 2019;299(6):1511-1524. doi: 10.1007/s00404-019-05123-x
361. Travaglino A, Raffone A, Saccone G, et al. PTEN as a predictive marker of response to conservative treatment in endometrial hyperplasia and early endometrial cancer. A systematic review and meta-analysis. Eur J Obstet Gynecol Reprod Biol. 2018;231:104-110. doi: 10.1016/j.ejogrb.2018.10.025
362. Mutter GL, Ince TA, Baak JP, Kust GA, Zhou X-P, Eng C. Molecular identification of latent precancers in histologically normal endometrium. Cancer Res. 2001;61(11):4311-4314.
363. Shang Y. Hormones and cancer. Cell Res. 2007;17(4):277-279. doi: 10.1038/cr.2007.26
364. Lang D, Powell SK, Plummer RS, Young KP, Ruggeri BA. PAX genes: roles in development, pathophysiology, and cancer. Biochem Pharmacol. 2007;73(1):1-14. doi: 10.1016/j.bcp.2006.06.024
365. Wu H, Chen Y, Liang J, et al. Hypomethylation-linked activation of PAX2 mediates tamoxifen-stimulated endometrial carcinogenesis. Nature. 2005;438(7070):981-987. doi: 10.1038/nature04225
366. Chen H, Li L, Liu H, et al. PAX2 is regulated by estrogen/progesterone through promoter methylation in endometrioid adenocarcinoma and has an important role in carcinogenesis via the AKT/mTOR signaling pathway. J Pathol. 2024. [online ahead of print] doi: 10.1002/path.6249
367. Stuart ET, Haffner R, Oren M, Gruss P. Loss of p53 function through PAX-mediated transcriptional repression. EMBO J. 1995;14(22):5638-5645. doi: 10.1002/j.1460-2075.1995.tb00251.x
368. Allison KH, Upson K, Reed SD, et al. PAX2 loss by immunohistochemistry occurs early and often in endometrial hyperplasia. Int J Gynecol Pathol. 2012;31(2):151-159. doi: 10.1097/PGP.0b013e318226b376
369. Strissel PL, Ellmann S, Loprich E, et al. Early aberrant insulin-like growth factor signaling in the progression to endometrial carcinoma is augmented by tamoxifen. Int J Cancer. 2008;123(12):2871-2879. doi: 10.1002/ijc.23900
370. Trabzonlu L, Muezzinoglu B, Corakci A. BCL-2 and PAX2 expressions in EIN which had been previously diagnosed as non-atypical hyperplasia. Pathol Oncol Res. 2019;25(2):471-476. doi: 10.1007/s12253-017-0378-0

371. Kahraman K, Kiremitci S, Taskin S, Kankaya D, Sertcelik A, Ortac F. Expression pattern of PAX2 in hyperplastic and malignant endometrium. Arch Gynecol Obstet. 2012;286(1):173-178. doi: 10.1007/s00404-012-2236-3.
372. Quick CM, Laury AR, Monte NM, Mutter GL. Utility of PAX2 as a marker for diagnosis of endometrial intraepithelial neoplasia. Am J Clin Pathol. 2012;138(5):678-684. doi: 10.1309/AJCP8OMLT7KDWLMF
373. Upson K, Allison KH, Reed SD, et al. Biomarkers of progestin therapy resistance and endometrial hyperplasia progression. Am J Obstet Gynecol. 2012;207(1):36.e1- e8. doi: 10.1016/j.ajog.2012.05.012
374. Joiner AK, Quick CM, Jeffus SK. PAX2 expression in simultaneously diagnosed WHO and EIN classification systems. Int J Gynecol Pathol. 2015;34(1):40-46. doi: 10.1097/PGP.0000000000000185
375. Raffone A, Travaglino A, Saccone G, et al. PAX 2 in endometrial carcinogenesis and in differential diagnosis of endometrial hyperplasia: A systematic review and meta-analysis of diagnostic accuracy. 2019;98(3):287-299. doi: 10.1111/aogs.13512
376. Gilks CB, Oliva E, Soslow RA. Poor interobserver reproducibility in the diagnosis of high-grade endometrial carcinoma. Am J Surg Pathol. 2013;37(6):874-881. doi: 10.1097/PAS.0b013e31827f576a
377. McAlpine J, Leon-Castillo A, Bosse T. The rise of a novel classification system for endometrial carcinoma; integration of molecular subclasses. J Pathol. 2018;244(5):538-549. doi: 10.1002/path.5034
378. Han G, Sidhu D, Duggan MA, et al. Reproducibility of histological cell type in high-grade endometrial carcinoma. Mod Pathol. 2013;26(12):1594-1604. doi: 10.1038/modpathol.2013.102
379. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67-73. doi: 10.1038/nature12113
380. Stelloo E, Nout RA, Osse EM, et al. Improved risk assessment by integrating molecular and clinicopathological factors in early-stage endometrial cancer—combined analysis of the PORTEC cohorts. Clin Cancer Res. 2016;22(16):4215-4224. doi: 10.1158/1078-0432.CCR-15-2878
381. Talhouk A, McConechy MK, Leung S, et al. A clinically applicable molecular-based classification for endometrial cancers. Br J Cancer. 2015;113(2):299-310. doi: 10.1038/bjc.2015.190

382. Mais V, Peiretti M. Immunohistochemical Markers in Endometrial Cancer. Cancers (Basel). 2021;13(3):505. doi: 10.3390/cancers13030505
383. Mais V, Peiretti M. Immunohistochemical Markers in Endometrial Cancer: Latest Updates. Cancers (Basel). 2023;15(17):4202. doi: 10.3390/cancers15174202
384. Ring KL, Mills AM, Modesitt SC. Endometrial hyperplasia. Obstet Gynecol. 2022;140(6):1061-1075. doi: 10.1097/AOG.0000000000004989
385. Allison KH, Tenpenny E, Reed SD, Swisher EM, Garica RL. Immunohistochemical markers in endometrial hyperplasia: Is there a panel with promise?: A review. Appl Immunohistochem Mol Morphol. 2008;16(4):329-343. doi: 10.1097/PAI.0b013e318159b88e
386. Trimble CL, Leitao M, Lu K, et al. Management of endometrial precancers. Obstet Gynecol. 2012;120(5):1160-1175. doi: 10.1097/aog.0b013e31826bb121
387. Piedimonte S, Rosa G, Gerstl B, et al. Evaluating the use of machine learning in endometrial cancer: A systematic review. Int J Gynecol Cancer. 2023;33(9):1383-1393. doi: 10.1136/ijgc-2023-004622
388. Zhang Y, Gong C, Zheng L, Li X, Yang X. Deep Learning for Intelligent Recognition and Prediction of Endometrial Cancer. J Healthc Eng. 2021;1148309. doi: 10.1155/2021/1148309
389. Takahashi Y, Sone K, Noda K, et al. Automated system for diagnosing endometrial cancer by adopting deep-learning technology in hysteroscopy. PLoS One. 2021;16(3):e0248526. doi: 10.1371/journal.pone.0248526
390. Vlachokosta AA, Asvestas PA, Gkrozou F, Lavasidis L, Matsopoulos GK, Paschopoulos M. Classification of hysteroscopical images using texture and vessel descriptors. Med Biol Eng Comput. 2013;51(8):859-867. doi: 10.1007/s11517-013-1058-1
391. Li D, Hu R, Li H, et al. Performance of automatic machine learning versus radiologists in the evaluation of endometrium on computed tomography. Abdom Radiol (NY). 2021;46(11):5316-5324. doi: 10.1007/s00261-021-03210-9
392. Xia Z, Zhang L, Liu S, Ran W, Liu Y, Tu J. Deep learning-based hysteroscopic intelligent examination and ultrasound examination for diagnosis of endometrial carcinoma. J Supercomput. 2022;78:11229-11244.
393. Zhang Y, Wang Z, Zhang J, et al. Deep learning model for classifying endometrial lesions. J Transl Med. 2021;19(1):10. doi: 10.1186/s12967-020-02660-x

394. Downing MJ, Papke Jr DJ, Tyekucheva S, Mutter GL. A new classification of benign, premalignant, and malignant endometrial tissues using machine learning applied to 1413 candidate variables. Int J Gynecol Pathol. 2020;39(4):333-343. doi: 10.1097/PGP.0000000000000615
395. Papke Jr DJ, Lohmann S, Downing M, Hufnagl P, Mutter GL. Computational augmentation of neoplastic endometrial glands in digital pathology displays. J Pathol. 2020;253(3):257-267. doi: 10.1002/path.5586.
396. Fell C, Mohammadi M, Morrison D, et al. Detection of malignancy in whole slide images of endometrial cancer biopsies using artificial intelligence. PLoS One. 2023;18(3):e0282577. doi: 10.1371/journal.pone.0282577
397. Liao X, Zheng X, He J, Li Q. Computer-aided decision-making system for endometrial atypical hyperplasia based on multi-modal and multi-instance deep convolution neural networks. Soft Computing. 2021. https://doi.org/10.1007/s00500-021-06576-6
398. Rewcastle E, Gudlaugsson E, Lillesand M, Skaland I, Baak JP, Janssen EA. Automated Prognostic Assessment of Endometrial Hyperplasia for Progression Risk Evaluation Using Artificial Intelligence. Mod Pathol. 2023;36(5):100116.
399. Cheng J, Liu Y, Huang W, Hong W, Wang L, Ni D. Identifying novel prognostic markers and genotype-phenotype associations in endometrioid endometrial carcinoma by computational analysis of histopathological images. Medicine in Omics. 2021;1:100005. https://doi.org/10.1016/j.meomic.2021.100005
400. Makris GM, Pouliakis A, Siristatidis C, et al. Image analysis and multi-layer perceptron artificial neural networks for the discrimination between benign and malignant endometrial lesions. Diagn Cytopathol. 2017;45(3):202-211. doi: 10.1002/dc.23649
401. Erdemoglu E, Serel TA, Karacan E, et al. Artificial Intelligence for Prediction of Endometrial Intraepithelial Neoplasia/Endometrial Cancer Risk in Pre-and Postmenopausal Women. AJOG Glob Rep. 2023;3(1):100154. doi: 10.1016/j.xagr.2022.100154.
402. Giannella L, Cerami LB, Setti T, Bergamini E, Boselli F. Prediction of endometrial hyperplasia and cancer among premenopausal women with abnormal uterine bleeding. Biomed Res Int. 2019:8598152 doi: 10.1155/2019/8598152

403. Kuai D, Tang Q, Tian W, Zhang H. Rapid identification of endometrial hyperplasia and endometrial endometrioid cancer in young women. Discov Oncol. 2023;14(1):121. doi: 10.1007/s12672-023-00736-w
404. Ruan H, Chen S, Li J, et al. Development and Validation of a Nomogram Prediction Model for Endometrial Malignancy in Patients with Abnormal Uterine Bleeding. Yonsei Med J. 2023;64(3):197-203. doi: 10.3349/ymj.2022.0239
405. Quasar Collaborative Group, Gray R, Barnwell J, et al. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. Lancet. 2007;370(9604):2020-2029. doi: 10.1016/S0140-6736(07)61866-2
406. Mitchard JR, Love SB, Baxter KJ, Shepherd NA. How important is peritoneal involvement in rectal cancer? A prospective study of 331 cases. Histopathology. 2010;57(5):671-679. doi: 10.1111/j.1365-2559.2010.03687.x
407. Ersvær E, Hveem TS, Vlatkovic L, et al. Prognostic value of DNA ploidy and automated assessment of stroma fraction in prostate cancer. Int J Cancer. 2020;147(4):1228-1234. doi: 10.1002/ijc.32832
408. Ersvær E, Kildal W, Vlatkovic L, et al. Prognostic value of mitotic checkpoint protein BUB3, cyclin B1, and pituitary tumor-transforming 1 expression in prostate cancer. Mod Pathol. 2020;33(5):905-915. doi: 10.1038/s41379-019-0418-2
409. Andersen S, Richardsen E, Nordby Y, et al. Disease-specific outcomes of radical prostatectomies in Northern Norway; a case for the impact of perineural infiltration and postoperative PSA-doubling time. BMC Urol. 2014;14:1-11. doi: 10.1186/1471-2490-14-49
410. Rakaee M, Busund L-TR, Jamaly S, et al. Prognostic value of macrophage phenotypes in resectable non–small cell lung cancer assessed by multiplex immunohistochemistry. Neoplasia. 2019;21(3):282-293. doi: 10.1016/j.neo.2019.01.005
411. Kvikstad V, Mangrud OM, Gudlaugsson E, et al. Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas. Diagn Pathol. 2019;14(1):90. doi: 10.1186/s13000-019-0868-3.
412. Lillesand M, Kvikstad V, Mangrud OM, et al. Mitotic activity index and CD25+ lymphocytes predict risk of stage progression in non-muscle invasive

bladder cancer. PLoS One. 2020;15(6):e0233676. doi: 10.1371/journal.pone.0233676

413. Petersen V, Baxter K, Love S, Shepherd N. Identification of objective pathological prognostic determinants and models of prognosis in Dukes' B colon cancer. Gut. 2002;51(1):65-69. doi: 10.1136/gut.51.1.65
414. Hveem TS IM, Kalsnes J, Julbø F, Pradha M, Kleppe A, De Raedt S, Skrede OJ, Torheim T, Nesheim JA, Mohn HM, Askautrud HA, Cyll K, Kildal W, Rewcastle E, Lillesand M, Kvikstad V, Janssen E, Jones R, Brustugun OT, Wæhre H, Brennhovd B, Haug ES, Busund LTR, Lindemann K, Kristensen G, Shepherd NA, Novelli M, Liestøl K, Kerr D, Danielsen HE. Applicability of mitotic figure counting by deep learning: a development and pan-cancer validation study. Unpublished Work; 2024.
415. Rewcastle E, Varhaugvik AE, Gudlaugsson E, et al. Assessing the prognostic value of PAX2 and PTEN in endometrial carcinogenesis. Endocr Relat Cancer. 2018;25(12):981-991. doi: 10.1530/ERC-18-0106
416. Ortiz Hidalgo C. Immunohistochemistry in Historical Perspective: Knowing the Past to Understand the Present. Methods Mol Biol. 2022:2422:17-31. doi: 10.1007/978-1-0716-1948-3_2
417. Schacht V, Kern JS. Basics of immunohistochemistry. J Invest Dermatol. 2015;135(3):1-4. doi: 10.1038/jid.2014.541
418. De Matos LL, Trufelli DC, De Matos MGL, da Silva Pinhal MA. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. Biomark Insights. 2010;5:9-20. doi: 10.4137/bmi.s2185
419. Magaki S, Hojat SA, Wei B, So A, Yong WH. An introduction to the performance of immunohistochemistry. Methods Mol Biol. 2019;1897:289-298. doi: 10.1007/978-1-4939-8935-5_25
420. Combs SE, Han G, Mani N, Beruti S, Nerenberg M, Rimm DL. Loss of antigenicity with tissue age in breast cancer. Lab Invest. 2016;96(3):264-269. doi: 10.1038/labinvest.2015.138
421. Taylor CR, Rudbeck L. Immunohistochemical Staining Methods. 6th ed. Agilent; Dako; 2021.
422. Cree IA, Tan PH, Travis WD, et al. Counting mitoses: SI(ze) matters! Mod Pathol. 2021;34(9):1651-1657. doi: 10.1038/s41379-021-00825-7.
423. International Ki67 in Breast Cancer Working Group. International Ki67 in Breast Cancer Working Group. Accessed: February 13, 2024. https://www.ki67inbreastcancerwg.org/

424. Travaglino A, Raffone A, Saccone G, et al. PTEN immunohistochemistry in endometrial hyperplasia: which are the optimal criteria for the diagnosis of precancer? APMIS. 2019;127(4):161-169. doi: 10.1111/apm.12938.
425. Steinbakk A, Skaland I, Gudlaugsson E, et al. The prognostic value of molecular biomarkers in tissue removed by curettage from FIGO stage 1 and 2 endometrioid type endometrial cancer. Am J Obstet Gynecol. 2009;200(1):78.e1-8. doi: 10.1016/j.ajog.2008.07.020
426. Visiopharm. About us. Accessed: February 14, 2024. https://visiopharm.com/about-us/
427. Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. J Intern Med. 2020;288(1):62-81. doi: 10.1111/joim.13030
428. Usman HA, Abidin FAZ. Digital image analysis of immunohistochemistry KI-67 using QuPath software in breast cancer. Jurnal Kedokteran dan Kesehatan Indonesia. 2021;12(1):34-43.
429. Acs B, Pelekanou V, Bai Y, et al. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. Lab Invest. 2019;99(1):107-117. doi: 10.1038/s41374-018-0123-7
430. Boyaci C, Sun W, Robertson S, Acs B, Hartman J. Independent clinical validation of the automated Ki67 scoring guideline from the International Ki67 in Breast Cancer Working Group. Biomolecules. 2021;11(11):1612. doi: 10.3390/biom11111612.
431. Robertson S, Azizpour H, Smith K, Hartman JJTR. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. 2018;194:19-35. doi: 10.1016/j.trsl.2017.10.010
432. Dawe M, Shi W, Liu TY, et al. Reliability and Variability of Ki-67 Digital Image Analysis Methods for Clinical Diagnostics in Breast Cancer. Lab Invest. 2024:100341. doi: 10.1016/j.labinv.2024.100341
433. Abeler VM, Røyne O, Thoresen S, Danielsen HE, Nesland JM, Kristensen GB. Uterine sarcomas in Norway. A histopathological and prognostic survey of a total population from 1970 to 2000 including 419 patients. Histopathology. 2009;54(3):355-364. doi: 10.1111/j.1365-2559.2009.03231.x
434. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. doi: 10.1016/j.jcm.2016.02.012.
435. Usubutun A, Mutter GL, Saglam A, et al. Reproducibility of endometrial intraepithelial neoplasia diagnosis is good, but influenced by the diagnostic

style of pathologists. Mod Pathol. 2012;25(6):877-884. doi: 10.1038/modpathol.2011.22

436.428. Mutter GL, Lin M-C, Fitzgerald JT, et al. Altered PTEN expression as a diagnostic marker for the earliest endometrial precancers. J Natl Cancer Inst. 2000;92(11):924-930. doi: 10.1093/jnci/92.11.924

437. Kanamori Y, Kigawa J, Itamochi H, et al. PTEN expression is associated with prognosis for patients with advanced endometrial carcinoma undergoing postoperative chemotherapy. Int J Cancer. 2002;100(6):686-689. doi: 10.1002/ijc.10542

438. Uegaki K, Kanamori Y, Kigawa J, et al. PTEN-positive and phosphorylated-Akt-negative expression is a predictor of survival for patients with advanced endometrial carcinoma. Oncol Rep. 2005;14(2):389-392.

439. Mackay HJ, Gallinger S, Tsao MS, et al. Prognostic value of microsatellite instability (MSI) and PTEN expression in women with endometrial cancer: results from studies of the NCIC Clinical Trials Group (NCIC CTG). Eur J Cancer. 2010;46(8):1365-1373. doi: 10.1016/j.ejca.2010.02.031

440. Akiyama-Abe A, Minaguchi T, Nakamura Y, et al. Loss of PTEN expression is an independent predictor of favourable survival in endometrial carcinomas. Br J Cancer. 2013;109(6):1703-1710. doi: 10.1038/bjc.2013.455

441. Ørbo A, Arnes M, Hancke C, Vereide AB, Pettersen I, Larsen K. Treatment results of endometrial hyperplasia after prospective D-score classification: a follow-up study comparing effect of LNG-IUD and oral progestins versus observation only. Gynecol Oncol. 2008;111(1):68-73. doi: 10.1016/j.ygyno.2008.06.014

442. Sanderson PA, Esnal-Zufiaurre A, Arends MJ, et al. Improving the diagnosis of endometrial hyperplasia using computerized analysis and immunohistochemical biomarkers. Front Reprod Health. 2022;4:896170 doi: 10.3389/frph.2022.896170

443. Robertson S, Acs B, Lippert M, Hartman J. Prognostic potential of automated Ki67 evaluation in breast cancer: different hot spot definitions versus true global score. Breast Cancer Res Treat. 2020;183(1):161-175. doi: 10.1007/s10549-020-05752-w

444. Egeland NG. Discovery and Validation of Biomarkers in Breast Cancer. PhD Dissertation. University of Stavanger; 2020. Accessed February 14, 2024.

445. Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer. 2021;21(3):199-211. doi: 10.1038/s41568-020-00327-9

446. International Medical Device Regulators Forum. Software as a Medical Device (SaMD): Key Definitions. 2013. December 9, 2013. Accessed February 14, 2024. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf
447. Niemiec E. Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices? Digit Health. 2022;8:20552076221089079. doi: 10.1177/20552076221089079
448. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27(4):582-584.
449. U.S. Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. January 2021. Accessed February 14, 2024. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
450. Woerl A-C, Eckstein M, Geiger J, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. Eur Urol. 2020;78(2):256-264. doi: 10.1016/j.eururo.2020.04.023
451. Flinner N, Gretser S, Quaas A, et al. Deep learning based on hematoxylin–eosin staining outperforms immunohistochemistry in predicting molecular subtypes of gastric adenocarcinoma. J Pathol. 2022;257(2):218-226. doi: 10.1002/path.5879
452. Liu H, Xu W-D, Shang Z-H, et al. Breast cancer molecular subtype prediction on pathological images with discriminative patch selection and multi-instance learning. Front Oncol. 2022;12:858453. doi: 10.3389/fonc.2022.858453
453. Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. J Med Internet Res. 2022;24(1):e32939. doi: 10.2196/32939.
454. Paige. Paige and Visiopharm Work Together to Deliver Advanced AI Cancer Diagnostic Support Through the Paige Platform. Paige; May 4, 2023. Accessed February 14, 2024. https://paige.ai/paige-and-visiopharm-work-together-to-deliver-advanced-ai-cancer-diagnostic-support-through-the-paige-platform/

455. Paige. Paige and Mindpeak Launch an Integrated Solution for Enhanced Cancer Diagnosis. Paige; March 9, 2023. Accessed July 4, 2023. https://paige.ai/paige-and-mindpeak-launch-an-integrated-solution-for-enhanced-cancer-diagnosis/
456. Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. Histopathology. 2018;72(4):662-671. doi: 10.1111/his.13403
457. Williams BJ, Knowles C, Treanor D. Maintaining quality diagnosis with digital pathology: a practical guide to ISO 15189 accreditation. J Clin Pathol. 2019;72(10):663-668. doi: 10.1136/jclinpath-2019-205944
458. Boeken T, Feydy J, Lecler A, et al. Artificial intelligence in diagnostic and interventional radiology: Where are we now? Diagn Interv Imaging. 2022;104(1):1-5. doi: 10.1016/j.diii.2022.11.004
459. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA. 2016;316(22):2353-2354. doi: 10.1001/jama.2016.17438
460. Mukhjeree S. The Gene: An Intimate History. Vintage; 2016.

Appendix 2 – Mitotic Count

Table: Mitotic Count (score) according to diameter of the high-power field and its corresponding area, as described by the WHO Classification of Breast Tumours (2019)

Field diameter (mm)	Field area (mm^2)	Mitotic count (score)		
		1	2	3
0.40	0.126	≤ 4	5 – 9	≥ 10
0.41	0.132	≤ 4	5 – 9	≥ 10
0.42	0.138	≤ 5	6 – 10	≥ 11
0.43	0.145	≤ 5	6 – 10	≥ 11
0.44	0.152	≤ 5	6 – 11	≥ 12
0.45	0.159	≤ 5	6 – 11	≥ 12
0.46	0.166	≤ 6	7 – 12	≥ 13
0.47	0.173	≤ 6	7 – 12	≥ 13
0.48	0.181	≤ 6	7 – 13	≥ 14
0.49	0.188	≤ 6	7 – 13	≥ 14
0.50	0.196	≤ 7	8 – 14	≥ 15
0.51	0.204	≤ 7	8 – 14	≥ 15
0.52	0.212	≤ 7	8 – 15	≥ 16
0.53	0.221	≤ 8	9 – 16	≥ 17
0.54	0.229	≤ 8	9 – 16	≥ 17
0.55	0.237	≤ 8	9 – 17	≥ 18
0.56	0.246	≤ 8	9 – 17	≥ 18
0.57	0.255	≤ 9	10 – 18	≥ 19
0.58	0.264	≤ 9	10 – 19	≥ 20
0.59	0.273	≤ 9	11 – 20	≥ 20
0.60	0.283	≤ 10	11 – 21	≥ 21
0.61	0.292	≤ 10	12 – 22	≥ 22
0.62	0.302	≤ 11	12 – 22	≥ 23
0.63	0.312	≤ 11	12 – 23	≥ 23
0.64	0.322	≤ 11	12 – 23	≥ 24
0.65	0.332	≤ 12	13 – 24	≥ 25
0.66	0.342	≤ 12	13 – 24	≥ 25
0.67	0.352	≤ 12	13 – 25	≥ 26
0.68	0.363	≤ 13	14 – 26	≥ 27
0.69	0.374	≤ 13	14 – 27	≥ 28

Appendix 3 – QuPath Ki67 Cell Classifier Script

```
setImageType('BRIGHTFIELD_H_DAB');
setColorDeconvolutionStains('{"Name" : "H-DAB default", "Stain 1" : "Hematoxylin", "Values 1" : "0.65111 0.70119 0.29049 ", "Stain 2" : "DAB", "Values 2" : "0.26917 0.56824 0.77759 ", "Background" : " 255 255 255 "}');
runPlugin('qupath.imagej.detect.cells.PositiveCellDetection', '{"detectionImageBrightfield": "Optical density sum",  "requestedPixelSizeMicrons": 1.0,  "backgroundRadiusMicrons": 10.0,  "medianRadiusMicrons": 0.1,  "sigmaMicrons": 1.5,  "minAreaMicrons": 15.0,  "maxAreaMicrons": 400.0,  "threshold": 0.1,  "maxBackground": 2.0,  "watershedPostProcess": true,  "excludeDAB": false,  "cellExpansionMicrons": 5.0,  "includeNuclei": true,  "smoothBoundaries": true,  "makeMeasurements": true,  "thresholdCompartment": "Nucleus: DAB OD mean",  "thresholdPositive1": 0.2,  "thresholdPositive2": 0.4,  "thresholdPositive3": 0.6000000000000001,  "singleThreshold": true}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": false}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": true}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": false}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": false}');
```

PAPER I

Rewcastle, E., Varhaugvik, A.E., Gudlaugsson, E., Steinbakk, A., Skaland, I., van Diermen, B. Baak, J.P. & Janssen, E.A.M

(2018) Assessing the prognostic value of PAX2 and PTEN in endometrial carcinogenesis. *Endocrine-Related Cancer* 25(12):981-991. DOI: 10.1530/ERC-18-0106.

This paper is not available in the repository due to copyright restrictions.

PAPER II

Mod Pathol 36 (2023) 100116

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Automated Prognostic Assessment of Endometrial Hyperplasia for Progression Risk Evaluation Using Artificial Intelligence

Emma Rewcastle[a,b,*], Einar Gudlaugsson[a], Melinda Lillesand[a,b], Ivar Skaland[a], Jan P.A. Baak[a,c], Emiel A.M. Janssen[a,b]

[a] *Department of Pathology, Stavanger University Hospital, Stavanger, Norway;* [b] *Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway;* [c] *Dr. Med. Jan Baak AS, Tananger, Norway*

ARTICLE INFO

Article history:
Received 20 September 2022
Revised 20 December 2022
Accepted 18 January 2023
Available online 1 February 2023

Keywords:
gynecologic cancer
endometrial cancer
prognostic biomarkers
artificial intelligence

ABSTRACT

Endometrial hyperplasia is a precursor to endometrial cancer, characterized by excessive proliferation of glands that is distinguishable from normal endometrium. Current classifications define 2 types of EH, each with a different risk of progression to endometrial cancer. However, these schemes are based on visual assessments and, therefore, subjective, possibly leading to overtreatment or undertreatment. In this study, we developed an automated artificial intelligence tool (ENDOAPP) for the measurement of morphologic and cytologic features of endometrial tissue using the software Visiopharm. The ENDOAPP was used to extract features from whole-slide images of PAN-CK$^+$ −stained formalin-fixed paraffin-embedded tissue sections from 388 patients diagnosed with endometrial hyperplasia between 1980 and 2007. Follow-up data were available for all patients (mean = 140 months). The most prognostic features were identified by a logistic regression model and used to assign a low-risk or high-risk progression score. Performance of the ENDOAPP was assessed for the following variables: images from 2 different scanners (Hamamatsu XR and S60) and automated placement of a region of interest versus manual placement by an operator. Then, the performance of the application was compared with that of current classification schemes: WHO94, WHO20, and EIN, and the computerized-morphometric risk classification method: D-score. The most significant prognosticators were percentage stroma and the standard deviation of the lesser diameter of epithelial nuclei. The ENDOAPP had an acceptable discriminative power with an area under the curve of 0.765. Furthermore, strong to moderate agreement was observed between manual operators (intraclass correlation coefficient: 0.828) and scanners (intraclass correlation coefficient: 0.791). Comparison of the prognostic capability of each classification scheme revealed that the ENDOAPP had the highest accuracy of 88%-91% alongside the D-score method (91%). The other classification schemes had an accuracy between 83% and 87%. This study demonstrated the use of computer-aided prognosis to classify progression risk in EH for improved patient treatment.

© 2023 THE AUTHORS. Published by Elsevier Inc. on behalf of the United States & Canadian Academy of Pathology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

In 2020, more than 417,000 new cases of endometrial cancer were reported globally, with the highest incidence recorded in Asia followed by Europe, in Norway, 764 new cases were

* Corresponding author.
E-mail address: emma.rewcastle@sus.no (E. Rewcastle).

0893-3952/© 2023 THE AUTHORS. Published by Elsevier Inc. on behalf of the United States & Canadian Academy of Pathology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
https://doi.org/10.1016/j.modpat.2023.100116

reported.[1,2] Approximately 80% of endometrial cancers are preceded by endometrial hyperplasia.[3] According to the World Health Organization's 2020 guidelines (WHO20),[4] endometrial hyperplasia can be classified into 2 types: endometrial hyperplasia without atypia (HwA) and endometrial atypical hyperplasia (EAH). HwA is defined as the pathologic proliferation of endometrial glands without cytologic atypia, whereas EAH is defined as an excess of endometrial glands in comparison with stroma alongside changes to epithelial cytology that is distinct from surrounding normal endometrium.[4] Risk of progression to cancer varies between the 2 types: HwA is considered low risk (<5% progression risk)[4-6] and EAH high risk (8%-40% progression risk).[4,7] Additionally, EAH has a high risk of concurrent endometrial cancer.[4,6,8,9] Therefore, these diagnoses receive different treatments. High-risk EAH is commonly treated with surgery such as hysterectomy, whereas low-risk HwA is assigned conservative management with monitoring and progestin treatment. A weakness of the WHO20 scheme is the dependence on subjective qualitative criteria, resulting in poor reproducibility.

Another diagnostic classification scheme, which distinguishes high-risk endometrial intraepithelial neoplasia (EIN) from low-risk non-EIN, also experiences this shortcoming. A computerized, morphometric, semiautomatic assessment tool, D-score, is used to objectively evaluate the progression risk of patients with hyperplasia. This method measures key features of endometrial hyperplasia, including percentage stroma and variation in diameter of epithelial nuclei, to define progression risk: low or high. The method was both reproducible and prognostic in cohorts from the United States, Netherlands, and Norway.[10-12] However, it is time consuming and not widely available.

With the digital transformation of pathology departments across the globe come new opportunities for integration of artificial intelligence (AI) tools for computer-assisted prognostic assessment. Such applications can contribute to increased efficiency in routine workflows by automating time-consuming tasks, such as the D-score method, and have the potential to improve diagnostics through more objective and reproducible methods. In this study, we developed an AI-assistance tool that can classify endometrial hyperplasia into low-risk and high-risk forms to reduce overtreatment and undertreatment of patients.

Materials and Methods

Study Population

The study was retrospective and received approval from the Regional Ethics Committee of Health West Norway (2010/2464) and informed consent waived. The patient database consisted of 602 cases who received a diagnosis of endometrial hyperplasia between 1980 and 2007 at Stavanger University Hospital, Norway. Inclusion criteria for this study were as follows: (1) original diagnosis of endometrial hyperplasia by WHO94, (2) at least 1 follow-up sample, for nonprogression cases required to be ≥6 months after the initial diagnosis, (3) material available as a formalin-fixed paraffin-embedded (FFPE) tissue block, and (4) a minimum area of 4.0 mm^2 of endometrial tissue in the whole-slide image (WSI). Of the 467 cases that met inclusion criteria, 388 had D-scores available. Patients were treated according to the Norwegian National Guidelines at the time of primary diagnosis (total: n = 388; no treatment: n = 317; hormone/progesterone therapy: n = 71). Cases were previously diagnosed using the WHO94 guidelines.[3,13] Each case was reviewed by a pathologist (E.G.), on original hematoxylin-eosin and saffron (HES)–stained slides, according to the WHO20 guidelines[4] and EIN criteria.[13]

Immunohistochemistry (PAN-CK$^+$), Image Acquisition, and D-score

Tissue sections were cut from FFPE tissue blocks at a thickness of 3.0 μm. Sections were mounted on SuperFrost Plus slides (Menzel Gläser). After incubation at 60 °C for 1 hour, slides were transferred to the Dako Omnis. Slides were stained for pan cytokeratin plus (PAN-CK$^+$) using a preoptimized protocol. The PAN-CK$^+$ antibody (clone AE1/AE3+CK8/18; Biocare Medical) was used at a dilution of 1:150 and visualized with the EnVision FLEX detection system (Dako). Sections were counterstained with hematoxylin for detection of nuclei.

Whole PAN-CK$^+$–stained sections were scanned at 400× magnification using the Hamamatsu Nanozoomer XR (Hamamatsu Photonics) at Haukeland University Hospital, Bergen, Norway, and the Hamamatsu Nanozoomer S60 at Stavanger University Hospital.

D-scores were calculated on HES-stained sections as previously described.[14] Lesions scored as ≥0 were defined as low risk and lesions scored <0 as high risk.

Training and Tuning of an AI Application

Digitized WSIs were uploaded to the Visiopharm system (version 2020.08; Visiopharm A/S). An in-house application (ENDOAPP) was developed to detect specific morphologic features based on staining patterns using a combination of classifiers, filters, and postprocessing steps. The ENDOAPP consists of several individual algorithms/stages, which can be run separately or in batches (Fig. 1). The training data set (Fig. 2) was used to develop the ENDOAPP. Manual annotations were required for stages 1, 2, and 6. Annotations were added iteratively until the training was considered satisfactory (satisfactory alignment of feature and output label) as assessed by the operator (E.R.).

The algorithms were developed and trained as follows:

- Stage 1: Segmentation of images into tissue and nontissue using a threshold classifier trained on manually annotated labels of representative regions.
- Stage 2: Further segmentation of tissue into stroma, lumen, and epithelial (gland) compartments using a K-means classifier. Manually annotated labels (brown pixels: PAN-CK$^+$ [gland], blue pixels: hematoxylin [stroma], and white pixels: absence of stain [background and lumen]) were used to train the classifier. Segmentation of lumen compartments used relational and size criteria (eg, surrounded by gland label).
- Stage 3: A heat map was generated using a built-in package to detect dense regions of green label (gland compartment) for delineation of a 4.0-mm^2 region of interest (ROI) around the densest region, the hotspot.
- Stage 4: Extraction of measurements of the segmented compartments within the hotspot.
- Stage 5: Outlining of the gland compartment within the hotspot with a ROI.
- Stage 6: Segmentation of the gland compartment into nuclei and background using a K-means classifier. Manually annotated labels of brown pixels (background) and blue pixels (hematoxylin [nuclei]) were used to train the classifier.
- Stage 7: Extraction of measurements of the segmented nuclei.

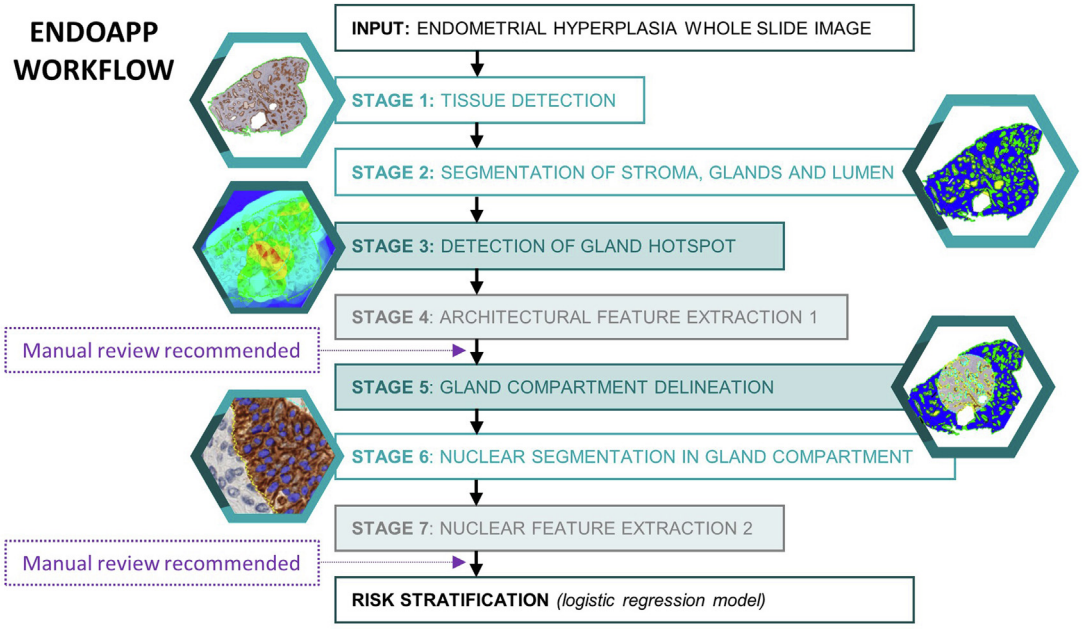


Figure 1.
The ENDOAPP workflow for progression risk assessment of endometrial hyperplasia. The workflow is divided into 7 stages, which run independently or in batch. The ENDOAPP can be paused, reviewed, or rerun at any stage. A manual review is performed after stages 4 and 7. Stage 1: Segmentation of tissue and nontissue, followed by delineation of tissue with a region of interest (ROI). Any tissue region smaller than 4.0-mm^2 total area was excluded. Stage 2: Segmentation of the stromal compartment (blue label), epithelial/gland compartment (green label), and lumen compartment (yellow label). Stage 3: A heat map is generated to detect the largest area of green label (gland). A single, circular, 4.0-mm^2 ROI is placed automatically around the hotspot. Stage 4: Measurements of all labeled features is performed within the hotspot ROI. Stage 5: The gland/epithelial compartment is delineated with a new ROI. Stage 6: The gland compartment is segmented into nuclei (blue label) and background (no label). Stage 7: Feature measurement and extraction for all segmented nuclei within the gland compartment. After completion of stage 7, all features extracted are input into the risk stratification model for calculation of progression risk score.

The features extracted in stage 4 and 7 are outlined in Table 1. The tuning set (Fig. 2) was used to monitor and evaluate ENDOAPP performance after adjustments to the algorithms.

Comparison of a Manual and Automatic ROI Method

To assess the prognostic value of the ENDOAPP, a manually placed ROI, by a trained operator, was compared with the automated ROI placement by the ENDOAPP. This was performed on Hamamatsu XR scanned images from the development data set (Fig. 2). The automatic method proceeded as described (Fig. 1). For the manual method, 3 independent operators (E.A.M.J., E.R., E.G.), blinded to the outcome and previously trained in the D-score analysis (E.G. is an experienced pathologist), were instructed to place a 4.0-mm^2 circular ROI on an unclassified WSI. The operators were instructed to use the following criteria: the ROI must include at least 2.0 mm^2 of endometrial tissue, fragmented regions (Supplementary Fig. S1) should be avoided where possible, and the ROI should contain the area of tissue the operator deemed to have the highest density of glands. Stages 2-7 of the ENDOAPP was run, excluding stage 3 (hotspot detection) (Fig. 1). To avoid retention bias for the operator (E.R.), a mandatory washout period of 6 weeks was required between reviewing any results/classified images and performing manual analysis.

For both manual and automated ROI placement, a classified image was reviewed (E.R.) after stages 4 and 7 and the case assigned either pass or fail (Fig. 1). If the analysis failed owing to tissue or scanning artifacts (Supplementary Fig. S2), a new ROI was placed by the original operator if a similarly dense region was available. If no similar region was available, the analysis was classified as a fail. For the automated method, artifacts, including benign mimics such as metaplasia, were excluded from the tissue ROI (stage 1) and the ENDOAPP rerun from stage 2. If no similarly available region was present, the ENDOAPP was not rerun and the case classified as a fail. Performance was recorded for both manual and automatic methods.

Statistical Analysis

All statistical analyses were performed using SPSS for Windows (version 26.0.0; IBM SPSS Statistics). For all analyses a P value of $<.05$ was considered significant and assumptions were verified for each statistical test. The Kruskall-Wallis and Pearson χ^2 tests were used to determine differences between the original database cohort ($N = 602$) and the study database ($n = 388$). The development data set was used for all statistical analyses (Fig. 2).

Feature Assessment

Preliminary feature assessment was performed on data from WSI from the Hamamatsu XR. Descriptive statistics and crosstables were generated for all extracted variables as defined in Table 1. Patients were grouped according to progression or no progression to endometrial carcinoma. Nonprogression was defined by no diagnosis of endometrial endometrioid carcinoma in the follow-up period and progression as a diagnosis of endometrial endometrioid carcinoma in the follow-up period.

Risk Classification Modeling

A logistic regression analysis was used to identify any meaningful relationship between the outcome (progression status) and

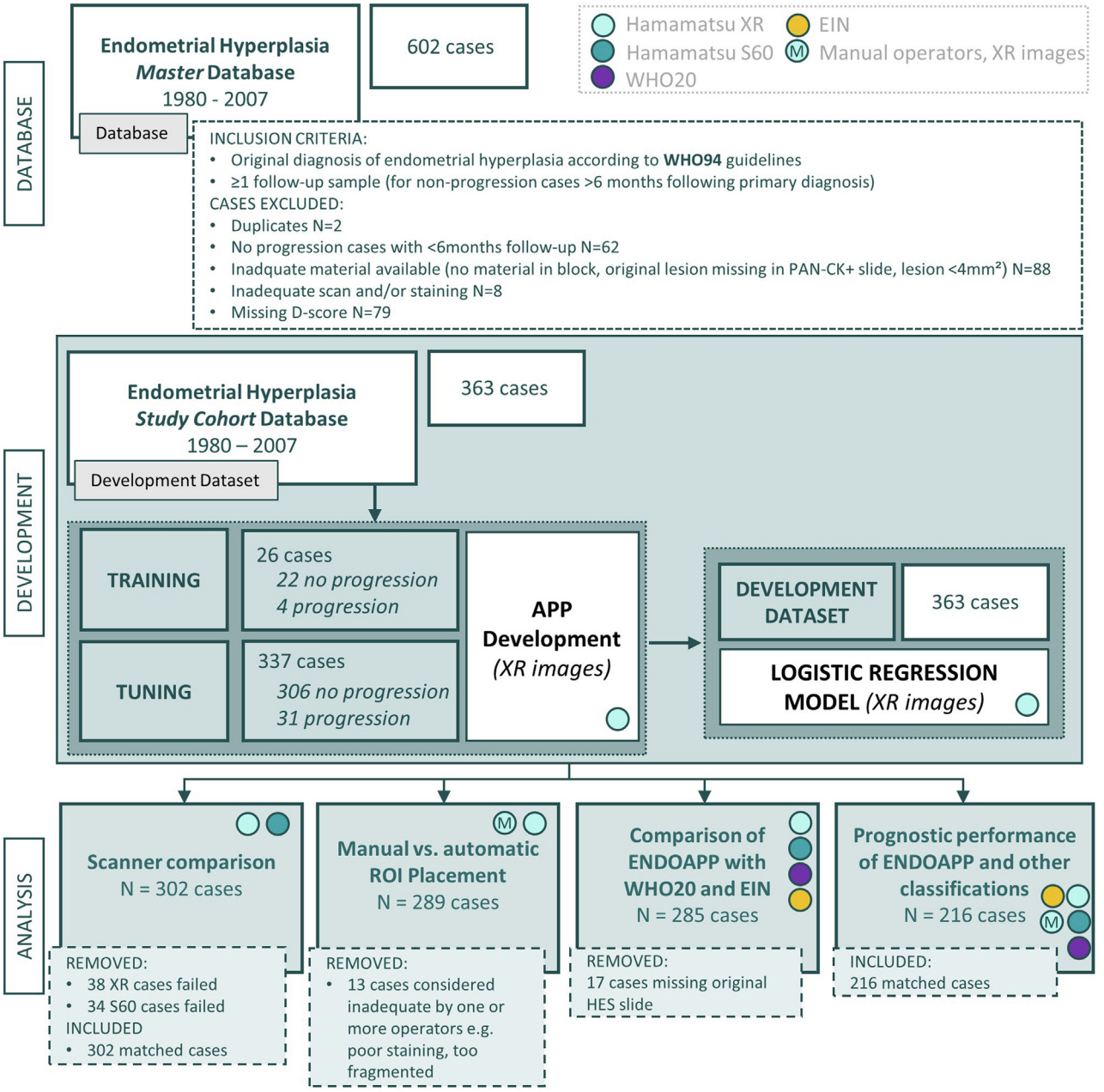


Figure 2.
Overview of cases from the original database of 602 patients with a primary diagnosis of endometrial hyperplasia between the years 1980 and 2007 diagnosed at Stavanger University Hospital, Stavanger, Norway. Inclusion and exclusion criteria for the development data set are described. Training and tuning data sets were generated from the 363 cases available from the development data set and prevalence of progression was aimed to be between 0.1 and 0.25 in each group. The ENDOAPP was trained on annotations from the training data set. The tuning data set was used to evaluate the training and identify any adjustments required. The full development data set was used to develop the logistic regression model for risk stratification. Further comparative analysis was performed to assess performance and prognostic capability across classifications. Cases that were included or excluded are described.

the predictor variables. The predictor variables tested were age; WHO20, WHO94, and EIN classifications; and the features described in Table 1. The best performing combination of features recommended using the backward Wald method was used in the final logistic regression model. The feature standard deviation of the lesser diameter of the nuclei (LDSD) was transformed by a factor of 10 to avoid values between 0 and 1.

Survival Analysis

A receiver operating characteristic (ROC) curve was created to identify potential risk category cut-offs. The area under the ROC curve (AUC) was used to evaluate the model's discriminative ability according to the guidelines by Hosmer et al.[15] Kaplan-Meier curves were generated using the most optimal cutoff. Significant differences were assessed using the log-rank test. The end point was progression-free survival, with progression to endometrial cancer defining an event. Patients with no progression were censored according to the last known follow-up date or if they received a hysterectomy, censored to the date of the procedure. Kaplan-Meier survival curves were generated using R studio (2022.02.0 Build 443).

Table 1
Overview of the features measured and extracted by the application

Feature	Units	Type of measurement	Type of measure
Structural features			
Percentage stroma	%	Area	Total
Percentage gland	%	Area	Total
Percentage lumen	%	Area	Total
Ratio of *glands+lumen:stroma*		Area	Total
Perimeter length of glands	mm	Linear	Total; mean
Interface length of glands	mm	Linear	Total; mean
Gland count		Count	Total
Ellipticalness[a] (gland)	0-1	Score	Mean
Form factor[b] (gland)	0-1	Score	Mean
Nuclear features			
Lesser diameter	μm	Linear	Mean ± SD
Minor axis	μm	Linear	Mean ± SD
Largest diameter	μm	Linear	Mean
Major axis	μm	Linear	Mean
Ellipticalness (nuclei)	0-1	Score	Mean
Form factor (nuclei)	0-1	Score	Mean

[a] A measure of how elliptical an object is.
[b] A measure of how round an object is.

Performance Metrics

For each classification scheme (ENDOAPP, WHO94, WHO20, EIN, and D-score), performance metrics were calculated: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Accuracy was calculated based on a prevalence of progression of 0.25, as estimated by the Norwegian Department of Health,[16] and 0.1, as estimated from the study cohort.

Concordance Analysis

To assess reproducibility of the model, Cohen κ (categorical variables) and the intraclass correlation coefficient (ICC) were calculated (ICC: absolute agreement, 2-way mixed effects model with 95% confidence intervals). Agreement levels, as defined by Koo and Li,[17] was assessed between XR and S60, operators of the manual method, manual and automatic ROI placement, and the model and D-scores (κ only).

Results

Case Overview

From the original database of 602 endometrial hyperplasia cases, 388 cases met the inclusion criteria and had D-scores available. Of the 388 cases, 25 were removed owing to the following reasons: 2 duplicates, 3 cases failed scanning (Hamamatsu XR only), 15 cases had no tissue remaining in the FFPE block, and 5 cases had inadequate staining, leaving 363 cases for analysis, of which 328 had no progression to endometrial cancer (90%) and 35 cases had progression (10%). Between the original 602 cases and final 388 cases, there was little to no significant difference in the proportion of cases according to diagnostic category, progression status, or age (Table 2). Furthermore, there was no significant survival difference between treated and untreated patients according to diagnostic category: WHO94, WHO20, and EIN (Kaplan-Meier, $P > .1$).

ENDOAPP Performance

The ENDOAPP ran successfully on 325 WSIs (n = 363, 90%) scanned on the Hamamatsu XR (XR) and for 332 WSI (n = 366, 91%) scanned on the Hamamatsu S60 (S60). For both WSIs scanned on the XR and S60, 8% (n = 49) of cases failed because of either out-of-focus regions resulting in failure of stage 6 (Fig. 1) or lesions being smaller than 2.0 mm^2 (n = 21, 2%). WSIs scanned using the XR had a higher edit rate after review than those by the S60 (44% vs 33%). The main types of edits included removal of artifacts such as benign mimics such as metaplasia (2 cases) and fragmented regions and correcting erroneous labels due to suboptimal staining intensity caused by too thick sections. Edits were performed after a manual review (Fig. 1). Thus, 325 cases were available for further analysis for the XR and 332 cases for the S60. For comparative analysis between the 2 scanners, 302 cases were available (Fig. 2).

Prognostic Feature Extraction

The ENDOAPP measured and extracted 15 different features (Table 1). A logistic regression analysis was performed to identify any associations between the features and progression, with age included. Five features were identified as significant predictors of progression, independently (Table 3). These features were further analyzed in a multivariate logistic regression. Percentage stroma and LDSD were recommended as optimal predictor variables of progression ($P < .001$ and $P = .005$, respectively). This model had acceptable discriminative power, with an AUC of 0.765 (95% CI, 0.655-0.875) in an ROC curve analysis. The most optimal cutoff was used, and cases were defined as low risk (≤0) or high risk (>0), with a score range of −5 to 3.

Survival Analysis

All included cases had follow-up data available, ranging from 10 to 366 months (mean = 144 ± 75.8 months) for no progression

Table 2
Overview of cases included in the original patient database and the final study database

Characteristics	Original database	Study database	Statistical difference
N	602	388	
Age (y)			H = 3.600, *P* =.058
Mean	53	51	
Range	21-98	24-88	
Progression (% cases)			$\chi^2 = 0.694$, *P* =.707
No progression	86	89	
Cancer	6	8	
Concurrent cancer	3	3	
Lost to follow-up	5	0	
Original diagnosis (WHO94) (% cases)			$\chi^2 = 5.745$, *P* =.332
Simple hyperplasia	64	67	
Simple hyperplasia with atypia	4	3	
Complex hyperplasia	21	21	
Complex hyperplasia with atypia	9	7	
Unclear hyperplasia	2	2	

cases and 0 to 247 months (mean = 70 ± 68.4 months) for progression cases. The Kaplan-Meier survival analysis revealed that high-risk patients (as defined by the ENDOAPP) showed significantly lower progression-free survival than low-risk patients did over a period of 20 years (Fig. 3A). The ENDOAPP, using XR-scanned images, had a sensitivity of 50% and specificity of 92% (n = 325).

Comparison of Scanners

In addition to XR-scanned images, the ENDOAPP was run on WSIs obtained from the S60 scanner. Similar to results from the XR, S60 classifications revealed a significant separation in progression-free survival between low-risk and high-risk patients (Fig. 3). In addition, there was moderate absolute agreement between scanners (ICC: 0.770; 95% CI, 0.282-0.899). The agreement of the feature percentage stroma between the scanners was strong (ICC: 0.890; 95% CI, 0.864-0.912; $P < .001$), whereas the agreement of LDSD was weak (ICC: 0.536; 95% CI, −0.081 to −0.801). Sensitivity and specificity of the model for the S60 were similar to those of the XR (42% and 50% and 96% and 92%, respectively).

Comparison of Manual and Automatic ROI Placement

In addition to automated ROI placement, 6 operators performed manual ROI placement and evaluated the variability and performance of the ENDOAPP. There was strong agreement (risk scores) between the 6 operators for manual placement (ICC: 0.828; 95% CI, 0.789-0.861; $P < .001$) and moderate agreement between the operators and the automated method (ICC: 0.791; 95% CI, 0.752-0.827; $P < .001$). The ROIs chosen by the operators tended to overestimate the risk score, with higher a number of false-positive results than the automated method. Following further investigation of falsely predicted cases, most (58%) had borderline scores (0 ± 0.5). The remainder were either highly fragmented or showed uniform morphology throughout the tissue. We observed little difference in morphology between HES-stained and PAN-CK$^+$−stained sections for most of the cases (97%). Comparison of the ENDOAPP with the D-score method revealed minimal (XR) (κ: 0.361; 95% CI, 0.194-0.528; $P < .001$) to moderate agreement (S60) (κ: 0.506; 95% CI, 0.328-0.684; $P < .001$).

Comparison of Classification Schemes

Risk classification by all methods (WHO94, WHO20, EIN, D-score, XR-ENDOAPP, and S60-ENDOAPP) was available for 285 matched cases (216 for additional manual operator comparison) (Fig. 2). Performance metrics were calculated for each method (Table 4). The number of false-negative results was similar for all classification methods (Table 4). There was notable variation in the number of false-positive results among the classification schemes, with fewest recorded by the ENDOAPP (S60) and D-score method and the highest recorded by the EIN scheme (Table 4). Overall, the highest accuracy was reported for the ENDOAPP (S60) and D-score method at 91% (Table 4). NPVs were similar across all classifications, whereas PPVs were highest for the ENDOAPP (S60) and D-score method at 52% and 54%, respectively (Table 4). When the metrics were adjusted to only include matched cases (n = 216), our observations remained relatively unchanged (Table 4, data in parentheses). Furthermore, when WHO20 and EIN were added to the logistic regression model, only WHO20 was recommended as an additional predictor; however, there was no

Table 3
Prognostic features for assessment of risk of progression of endometrial hyperplasia evaluated independently by a logistic regression model

Feature	Units	OR	95% CI	*P*
Age	<55, ≥55 y	2.902	1.304-6.456	.009
Percentage stroma	%	0.930	0.901-0.959	<.001
Percentage lumen	%	1.120	1.055-1.189	<.001
Ratio[a]		2.990	1.559-5.735	.001
SD of the lesser diameter of gland nuclei ×10	μm	1.687	1.170-2.431	.005

[a] The ratio of glands+lumen to stroma.

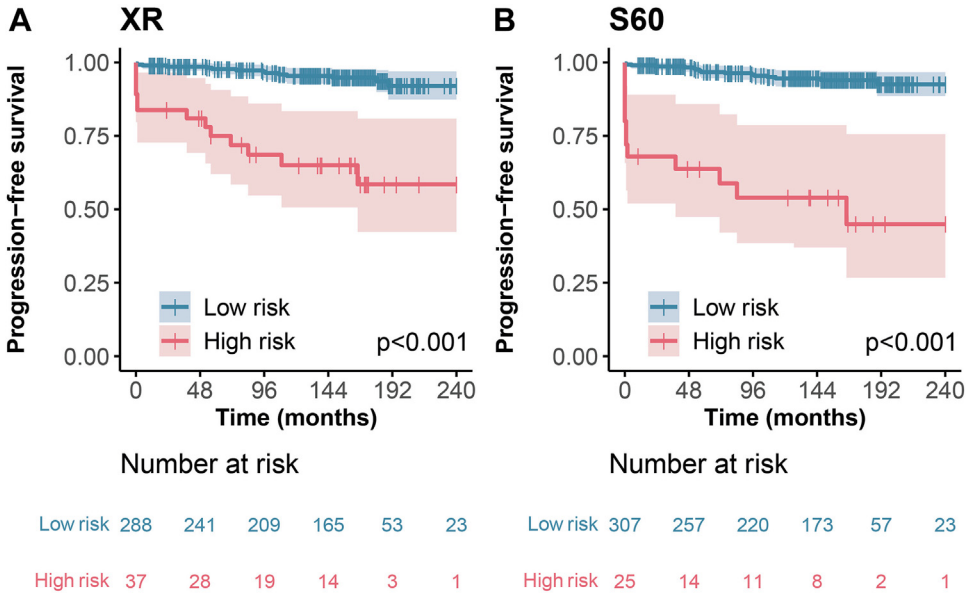


Figure 3.
Kaplan-Meier curves demonstrating progression-free survival in patients diagnosed with endometrial hyperplasia, stratified by risk score according to the ENDOAPP logistic regression model. (A) Progression-free survival estimates extracted from XR scanned images (n = 325). (B) Progression-free survival estimates extracted from S60 scanned images (n = 332).

noteworthy improvement to the resulting NPVs and PPVs, nor accuracy, compared with those by the ENDOAPP (data not shown). When progressive cases were subdivided into concurrent and nonconcurrent cases, the D-score method had the fewest misclassifications (n = 1) in comparison with WHO20 and EIN, which recorded the most (n = 3) (Table 5). In summary, the ENDOAPP (S60) and D-score method obtained higher performance metrics than the other classifications.

Discussion

AI has been used to identify and diagnose endometrial hyperplasia and carcinoma in histologic tissue,[18-22] but prognostic assessment has not yet been studied in depth. In this study, we use AI to classify the risk of progression of endometrial hyperplasia. The ENDOAPP is an improvement on traditional prognostic guidelines and has equal prognostic values to those of the established, quantitative D-score method in the study cohort.

The ENDOAPP performed successfully on 89%-91% of WSIs, where reasons for failure constituted preanalytical errors such as out-of-focus scans, too thick sections, and too small lesions. If these errors are corrected, fewer cases would fail at the analytical stage. In digital pathology optimization of preanalytical protocols should be prioritized, including steps for quality control of WSI, before implementation of AI algorithms.[23]

At present, 3 classification systems are used in the clinic to assess prognosis of endometrial hyperplasia: the WHO20, EIN, and D-score. A meta-analysis by Raffone et al[24] debates that there is not enough evidence to suggest that either WHO20 or EIN is

Table 4
Performance metrics for the ENDOAPP according to scanner type (XR, S60) and 3 operators for manual ROI placement, and performance metrics for D-score and the diagnostic classification schemes: WHO94, WHO20, and EIN

Performance metrics	ENDOAPP			D-score	WHO94	WHO20	EIN
	XR	S60	Operators				
N	325 (216)	332 (216)		363 (216)	378 (216)	349 (216)	348 (216)
True-positive results	14 (6)	13 (5)		13 (6)	18 (5)	16 (5)	18 (8)
True-negative results	274 (182)	289 (194)		317 (193)	317 (179)	285 (183)	272 (174)
False-positive results	23 (18)	12 (6)		11 (7)	25 (21)	30 (17)	42 (26)
False-negative results	14 (10)	18 (11)		22 (10)	22 (11)	18 (11)	16 (8)
Sensitivity (%)	50 (38)	42 (31)	50-58 (44-56)	37 (38)	39 (31)	47 (31)	53 (50)
Specificity (%)	92 (91)	96 (97)	95-96 (84-86)	97 (97)	93 (90)	90 (92)	87 (87)
NPV (%)	95 (95)	94 (95)	95-96 (95-96)	94 (95)	94 (94)	94 (94)	94 (96)
PPV (%)	38 (25)	52 (45)	24-26 (19-21)	54 (46)	36 (19)	35 (23)	30 (24)
Accuracy (0.25)[a] (%)	82 (78)	82 (81)	76-79 (75-77)	82 (82)	79 (75)	80 (76)	78 (78)
Accuracy (0.1)[b] (%)	88 (86)	91 (90)	82-83 (81)	91 (91)	87 (84)	86 (85)	83 (83)

Values are both unadjusted and adjusted (in parentheses) for matched cases based on data availability.
NPV, negative predictive value; PPV, positive predictive value.
[a] Accuracy based on a prevalence of progression of 0.25 estimated by the Norwegian Department of Health.[16]
[b] Accuracy based on a prevalence of progression of 0.1 estimated from the study cohort.

Table 5
Diagnosis of concurrent progression cases according to WHO94, WHO20, and EIN classification schemes and risk classification according to D-score, ENDOAPP (XR), and ENDOAPP (S60)

WHO94		WHO20		EIN		D-score		ENDOAPP (XR)		ENDOAPP (S60)	
No. of cases according to diagnostic category = 9											
CH	1	HwA	3	Non-EIN	3	Low	1	Low	2	Low	2
CAH	8	EAH	5	EIN	5	High	8	High	7	High	7
		Cancer	1	Cancer	1						

CAH, complex atypical hyperplasia; CH, complex hyperplasia.

superior to the other but conclude that the D-score method has the highest prognostic value.[24-30] In this study, the 2 most prognostic features in a multivariate logistic regression model were percentage stroma and LDSD. This is not unexpected because both features are also measured for the D-score analysis.[12] The final ENDOAPP incorporated both these features.

In comparison with canonical classifications, we assessed the prognostic value of the ENDOAPP. In the clinic, a test with a high PPV is desired when treatment is highly invasive. Therefore, it is prudent to develop a model that has a low likelihood of producing false-positive results: patients who would be likely to be overtreated. Ideally, with a progression prevalence of 10%, an optimal PPV is approximately 50%, with an NPV of 99%. The ENDOAPP achieved a PPV of 38%-52%, whereas the traditional schemes (EIN, WHO20, and WHO94) had a lower PPV of 30%-35% in the study cohort, also lower than the target. Ørbo et al[12] noted for the WHO94 scheme an NPV and PPV of 94% and 44%, respectively, which is in line with our observations for the WHO94 (NPV: 94%, PPV: 36%) and WHO20 (NPV: 94%, PPV: 35%). It is also comparable with the findings of Sanderson et al,[20] who reported an NPV of 91.6% for WHO94 and 98.4% for WHO14/EIN scheme. At present, we acknowledge that the evaluation of the ENDOAPP has been limited to a development data set with a small number of progression cases. Therefore, we cannot make any definitive conclusions on clinical implementation yet. To assess clinical relevance, validation of the ENDOAPP should be performed on an independent test set with a sufficient number of representative samples (eg, benign mimics and progressive cases).

A reference or gold standard is required when evaluating a new test or biomarker for its predictive and prognostic values. In our case, the gold standard could be D-scores or one of the other classification schemes. We demonstrated weak to minimal agreement between the ENDOAPP and D-score method. However, as noted by Pepe et al,[31] assessing the agreement between 2 tests does not consider the target of the test: in this case, presence or absence of progression. Therefore, we find it more appropriate to use survival data as our "ground truth" and compare the performance metrics of each test. We report similar PPVs for the ENDOAPP (38%-52%) and D-score (54%), which is comparable with the literature where the PPV for the D-score method varies between 38% and 67%,[12,13,32,33] with an average of 50.8% reported.[14] Although the NPV reported in this study is a little lower (ENDOAPP: 94%-95%, D-score: 94%) than the average 100% reported by Baak et al.[14] The similarities between the D-score method and ENDOAPP suggest that our method thus far is equivalent to the D-score method.

In addition to assessing the prognostic value of our method, we wish to assess the reproducibility of the ENDOAPP in comparison with other classification schemes. Since the transition from a multitiered classification (WHO94/WHO03) to a 2-tiered classification (WHO14/WHO20 and EIN), the consensus is that reproducibility of endometrial hyperplasia classifications has improved.[25,29,34] In this study, we report good reproducibility with a moderate agreement between different operators and scanners using a 2-tiered system (ICC: 0.791). Usubutun et al[35] evaluated the reproducibility of a 3-class EIN scheme (benign, EIN, and carcinoma) and reported an average weighted κ value of 0.72, while Ordi et al[29] reported a full agreement among pathologists at 70% for WHO03[36] and 69% for EIN. The reproducibility observed in the literature is similar to what we observe between the scanners (ICC: 0.770). Moreover, of the 2 features suggested by the logistic regression model, LDSD, a measure of cytologic atypia, had a weaker agreement (ICC: 0.536) in comparison with percentage stroma (ICC: 0.890). This is consistent with observations that cytological atypia is the diagnostic criteria with the highest disagreement between pathologists.[13,37–40]

We acknowledge that there are limitations with the ENDOAPP. Although we argue that automated methods are objective, we were unable to remove subjective influence completely. Owing to the nature of sampling techniques, endometrial tissue is often subject to mechanical stress, resulting in either a fragmented appearance or artifact crowding due to compression of glands, which makes it difficult to obtain a whole, representative sample. Fragmented regions often lose the stromal compartment, leaving disjointed and incomplete epithelium that appears as dense clusters. This can result in a false-positive score, and such artifacts should be removed from the analysis. The ENDOAPP is able to remove only smaller fragmented regions (<4.0 mm^2). Therefore, an operator must identify and remove them during the review process. This requires knowledge of endometrial tissue for correct identification. We observed that fragmented regions were still approved for analysis because it was sometimes difficult to differentiate between sampling artifacts and true tissue. Fragment artifacts represented 18% of misclassified cases. A more advanced AI approach such as deep learning may be more suitable to perform this task, although fragmented lesions have also been noted as a source of error in a deep learning model.[22] Nevertheless, scores achieved through automated placement and manual placement showed strong agreement, and therefore, any variation introduced owing to manual editing may have little impact. However, operators had a tendency to overestimate, with a higher number of false-positive results than the automated method. In difficult cases, manual placement may be acceptable. As discussed, a deep learning neural network may be a favorable approach because it would allow utilization of hematoxylin and eosin–stained slides over PAN-CK^+–stained slides. This would remove the extra step of immunohistochemistry, saving on cost and time, and may be more relevant to pathologists who primarily diagnose on hematoxylin and eosin–stained slides.

Regarding clinical implementation, the ENDOAPP must be validated on an independent test set before any conclusions of realistic clinical implementation can be made. Practical aspects should also be considered, such as integration of the ENDOAPP into a pathology workflow and feasibility and cost of software integrations into the existing laboratory systems. Implementation of an AI algorithm will be dependent on the cost-benefits judged

by the demands of the individual institution. In laboratories where the D-score method is used, implementation of the ENDOAPP may be very relevant because there is a considerable time-saving aspect: shortening analysis time from 15 to 60 minutes per case to 2 to 15 minutes. The tradeoff between the logistics of integration, time spent on the analysis, and the potential prognostic effect compared with the shortcomings of present-day diagnostics should be weighed carefully.

Pathologists use diagnostic guidelines such as the WHO20 to diagnose endometrial hyperplasia. However, owing to interobserver variability, a "panel approach" is recommended because morphologic criteria alone may not be sufficient for accurate prognosis.[20,41,42] We propose that AI tools should also be considered for such panels, either for objective and quantitative evaluation of morphologic features or for assessment of prognostic biomarkers. The biomarkers PAX2 and PTEN have prognostic values and may be potential candidates for such a panel.[43,44] Endometrial tissue can be notoriously difficult to assess owing to the common presence of sampling artifacts, the influence of hormones, and the presence of benign mimics such as polyps and metaplasia. Therefore, a combinatorial approach using computerized analysis, molecular analysis and/or immunohistochemistry, which are less affected by the mentioned pitfalls, may be a more standardized and robust method than a single classification system.

In brief, use of the ENDOAPP intends to assist a pathologist in the prognostic assessment of endometrial hyperplasia. The ENDOAPP may improve the reproducibility of prognostic decisions and make them more objective through automation. At present, the ENDOAPP has been evaluated on a development data set, and it needs to be validated on an independent test set, where performance, value of clinical implementation, and practical usability should be assessed.

To conclude, the ENDOAPP presented in this study demonstrates equal prognostic value and improved practicality to the D-score method. Moreover, it has the potential to improve prognostic evaluation of endometrial hyperplasia for improved patient treatment.

Acknowledgments

The authors thank the technical assistance from Marit Nordhus and Silja Kavlie Fykse and the Units for Immunohistochemistry and Histology at Stavanger University Hospital for their technical support.

Author Contributions

E.R. and E.A.M.J. conceptualized the study. E.R., I.S., and E.A.M.J. contributed to the development of the methodology. E.R. developed the application. E.R., E.G., M.L., and E.A.M.J. provided data acquisition. E.R. performed statistical analysis. E.R., I.S., E.G., J.B., and E.A.M.J. interpreted data. E.G. and J.B. provided medical insight. E.R. wrote the manuscript, and all authors read and approved the final paper.

Data Availability

The data sets generated during the current study are not publicly available because of ethical and legal concerns. Anonymized data can be made available from the Stavanger University Hospital Institutional Data Access/Ethics Committee (contact via email: rek-vest@uib.no, REK vest, Rogaland, Vestland, Norway) for researchers who meet the criteria for access to confidential data.

Funding

The study was funded by the Helse Vest Strategic Research Fund as part of the Pathology in Western Norway project. The medical practice Dr Med Jan Baak AS did not provide any financial support and did not have any additional role in study design, data collection and analysis, and decision to publish or preparation of the manuscript.

Declaration of Competing Interest

The authors declare there are no conflicts of interest.

Ethics Approval and Consent to Participate

The study was retrospective and received approval from the Regional Ethics Committee of Health West Norway (2010/2464) and informed consent waived. The study was performed in accordance with the Declaration of Helsinki.

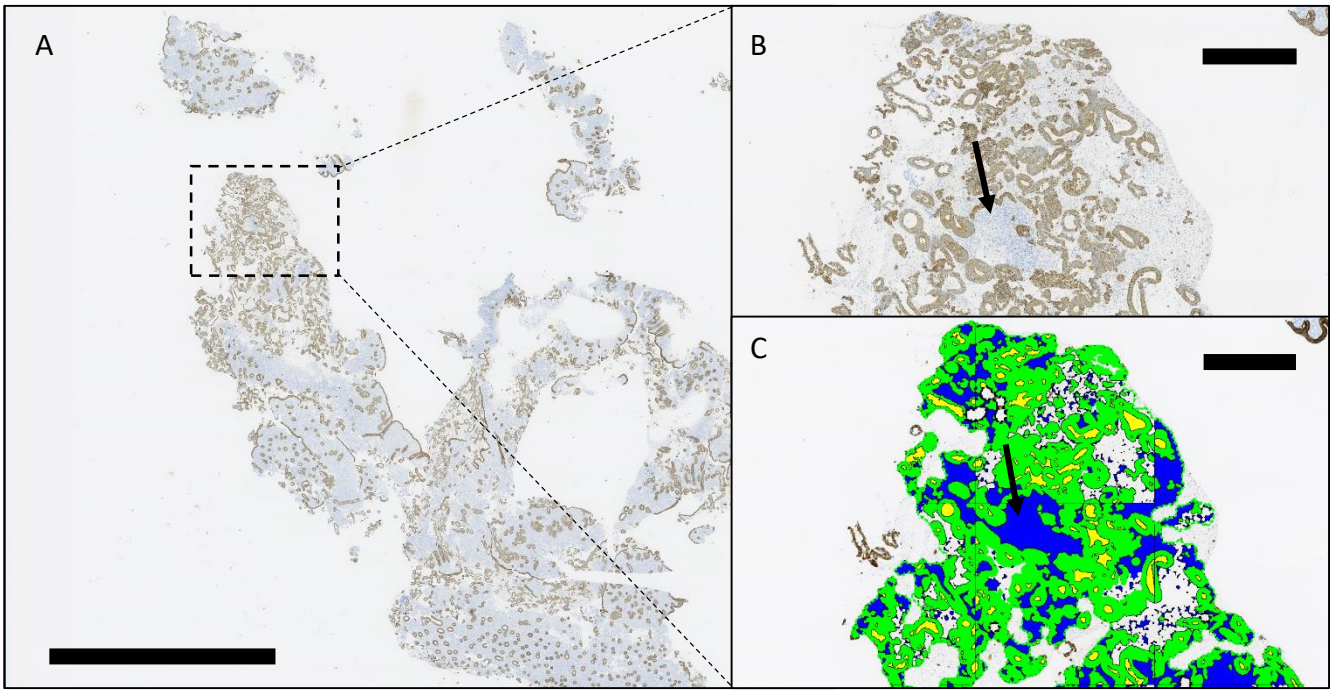
Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.modpat.2023.100116

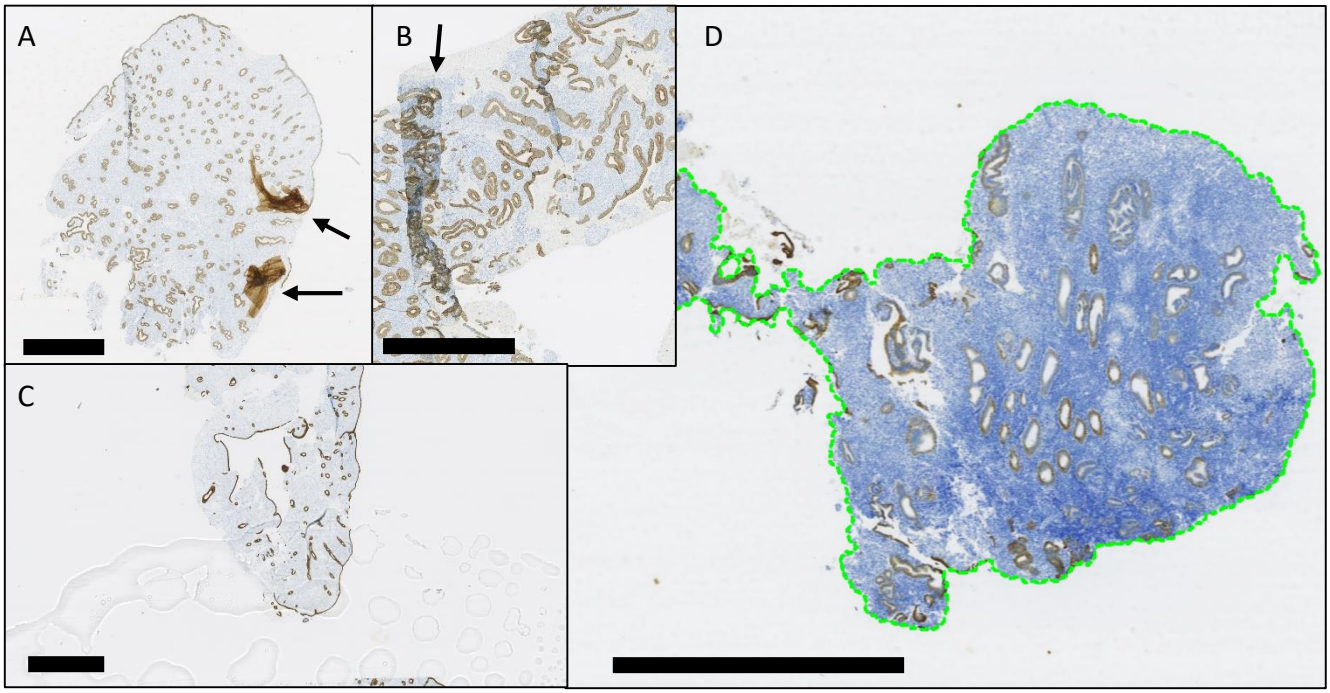
References

1. Cancer Registry of Norway. *Cancer in Norway 2020—Cancer Incidence, Mortality, Survival and Prevalence in Norway*. Cancer Registry of Norway; 2021.
2. Global Cancer Observatory. *Estimated Number of Cases Worldwide, Females, All Ages (excl. NMSC). [Online Database]*. International Agency for Research on Cancer, World Health Organization; 2020. Global Cancer Observatory (iarc.fr) Accessed April 25, 2022.
3. Kurman RJ, Kaminski PF, Norris HJ. The behavior of endometrial hyperplasia. A long-term study of "untreated" hyperplasia in 170 patients. *Cancer*. 1985;56:403–412.
4. WHO Classification of Tumours Editorial Board. *Female Genital Tumours; vol 4. 5th ed*. International Agency for Research on Cancer; 2020.
5. Nees LK, Heublein S, Steinmacher S, et al. Endometrial hyperplasia as a risk factor of endometrial cancer. *Arch Gynecol Obstet*. 2022;306:407–421.
6. Trimble CL, Kauderer J, Zaino R, et al. Concurrent endometrial carcinoma in women with a biopsy diagnosis of atypical endometrial hyperplasia. *Cancer*. 2006;106:812–819.
7. Urban RR, Reed SD. *Endometrial hyperplasia: management and prognosis*. In: *UpToDate*; 2021. UpToDate.
8. Chen YL, Cheng WF, Lin MC, Huang CY, Hsieh CY, Chen CA. Concurrent endometrial carcinoma in patients with a curettage diagnosis of endometrial hyperplasia. *J Fromos Med Assoc*. 2009;108:502–507.
9. Doherty MT, Sanni OB, Coleman HG, et al. Concurrent and future risk of endometrial cancer in women with endometrial hyperplasia: a systematic review and meta-analysis. *PLoS One*. 2020;15:e0232231.
10. Dunton CJ, Baak JP, Palazzo JP, van Diest PJ, McHugh M, Widra EA. Use of computerized morphometric analyses of endometrial hyperplasias in the prediction of coexistent cancer. *Am J Obstet Gynecol*. 1996;174:1518–1521.
11. Baak J, Wisse-Brekelmans E, Fleege J, Van Der Putten H, Bezemer P. Assessment of the risk on endometrial cancer in hyperplasia, by means of morphological and morphometrical features. *Pathol Res Pract*. 1992;188:856–859.
12. Ørbo A, Baak JP, Kleivan I, et al. Computerised morphometrical analysis in endometrial hyperplasia for the prediction of cancer development. A long term retrospective study from northern Norway. *J Clin Pathol*. 2000;53:697–703.
13. Baak J, Mutter G. EIN and WHO94. *J Clin Pathol*. 2005;58:1–6.

14. Baak JPA, Ørbo A, Van Diest PJ, et al. Prospective multicenter evaluation of the mophometric D-score for prediction of the outcome of endometrial hyperplasias. *Am J Surg Pathol*. 2001;25:930–935.
15. Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression; vol 398*. John Wiley & Sons; 2013.
16. Norwegian Directorate of Health. *Nasjonalt handlingsprogram med retningslinjer for gynekologisk kreft [National Guidelines for Gynecological Cancer]*. Norwegian Health Directorate; 2021. Accessed June 2021. Nasjonalt Handlingsprogram med retningslinjer for gynekologisk kreft - Helsedirektoratet.
17. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
18. Downing MJ, Papke Jr DJ, Tyekucheva S, Mutter GL. A new classification of benign, premalignant, and malignant endometrial tissues using machine learning applied to 1413 candidate variables. *Int J Gynecol Pathol*. 2020;39:333–343.
19. Papke Jr DJ, Lohmann S, Downing M, Hufnagl P, Mutter GL. Computational augmentation of neoplastic endometrial glands in digital pathology displays. *J Pathol*. 2020;253:257–267.
20. Sanderson PA, Esnal-Zufiaurre A, Arends MJ, et al. Improving the diagnosis of endometrial hyperplasia using computerized analysis and immunohistochemical biomarkers. *Front Reprod Health*. 2022;4:896170.
21. Sun H, Zeng X, Xu T, Peng G, Ma Y. Computer-aided diagnosis in histopathological images of the endometrium using a convolutional neural network and attention mechanisms. *IEEE J Biomed Health Inform*. 2019;24:1664–1676.
22. Zhao F, Dong D, Du H, et al. Diagnosis of endometrium hyperplasia and screening of endometrial intraepithelial neoplasia in histopathological images using a global-to-local multi-scale convolutional neural network. *Comput Methods Programs Biomed*. 2022;221:106906.
23. Steiner DF, Chen P-HC, Mermel CH. Closing the translation gap: AI applications in digital pathology. *Biochim Biophys Acta Rev Cancer*. 2021;1875:188452.
24. Raffone A, Travaglino A, Saccone G, et al. Endometrial hyperplasia and progression to cancer: which classification system stratifies the risk better? A systematic review and meta-analysis. *Arch Gynecol Obstet*. 2019;299:1233–1242.
25. Hecht JL, Ince TA, Baak JP, Baker HE, Ogden MW, Mutter GL. Prediction of endometrial carcinoma by subjective endometrial intraepithelial neoplasia diagnosis. *Mod Pathol*. 2005;18:324–330.
26. Lacey J, Ioffe O, Ronnett B, et al. Endometrial carcinoma risk among women diagnosed with endometrial hyperplasia: the 34-year experience in a large health plan. *Br J Cancer*. 2008;98:45–53.
27. Yang Y-F, Liao Y-Y, Peng N-F, Li L-Q, Xie S-R, Wang R-B. Prediction of coexistent carcinomas risks by subjective EIN diagnosis and comparison with WHO classification in endometrial hyperplasias. *Pathol Res Pract*. 2012;208:708–712.
28. Salman MC, Usubutun A, Boynukalin K, Yuce K. Comparison of WHO and endometrial intraepithelial neoplasia classifications in predicting the presence of coexistent malignancy in endometrial hyperplasia. *J Gynecol Oncol*. 2010;21:91–101.
29. Ordi J, Bergeron C, Hardisson D, et al. Reproducibility of current classifications of endometrial endometrioid glandular proliferations: further evidence supporting a simplified classification. *Histopathology*. 2014;64:284–292.
30. Travaglino A, Raffone A, Saccone G, et al. PTEN as a predictive marker of response to conservative treatment in endometrial hyperplasia and early endometrial cancer. A systematic review and meta-analysis. *Eur J Obstet Gynecol Reprod Biol*. 2018;231:104–110.
31. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst*. 2008;100:1432–1438.
32. Baak JP, Nauta J, Wisse-Brekelmans E, Bezemer P. Architectural and nuclear morphometrical features together are more important prognosticators in endometrial hyperplasias than nuclear morphometrical features alone. *J Pathol*. 1988;154:335–341.
33. Baak J, Kuik D, Bezemer P. The additional prognostic value of morphometric nuclear arrangement and DNA-ploidy to other morphometric and stereologic features in endometrial hyperplasias. *Int J Gynecol Cancer*. 1994;4:289–297.
34. Sherman ME, Ronnett BM, Ioffe OB, et al. Reproducibility of biopsy diagnoses of endometrial hyperplasia: evidence supporting a simplified classification. *Int J Gynecol Pathol*. 2008;27:318–325.
35. Usubutun A, Mutter GL, Saglam A, et al. Reproducibility of endometrial intraepithelial neoplasia diagnosis is good, but influenced by the diagnostic style of pathologists. *Mod Pathol*. 2012;25:877–884.
36. WHO Classification of Tumours Editorial Board. *Tumours of the Breast and Female Genital Organs*. 3rd ed. 4. International Agency for Research on Cancer; 2003.
37. Allison KH, Reed SD, Voigt LF, Jordan CD, Newton KM, Garcia RL. Diagnosing endometrial hyperplasia: why is it so difficult to agree? *Am J Surg Path*. 2008;32:691–698.
38. Skov B, Broholm H, Engel U, et al. Comparison of the reproducibility of the WHO classifications of 1975 and 1994 of endometrial hyperplasia. *Int J Gynecol Pathol*. 1997;16:33–37.
39. Bergeron C, Nogales FF, Masseroli M, et al. A multicentric European study testing the reproducibility of the WHO classification of endometrial hyperplasia with a proposal of a simplified working classification for biopsy and curettage specimens. *Am J Surg Path*. 1999;23:1102–1108.
40. Kendall BS, Ronnett BM, Isacson C, et al. Reproducibility of the diagnosis of endometrial hyperplasia, atypical hyperplasia, and well-differentiated carcinoma. *Am J Surg Path*. 1998;22:1012–1019.
41. Sanderson PA, Critchley HO, Williams AR, Arends MJ, Saunders PT. New concepts for an old problem: the diagnosis of endometrial hyperplasia. *Hum Reprod Update*. 2017;23:232–254.
42. Chen H, Strickland AL, Castrillon DH. Histopathologic diagnosis of endometrial precancers: updates and future directions. *Semin Diagn Pathol*. 2021;39:137–147.
43. Rewcastle E, Varhaugvik AE, Gudlaugsson E, et al. Assessing the prognostic value of PAX2 and PTEN in endometrial carcinogenesis. *Endocr Relat Cancer*. 2018;25:981–991.
44. Baak JPA, van Diermen B, Steinbakk A, et al. Lack of PTEN expression in endometrial intraepithelial neoplasia is correlated with cancer progression. *Hum Pathol*. 2005;36:555–561.



Supplementary figure 1: **A)** Example of endometrial hyperplasia (no progression) with highly fragmented regions due to mechanical stress during sampling. Scale bar 5mm. **B)** A close-up of a fragmented region consisting primarly of dense glands but lacking much of the stromal compartment. The only remaining stromal compartment is marked by a black arrow. Scale bar 500µm. **C)** The labeled image of the region shown in B, only the epithelial structures are labelled with minimal stroma. Measurement of this region would result in a higher than expected percentage stroma due to the loss of much of the stromal compartment. For this particular case, when the fragmented region was not removed manually it was selected for analysis by the ENDOAPP resulting in a percentage stroma of 32% and a high-risk score, false positive result. When removed the ENDOAPP selected an intact region resulting in a percentage stroma score of 60% and a low-risk result.



Supplementary figure 2: Examples of artifacts that may require manual removal: **A)** tissue artifact (scale bar 2.5mm) **B)** tissue fold (scale bar 1mm) **C)** coverslip glue artifact (scale bar 2mm) **D)** thick section resulting in a strong blue stain (hematoxylin) and weak brown stain (DAB) that in this case resulted in mislabelling by the ENDOAPP (scale bar 2mm).

PAPER III

The Ki67 Dilemma: Investigating Prognostic Cut-Offs and Reproducibility for Automated Ki67 Scoring in Breast Cancer

Emma Rewcastle[a,b,*], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]

[a] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b] Department of Pathology, Stavanger University Hospital, Stavanger, Norway.

*Corresponding author. E-mail address: emma.rewcastle@sus.no.

The authors declare there are no competing interests.

Keywords

Ki67, digital image analysis, prognostic, biomarkers, artificial intelligence

Abstract

Purpose

Quantification of Ki67 in breast cancer is a well-established prognostic and predictive marker, but inter-laboratory variability has hampered its clinical usefulness. This study compares the prognostic value and reproducibility of Ki67 scoring using four automated, digital image analysis (DIA) methods and two manual methods.

Methods

The study cohort consisted of 367 patients diagnosed between 1990 and 2004, with hormone receptor positive, HER2 negative, lymph node negative breast cancer. Manual scoring of Ki67 was performed using predefined criteria. DIA Ki67 scoring was performed using QuPath and Visiopharm® platforms. Reproducibility was assessed by the intraclass correlation coefficient (ICC). ROC curve survival analysis identified optimal cutoff values in addition to recommendations by the International Ki67 Working Group and Norwegian Guidelines. Kaplan-Meier curves, log-rank test and cox regression analysis assessed the association between Ki67 scoring and distant metastasis (DM) free survival.

Results

The manual hotspot and global scoring methods showed good agreement when compared to their counterpart DIA methods (ICC>0.780), and good to excellent agreement between different DIA hotspot scoring platforms (ICC: 0.781-0.906). Different Ki67 cutoffs demonstrate significant DM-free survival ($p<0.05$). DIA scoring had greater prognostic potential for DM-free survival using a 14% cutoff (HR: 3.054-4.077) than manual scoring (HR: 2.012-2.056). The use of a single cutoff for all scoring methods affected the distribution of prediction outcomes (e.g., false positives and negatives).

Conclusion

This study demonstrates that DIA scoring of Ki67 is superior to manual methods, but further study is required to standardize automated, DIA scoring and definition of a robust cut-off.

Introduction

Breast Cancer is the most common cancer globally and the fifth leading cause of cancer deaths [1]. In Norway, 4,247 new cases were reported in 2022 [2]. A panel of immunohistochemistry (IHC) biomarkers can classify breast cancer into four surrogate subtypes: hormone receptors (HR) (estrogen (ER) and progesterone (PR)), HER2 and Ki67. These classify luminal A-like (low Ki67, HR+, HER2-), luminal B-like (high Ki67, HR+, HER2-), HER2 positive (HER2+) and triple negative (HR-, HER2-) tumors [3, 4]. The endocrine-therapy sensitive luminal-like group is the largest category and may be further categorized, using the biomarker Ki67, into luminal A-like (low-Ki67) and luminal B-like (high-Ki67). The latter benefiting from additional chemotherapy, whilst the former does not [5].

The Ki67 score defines the percentage of positively stained tumor cells in a defined hotspot or global region [6, 7] and has prognostic, predictive, and monitoring potential [8-13]. Patients with a low Ki67 score have a low recurrence risk and may be spared from chemotherapy whilst patients with a high Ki67 score are associated with an increased risk of recurrence, higher mortality rate and may benefit more from adjuvant chemotherapy [14-20].

Despite its value as a prognostic biomarker, there is no global consensus on how to standardize Ki67 scoring. This includes standardization of pre-analytical and analytical conditions, a protocol for measurement of Ki67, and cutoff score for adjuvant treatment [6, 21-24]. In 2021 the St. Gallen consensus recommended the guidelines set by the International Ki67 in Breast Cancer Working Group (IKWG): patients with low Ki67 < 5% are not recommended for adjuvant chemotherapy whilst patients with a high Ki67 $\geq$ 30% are recommended [21]. However, treatment recommendations for the intermediate category (>5%, <30%) is still debated, and other cut-offs are still used.

With the rise of digital pathology and artificial intelligence (AI), digital scoring of Ki67 by automated algorithms or applications (APPs) have provided a new avenue for improved quantification. Various studies have demonstrated equal or improved reproducibility and accuracy using automated digital image analysis (DIA) compared to manual scoring methods [9, 25-30]. However, no recommendation

exists currently for automated DIA methods and established cutoffs for these methods have not yet been extensively validated.

In this study, we aim to compare the reproducibility and prognostic capacity of Ki67 score using four DIA scoring methods compared to using two manual methods (conventional-HS and global unweighted and weighted).

Materials and Methods

Study Cohort

The study received approval from the Regional Ethics Committee of Health West Norway (2010/1241) and informed consent waived. Patients who received a primary diagnosis of breast cancer between 1990-1998 (N=346) and 2000-2004 (N=253) at Stavanger University Hospital (SUH) were available for this study. Patients who received neoadjuvant treatment (N=8) were excluded. The following inclusion criteria were used to select cases: 1) archive tumor material available as a formalin-fixed paraffin embedded (FFPE) tissue block, 2) HR+ (ER $\geq$1%, and/or PR $\geq$10%) and HER2- status, and 3) at least one follow-up sample, for non-progression cases >6 months after initial diagnosis. Of the 591 cases, 64 (11%) were removed due to missing blocks/lack of tumor material, 19 (3%) were lost to follow-up and 141 HER2+/TNBC were excluded. In summary, 367 cases comprised the study cohort.

Datasets

The development dataset consisted of a training and tuning set (figure S1). A training set was used to create manual annotations to train classification algorithms. Five whole slide images (WSI) of good quality and representative of strong and weak staining, and high and low Ki67 positivity were selected. The WSIs were annotated with training labels for segmentation of key features: tissue and background, tumor and non-tumor, and positive brown and negative blue nuclei. Two of the four DIA methods required training (VIS1-HS, QuPath), whilst the two remaining methods (VIS2-HS, VIS2-G) were pre-

trained and validated. All DIA methods were run on the tuning dataset, which was used to monitor and evaluate algorithm performance, assess reproducibility, and define prognostic cutoffs.

Immunohistochemistry and Imaging

New 3µm tissue sections were cut, mounted on SuperFrost® Plus slides (Menzel Gläser, Braunschweig, Germany), and dried overnight at 37°C followed by 1 hour at 60°C. Sections were transferred deparaffinized to the Dako Omnis (Dako, Glostrup, Denmark). Antigen retrieval was performed using the EnVision FLEX Target Retrieval Solution High pH (Dako Omnis), heated at 97°C for 30 minutes. Sections were stained for Ki67 using a pre-optimized protocol using the Dako MIB-1 clone (dilution 1:50) and incubated for 20 minutes. Additionally, signal amplification was performed using the EnVision FLEX+ Mouse LINKER (Dako Omnis) with a 10-minute incubation.

Manual Hotspot Ki67 Quantification (conventional)

Using a microscope, the whole section was viewed at low power to identify the most proliferative region (hotspot) in the invasive tumor region. Non-invasive regions, necrotic regions and areas with high lymphocytic infiltration were avoided.

For the 1990-1998 cohort, a 40X objective was used to count positively stained tumor nuclei (brown) and the total number of tumor nuclei in the hotspot. For each case, at least 500 tumor cells were counted. If fewer than 500 tumor cells, adjacent fields of view (FOV) were counted. Ki67 score was calculated as the percentage positive Ki67 (Ki67 positive/(Ki67 positive + Ki67 negative)).

For the 2000-2004 cohort, the interactive QPRODIT system (Leica, Cambridge, UK) was used to score Ki67 as described by previously [31]. Within a hotspot tumor region, 250-350 FOV were defined, and a test grid used to classify each field as Ki67 positive or negative. A field was classified as positive if the first tumor cell that intersected with a grid point was positive, and vice versa for negative cells. Ki67 score was calculated as the percentage positive Ki67.

Both methods were grouped under the term conventional Ki67 score.

Manual Global Ki67 Quantification (Global unweighted and weighted)

Global weighted and unweighted scoring was performed according to the protocol set by the IKWG [32]. The IKWG Ki67 mobile counting tool was used (https://www.ki67inbreastcancerwg.org/). Using the NDP2.view2 image viewing software (v.2.9.29, Hamamatsu Photonics, Japan), a WSI of the Ki67-stained tissue was examined and the percentage area of negligible, low, medium, or high Ki67 was estimated and entered into the counting tool. Three to four circular annotations were placed to simulate a field of view, in each field type, as directed by the tool. In a typewriter fashion, 100 nuclei were counted as either negative or positive in each ROI. The unweighted and weighted global scores were recorded for each slide.

Digital Image Analysis (DIA)

Scanning: Whole sections stained with Ki67 were scanned at 40X magnification using the Hamamatsu Nanozoomer S60 (Hamamatsu Photonics, Hamamatsu City, Japan) at SUH.

Two platforms were used to score Ki67 on whole slide images (WSI): QuPath [33] and Visiopharm® (Version 2022.09.3.12885, Visiopharm A/S, Hørsholm, Denmark). The following hardware was used: Dell Precision 3640 Tower, Intel Core i9-10900, Nvidia GeForce RTX 2080 Ti.

VIS1-HS: An in-house APP (VIS1-HS) was developed using the Visiopharm® platform for quantification of Ki67 score (figure 1). Six standalone APPs were developed, that were batch run: 01 Tissue detection, 02 Positive nuclei detection, 03 Hotspot detection; 04 Tumor detection, 05 Nuclei segmentation, and 06 Hotspot Ki67 score quantification. Manual annotations from the training dataset were used to create labels to train the classifier APPs: (01) tissue and non-tissue training labels to train a tissue classifier, (02) positive nuclei labels to train a positive nuclei classifier, (04) tumor and non-tumor labels to train a tumor classifier, and (05) positive and negative nuclei labels to train a nuclei classifier. Training was performed until the classifier achieved accurate and consistent predictions, as determined by the operator. All classifier APPs used supervised K-means clustering. Specific parameters were trialed with the VIS1-HS APP (figure S1): the drawing radius for the heatmap

configuration (175µm, 400µm), number of hotspot ROI (1-5), and ROI size criteria (0.2mm^2, 1mm^2, minimum 550 tumor cell count, minimum 1000 tumor cell count).

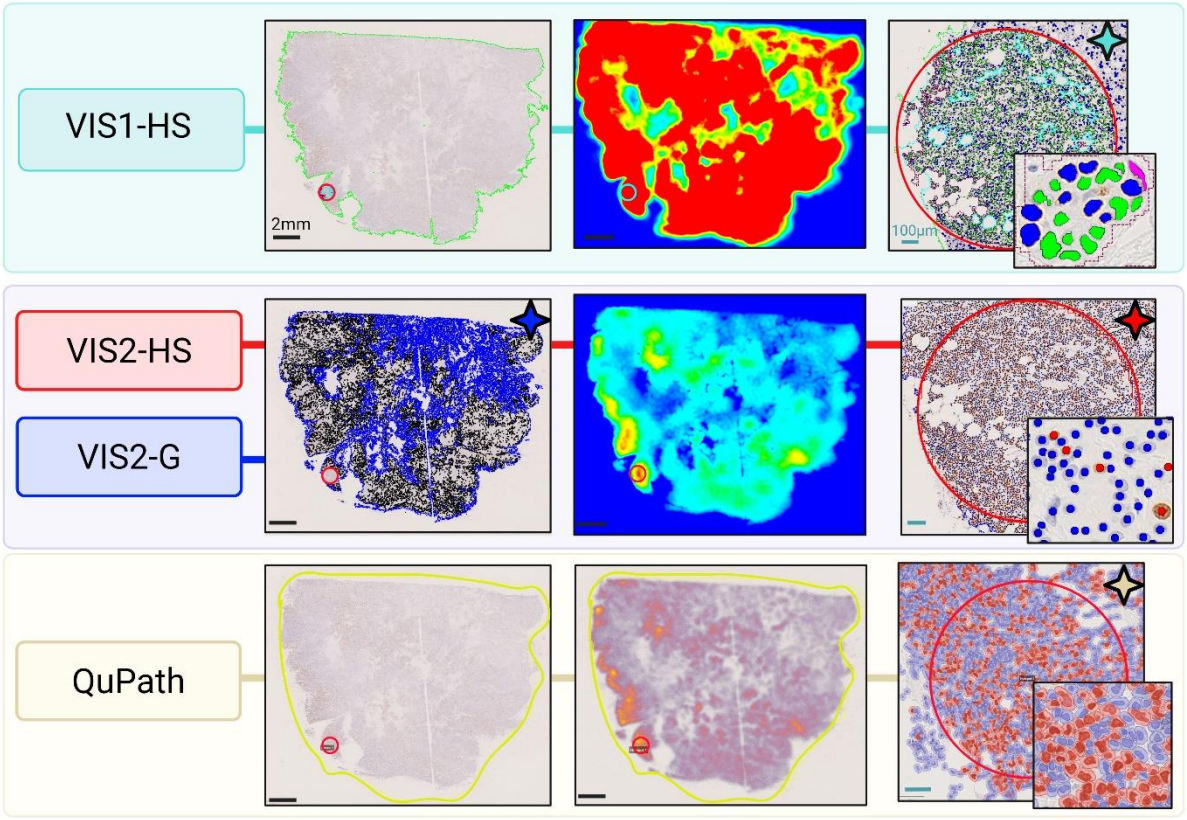


Figure 1*: Visualization of each digital image analysis Ki67 scoring method. The VIS1-HS method, developed in-house using the Visiopharm® commercial platform, utilized a 1mm^2 hotspot (red/cyan ROI) for Ki67 scoring. The VIS2 method, a commercial APP provided by Visiopharm®, measured both global Ki67 score in the entire tumor region – blue ROI (VIS2-G) and in a 0.96mm^2 hotspot – red ROI (VIS2-HS). The QuPath method required manual demarcation of the tissue (yellow ROI). The Ki67 score was calculated in a hotspot (red ROI) with minimum 500 tumor cells. Stars indicate when the Ki67 score was calculated for each method.*

VIS2-HS/G: Visiopharm® provided a commercial, CE-IVD Ki67 quantification APP (VIS2-HS/G) (figure 1). This method consists of six standalone APPs that were batch run, in the following order: 01 #10182 – IHC Tissue Detection, AI; 02 #10180 – Invasive Tumor Detection, AI; 03 #10180 – Invasive Tumor Postprocessing; 04 #10173 – Ki-67 Nuclei APP, Breast Cancer, AI; 05 #10114 Hot Spot Detection; 06 #10114 Hot Spot Quantification (table S1). This method generated both a global score (VIS2-G), which assessed the invasive tumor region of the entire WSI, and a hotspot score (VIS2-HS).

QuPath: A classifier for quantification of Ki67 score was developed using QuPath (figure 1), based on the protocol established by Acs and colleagues [34]. To train the cell classifier, positive and negative tumor cell nuclei were annotated manually on WSI from the training dataset. A tumor ROI was first manually drawn around the invasive tumor region and the cell classifier script (table S2) run. The density maps tool generated a heatmap and three ROIs were automatically placed on suspected hotspots. Each hotspot was reviewed, and the highest scoring hotspot was approved according to the following criterion: minimum 500 tumor cells and satisfactory labelling of positive and negative tumor nuclei. If the ROI contained less than 500 tumor cells, it was manually enlarged until this criterion was met. The QuPath method was semi-automated; manual delineation of the tumor region was required for each case, and the classifier and density maps were run separately per case.

Statistical analysis

Statistical analyses were performed with SPSS for Windows (version 26.0.0; IBM SPSS Statistics) and R studio (2023.06+561). Assumptions were verified for each test and $p<0.05$ was significant. Mann Whitney U and Kruskall Wallis tests were used to test for significant differences between patients with no distant metastases (DM) and with DM. Level of agreement, using the intraclass correlation coefficient (ICC), was assessed on transformed Ki67 scores (multiplied by a factor of 10 and log transformed). Receiving operating characteristic (ROC) curves were generated for each scoring method and the area under the curve (AUC) was used to assess a method's discriminative ability. Kaplan Meier survival analyses were performed to assess the prognostic value of DM-free survival and compared using log-rank (endpoint: first diagnosis of a DM in the follow-up or censored according to last-known follow-up date). Cox regression models for univariate and multivariate analysis was performed for: age, tumor size, mitotic activity index, Nottingham grade, operation type, adjuvant treatment, and Ki67 score.

Results

Overview

Of the 367 cases from the development dataset, 12 cases were ineligible for analysis due to poor quality material and 61 cases due to poor staining (figure S2). This left 294 cases (table S3). A further 9 cases failed analysis with QuPath due to: a false positive edge effect, high numbers of inflammatory cells, necrosis, and artefacts (figure S2).

Performance of DIA methods

For the VIS1-HS method various ROI specifications were trialed and differences in Ki67 scores for each was considered marginal (table S4). A $1mm^2$ ROI was selected for the final VIS1-HS method as it was the most consistent for detecting over 500 tumor cells.

All DIA methods required manual editing for some or all cases. For VIS1-HS, nearly all cases required manual removal of either DCIS, artefacts, inflammatory cell clusters or normal tissue. For VIS2-HS, 13% of cases (37/294) required a manual edit. The VIS2-G method required a manual edit in 28% of cases (81/294). For QuPath, all cases required a manual delineation of the tissue region for analysis, and 45% of cases (128/285) required a manual intervention for HS ROI placement or expansion.

Comparison of Ki67 score for four DIA scoring and two manual methods

For DIA and manual scoring methods the mean Ki67 score ranged from 9.5% (VIS2-G) to 16.2% (QuPath) (table 1, figure S3). For cases with no distant metastases (no DM) compared to cases with distant metastases (DM) in the follow-up, the greatest mean difference was recorded for QuPath (11.9%), followed by the VIS2-HS (9.6%), global weighted (8.6%), VIS1-HS (7.2%), VIS2-G (6.9%), global unweighted (6.8%), and conventional hotspot (5.9%) methods. Furthermore, QuPath recorded higher scores on average, with the VIS2-G method recording the lowest (table 1).

Table 1: Quantification of Ki67 and total tumor cell count according to four digital scoring methods, and two manual scoring methods (conventional hotspot, global unweighted and weighted).

Method	**N=**	**%Ki67 (all)**	**%Ki67 (No DM)**	**%Ki67 (DM)**	**Total TC count (all)**
Conventional hotspot	N=284 No DM: 243 DM: 41	13.7 (0-83)	12.8 (0-83)	18.7 (0-65)	No data
Global unweighted	N=294 No DM: 257 DM: 37	15.9 (0-100)	15.0 (0-100)	21.8 (0-57.3)	No data
Global weighted	N=294 No DM: 257 DM: 37	16.0 (0-100)	14.9 (0-100)	23.5 (0-77.9)	No data
VIS1-HS	N=294 No DM: 254 DM: 40	13.3 (0-79)	12.3 (0-79)	19.5 (2.8-74.6)	2907 (410-12727) [a]
VIS2-HS	N=294 No DM: 254 DM: 40	14.4 (0.2-89.8)	13.1 (0.2-.78.2)	22.7 (1.1-89.8)	4094 (251-11853) [b]
VIS2-G	N=294 No DM: 254 DM: 40	9.5 (0.1-70.3)	8.6 (0.1-.70.3)	15.5 (0.7-64.6)	122966 (531-1013920)
QuPath	N=285 No DM: 245 DM:40	16.2 (0-92.9)	14.6 (0-91.3)	26.5 (2.5-92.9)	662 (500-1697)

Mean values are reported with range in parentheses.

[a]*Two cases were <500 tumor cells.*

[b]*Three cases were <500 tumor cells.*

Reproducibility is one of the primary concerns regarding scoring of Ki67. In the study cohort, there was moderate to excellent agreement between all scoring methods (table 2). Comparison of all hotspot scoring methods revealed good to excellent agreement (ICC: 0.781–0.906). Agreement was on average higher between automated hotspot methods than between automated and manual hotspot methods (table 2). There was good agreement between the global DIA method and manual global method (ICC: 0.803–0.810). Agreement was lower between global and hotspot scoring methods (ICC: 0.636–0.759).

Table 2: Agreement of manual and digital image analysis Ki67 quantification methods as assessed by the intraclass correlation coefficient (ICC).

Method	Measure	Method					
		Conventional	*Global unweighted*	*Global weighted*	*VIS1-HS*	*VIS2-HS*	*VIS2-G*
Conventional	ICC (95% CI)						
Global unweighted	ICC (95% CI)	0.758 (0.698-0.806)					
Global weighted	ICC (95% CI)	0.746 (0.684-0.797)	0.986 (0.982-0.989)				
VIS1-HS	ICC (95% CI)	0.802 (0.743-0.848)	0.648 (0.559-0.719)	0.636 (0.535-0.715)			
VIS2-HS	ICC (95% CI)	0.846 (0.794-0.884)	0.759 (0.698-0.807)	0.752 (0.681-0.806)	0.874 (0.843-0.899)		
VIS2-G	ICC (95% CI)	0.834 (0.642-0.909)	0.803 (0.731-0.853)	0.810 (0.756-0.852)	0.735 (0.166-0.887)	0.863 (0.152-0.955)	
QuPath	ICC (95% CI)	0.781 (0.680-0.845)	0.729 (0.643-0.792)	0.709 (0.609-0.781)	0.835 (0.796-0.867)	0.906 (0.881-0.926)	0.786 (0.153-0.918)

Defining a prognostic threshold

In this study, a ROC curve was generated to assess DM-free survival and identify optimal cutoffs. The two methods with the highest AUC were QuPath, 0.721 (95% CI: 0.643-0.798) and VIS2-HS, 0.705 (95% CI: 0.625-0.785). The manual methods recorded the lowest AUC: conventional hotspot, 0.655 (95% CI: 0.569-0.741); global weighted, 0.648 (95% CI: 0.554-0.744); and global unweighted, 0.636 (95% CI: 0.541-0.731). A range of coordinates were selected from the ROC curve for binary categorization of Ki67 score, which revealed similar DM-free survival (table S5, figure S4). The manual methods demonstrated lower p-values (log-rank) for cut-offs around 10%, whereas automated methods reported lower p-values between 10-14% (table S5). A 14% cutoff was chosen for further evaluation, due to its recommendation by the Norwegian Guidelines [35].

Binary categorization of Ki67-14% for all methods was significantly associated with DM-free survival in a 20-year follow-up period (figure 2a). All methods demonstrated a significant separation of patients with DM-free survival for low (<14%) and high (≥14%) Ki67, with the manual methods reporting the smallest separation in comparison to DIA (figure 2a). Additionally, percentage agreement was highest for VIS1-HS and VIS2-HS methods and lowest between VIS2-G and QuPath methods (table S6).

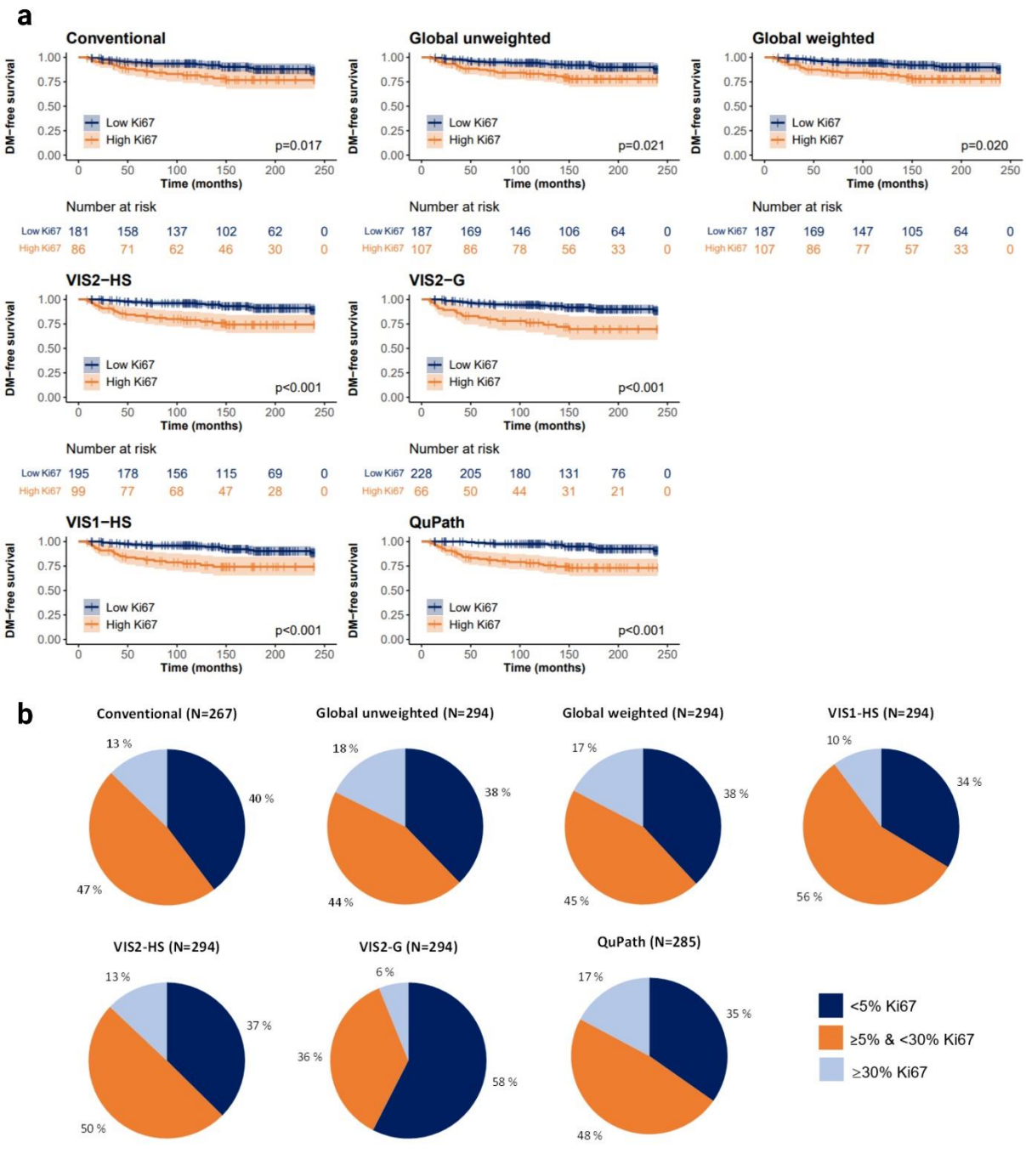


***Figure 2**: Evaluation of a prognostic cutoff. **a)** Distant-metastasis free survival of patients with HR+, HER2-, LN- breast cancer categorized by low Ki67 (<14%) and high Ki67 (≥14%) over a 20-year follow-up period. Scoring of Ki67 was performed by two manual scoring methods (conventional hotspot and global weighted and unweighted) and four digital image analysis methods: VIS1-HS, QuPath, VIS2-HS and VIS2-G. **b)** Proportion of cases assigned a low (<5%), intermediate (>5% & <30%) and high (≥ 30%) Ki67 score according to two manual scoring methods (conventional hotspot, global weighted and unweighted) and four DIA scoring methods (VIS1-HS, VIS2-HS, VIS2-G, and QuPath).*

International recommendations suggest <5% to assign low Ki67 and ≥30% to assign high Ki67, with remaining cases falling into an intermediate category. The VIS2-G method had the largest proportion

of low (<5%) Ki67 cases (58%) of all methods (figure 2b). QuPath and global weighted/unweighted methods had the highest proportion of high (≥30%) Ki67 cases (17-18%). The VIS2-G method reported the highest number of false negatives (low Ki67, DM), whilst the highest number of false positives was recorded by the manual global methods (high Ki67, no DM), respectively (figure 3). This was also observed at a 10-year follow-up.

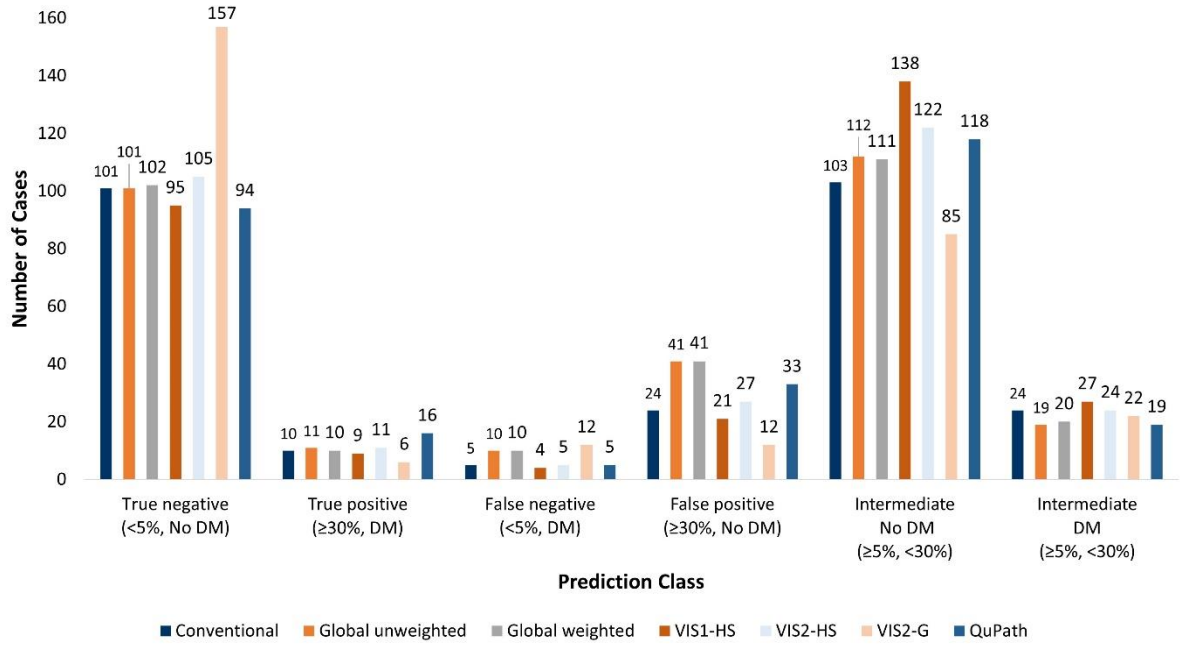


***Figure 3**: Prediction classes using <5% and ≥30% Ki67 thresholds set by the International Ki67 in Breast Cancer Working Group, as measured by a manual and digital image analysis scoring methods.*

Multivariate analysis

To assess the prognostic value of Ki67 scoring, we performed Cox regression analysis of each method alongside established prognostic markers. Of the variables tested, Nottingham grade, mitotic activity index (MAI10), tumor-size (2 cm), and Ki67 for all methods (14%) were statistically significant predictors of DM-free survival in univariate analysis (table 3). Hazard ratios (univariate) for all DIA methods (Ki67 14%) ranged from 3.054 to 4.077, with overlapping 95% confidence intervals whilst the manual methods had lower hazard ratios: 2.012-2.056 (table 3). In a multivariate analysis, Nottingham Grade, treatment and operation type were predictors in the final model.

Table 3: Distant-metastasis free survival in hormone receptor positive, HER2 negative, lymph node negative, breast cancer patients (*values reported in italics use continuous variables)*.

Characteristic	**All**		**Log rank p-value**	**HR**	**Cox regression p-value**	**95% CI**
	Event/At risk					
Age	N=316			*0.976*	*0.168*	*0.944-1.010*
<55	18/134	13%	0.676	1.140	0.677	0.617-2.107
≥55	24/182	13%				
Nottingham Grade	N=314					
Grade I (3-5)	6/108	6%	0.001*	2.441	0.054	0.984-6.053
Grade II (6-7)	21/149	14%		5.344	<0.001*	2.070-13.794
Grade III (8-9)	15/57	26%				
Tumor size[a]	N=314			*1.927*	*<0.001**	*1.404-2.645*
≤2cm	28/258	11%	0.003*	2.553	0.004*	1.341-4.862
>2cm	14/56	25%				
Operation Type	N=250					
Conservative/ Lumpectomy	15/152	10%	0.004*	2.524	0.005*	1.313-4.850
Mastectomy	23/98	23%				
MAI	N=309			*1.018*	*0.062*	*0.999-1.037*
<10	26/242	11%	0.005*	2.410	0.007*	1.276-4.553
≥10	15/67	22%				
Ki67						
Conventional	N=267			*1.014*	*0.041**	*1.001-1.028*
<14	20/181	11%	0.026*	2.012	0.029*	1.073-3.774
≥14	19/86	22%				
Global unweighted	N=294			*1.010*	*0.133*	*0.997-1-023*
<14	19/187	10%	0.021*	2.047	0.024*	1.099-3.813
≥14	21/107	20%				
Global weighted	N=294			*1.012*	*0.048**	*1.000-1.024*
<14	19/187	10%	0.020*	2.056	0.023*	1.104-3.829
≥14	21/107	20%				
VIS1-HS	N=294			*1.021*	*0.006**	*1.006-1.037*
<14	18/205	9%	<0.001*	3.054	<0.001*	1.633-5.711
≥14	22/89	25%				
VIS2-HS	N=294			*1.023*	*0.001**	*1.010-1.037*
<14	16/195	8%	<0.001*	3.245	<0.001*	1.721-6.116
≥14	24/99	24%				
VIS2-G	N=294			*1.025*	*0.004**	*1.008-1.042*
<14	22/228	10%	<0.001*	3.088	<0.001*	1.654-5.766
≥14	18/66	27%				
QuPath	N=285			*1.022*	*<0.001**	*1.010-1.034*
<14	12/175	7%	<0.001*	4.077	<0.001*	2.070-8.030
≥14	28/110	25%				

**Denotes significance p<0.05*

[a]*Only one case had a tumor sizes ≥5cm, so tumor size was grouped into ≤2cm and >2cm (T1 and T2), instead of three groups (T1, T2, T3).*

Discussion

We compare several automated DIA tools for global and hotspot Ki67 compared to two manual methods in HR+, HER2-, LN- breast tumors. Although Ki67 is considered an important biomarker in breast cancer, the concerns surrounding lack of standardization and poor reproducibility, have brought its value into question. In this study, we observed that commercial DIA tools (VIS2-HS, VIS2-G) required notably less manual editing compared to the in-house methods (VIS1-HS, QuPath).

Inter-platform variability demonstrated good to excellent agreement between all hotspot scoring methods and all global methods (ICC>0.8). Another inter-platform study reported excellent reproducibility (ICC>0.9) [26] and our observation of strong agreement between manual and DIA platforms is consistent with observations in the literature [7, 9, 36-42]

Although efforts have been made to standardize Ki67 scoring, both hotspot and global Ki67 score are still reported in the literature and in the Nordics [6, 38, 43, 44]. In addition, a range of cutoffs for defining low and high Ki67 are still reported (range: 10-20%) [17, 26, 45, 46]. We observed a range of prognostic cutoffs, with 14% being optimal and in agreement with Norwegian guidelines [6]. The hazard ratios (HR: 2.7-3.7) reported by Acs *et al*. [26] for DIA scoring on core needle biopsies and tissue microarrays were similar to those reported in our study (HR: 3.1-4.1) and Boyaci *et al*. [47] (HR: 2.6-4.2) for DIA scoring on surgical specimens. Furthermore, we observed that DIA methods had a greater discriminative capacity, using the AUC metric, than manual methods. This was reflected in the hazard ratios for Ki67 score (14%) and DM-free survival (DIA HR: 3.054-4.077 vs. Manual HR: 2.012-2.056). Another study observed similar hazard ratios for DIA scoring (hotspot HR: 6.88; global HR: 3.13) compared to a manual hotspot method (HR: 2.76), for recurrence free survival [48].

In 2021, the St. Gallen consensus adopted the IKWG recommendation of <5% (low) and >30% (high), with patients between 5-30% (intermediate) not recommended for treatment decisions by Ki67[38]. In our study, evaluation of Ki67 score using these thresholds revealed that QuPath had the largest proportion of Ki67 high cases and highest number of false positives (high Ki67, no DM). Whilst, the

VIS2-G, global scoring method, demonstrated the highest proportion of Ki67 low cases and highest number of false negatives (low Ki67, DM). This suggests that regardless of using a more restrictive cutoff, patients are still at risk of over- and under-treatment. Furthermore, assessing the number of false positives and negatives revealed differences in clinical consequence between methods. This is important for future implementation of DIA methods as the choice of method: hotspot or global scoring, can have notable differences in classification of patients.

A high total tumor count resulted in lower Ki67 scores. The average total number of tumor cells measured for global DIA score was >100,000, far more than for all other methods and it consistently reported lower Ki67 scores than the others. Norwegian guidelines and IKWG recommendations for Ki67 scoring recommend a minimum of 400 to 500 tumor cells scored [6, 49]. Our results suggest that Ki67 scores from tumor cell counts around 2000-4000 cells, generated by a $1mm^2$ ROI, were more consistent, with fewer potentially under- or over-treated cases. Observations from Robertson *et al*. revealed greater reproducibility with increasing tumor cell counts (from 200 to 1000 tumor cells)[48]. This suggests the need for caution when translating current manual methods to a DIA method. A larger study is required to affirm the optimal number of total tumor cells for Ki67 score by DIA.

As use of molecular signature tests for classification of breast cancer increases, the use of Ki67 for treatment decisions is called into question. Molecular testing has demonstrated prognostic and predictive value [50-54]. However, where such molecular panels are unavailable, considered costly, or introduce delays to treatment due to slower return of results, Ki67 could be an equivalent approach. Furthermore, Ki67 has the potential to be used as a screening tool for recommending molecular testing in intermediate cases (Ki67>5%, <30%) [38]. Majority consensus warrants use of both multigene panels and Ki67 score [21].

This study does not come without its limitations. The study utilized a retrospective cohort, and many of the tissue blocks were >20 years old, therefore the staining protocol had to be adjusted due to antigen decay. Previously stained Ki67 slides (from 2011) were used as a quality control. Additionally,

only one scanner type and one automated IHC-instrument was used whereas multiple different scanners and staining methods, from different locations and time periods, on different or the same tissue blocks, would be worth investigating. In the present study, we only used one global scoring DIA method (VIS2-G), and future work to compare global DIA reproducibility could be pursued. For a future study, it would be worthwhile to investigate DIA scoring of Ki67 in a prospective cohort, with molecular profile data, with a planned long-term follow-up, such as the EMIT study [55].

In summary, we report good agreement between manual and counterpart DIA scoring methods. DIA Ki67 scoring methods had a greater discriminative capacity for DM-free survival than manual methods. A range of cutoffs was prognostic for each method, but the choice of scoring method and cutoff can lead to notable differences in the number of patients to be treated or tested, emphasizing the need for further validation in a prospective cohort. Total tumor cell count contributed to changes in risk categorization using the recommended 5% and 30% threshold. Automated, DIA methods may improve reproducibility and prognostic value of Ki67 scoring in comparison to manual methods, if standardized.

Acknowledgements

The authors would like to thank Marit Nordhus for her technical assistance and the Units for Immunohistochemistry and Histology at Stavanger University Hospital for their technical support.

Ethics Approval and Consent to Participate

The study was retrospective and received approval from the Regional Ethics Committee of Health West Norway (2010/1241) and informed consent waived. The study was performed in accordance with the Declaration of Helsinki.

Author Contributions

Emma Rewcastle and Emiel AM Janssen designed the study. Emma Rewcastle, Ivar Skaland and Emiel AM Janssen contributed to the development of the methodology. Emma Rewcastle, Jan PA Baak, Einar Gudlaugsson, Silja K Fykse, and Emiel AM Janssen performed data acquisition. Emma Rewcastle

performed statistical analyses. Emma Rewcastle, Ivar Skaland, Einar Gudlaugsson, Jan PA Baak and Emiel AM Janssen interpreted the data. Einar Gudlaugsson and Jan PA Baak acted as medical consultants. Emma Rewcastle wrote the manuscript. All authors read, corrected, and approved the final paper.

Funding

The study was funded by the Helse Vest Strategic Research Fund as part of the Pathology in Western Norway project. The Visiopharm® CE-IVD Ki67 Application, used in this study, was provided free-of-charge for academic purposes. Visiopharm® did not provide any financial support and did not have any additional role in study design, data collection and analysis, and decision to publish or preparation of the manuscript.

Data Availability

The patient databases used in this study are not publicly available due to ethical and legal concerns. Anonymized data can be requested from Stavanger University Hospital Institutional Data Access/Ethics Committee (contact via email: rek-vest@uib.no, REK vest, Rogaland, Vestland, Norway) for researchers who meet the criteria for access to confidential data.

References

1. Sung H, Ferlay J, Siegel RL, *et al.* (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 71:209-249. https://doi.org/10.3322/caac.21660

2. Cancer Registry of Norway. Cancer in Norway 2022 - Cancer incidence, mortality, survival and prevalence in Norway. https://www.kreftregisteret.no/globalassets/cancer-in-norway/2022/cin_report-2022.pdf. Accessed 12 December 2023.

3. Yersal O, Barutca S (2014) Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol* 5:412-424. https://doi.org/10.5306%2Fwjco.v5.i3.412

4. WHO Classification of Tumors Editorial Board (2019) *Breast Tumours*. 5th edn. International Agency for Research on Cancer.

5. Curigliano G, Burstein HJ, Winer EP, *et al*. (2017) De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary

Therapy of Early Breast Cancer 2017. *Ann Oncol* 28:1700-1712. https://doi.org/10.1093/annonc/mdx308

6. Norwegian Directorate of Health (2021) Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft (National Guidelines for Diagnosis, Treatment and Follow-up of Patients with Breast Cancer). https://nbcgblog.files.wordpress.com/2021/03/nasjonalt-handlingsprogram-for-pasienter-med-brystkreft-01.03.2021-16-utgave.pdf. Accessed 12 December 2023.

7. Leung SC, Nielsen TO, Zabaglo LA, *et al*. (2019) Analytical validation of a standardised scoring protocol for Ki67 immunohistochemistry on breast cancer excision whole sections: an international multicentre collaboration. *Histopathology* 75:225-235. https://doi.org/10.1111/his.13880

8. Smith I, Robertson J, Kilburn L, *et al.* (2020) Long-term outcome and prognostic value of Ki67 after perioperative endocrine therapy in postmenopausal women with hormone-sensitive early breast cancer (POETIC): an open-label, multicentre, parallel-group, randomised, phase 3 trial. *Lancet Oncol* 21:1443-1454. https://doi.org/10.1016/S1470-2045(20)30458-7

9. Rimm DL, Leung SC, McShane LM, *et al*. (2019) An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer. *Mod Pathol* 32:59-69. https://doi.org/10.1038/s41379-018-0109-4

10. Cheang MC, Chia SK, Voduc D, *et al.* (2009) Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst* 101:736-750. https://doi.org/10.1093/jnci/djp082

11. De Azambuja E, Cardoso F, de Castro G, *et al.* (2007) Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12 155 patients. *Br J Cancer* 96:1504-1513. https://doi.org/10.1038/sj.bjc.6603756

12. Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA (2010) Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol* 11:174-183. https://doi.org/10.1016/S1470-2045(09)70262-1

13. Penault-Llorca F, André F, Sagan C, *et al.* (2009) Ki67 expression and docetaxel efficacy in patients with estrogen receptor-positive breast cancer. *J Clin Oncol* 27:2809-2815. https://doi.org/10.1200/JCO.2008.18.2808

14. Harbeck N, Rastogi P, Martin M, *et al*. (2021) Adjuvant abemaciclib combined with endocrine therapy for high-risk early breast cancer: updated efficacy and Ki-67 analysis from the monarchE study. *Ann Oncol* 32:1571-1581. https://doi.org/10.1016/j.annonc.2021.09.015

15. Lee AK, Loda M, Mackarem G, *et al*. (1997) Lymph node negative invasive breast carcinoma 1 centimeter or less in size (T1a, bN0M0) Clinicopathologic features and outcome. *Cancer* 79:761-771. https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4%3C761::AID-CNCR13%3E3.0.CO;2-Y

16. Klintman M, Strand C, Ahlin C, *et al.* (2013) The prognostic value of mitotic activity index (MAI), phosphohistone H3 (PPH3), cyclin B1, cyclin A, and Ki67, alone and in combinations, in node-negative premenopausal breast cancer. *PLoS One* 8:e81902. https://doi.org/10.1371/journal.pone.0081902

17. Abubakar M, Orr N, Daley F, *et al.* (2016) Prognostic value of automated KI67 scoring in breast cancer: a centralised evaluation of 8088 patients from 10 study groups. *Breast Cancer Res* 18:104. https://doi.org/10.1186/s13058-016-0765-6

18. Viale G, Regan MM, Mastropasqua MG, *et al*. (2008) Predictive value of tumor Ki-67 expression in two randomized trials of adjuvant chemoendocrine therapy for node-negative breast cancer. *J Natl Cancer Inst* 100:207-212. https://doi.org/10.1093/jnci/djm289

19. Viale G, Giobbie-Hurder A, Regan MM, *et al*. (2008) Prognostic and predictive value of centrally reviewed Ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: results from Breast International Group Trial 1-98 comparing adjuvant tamoxifen with letrozole. *J Clin Oncol* 26:5569-5575. https://doi.org/10.1200%2FJCO.2008.17.0829

20. Petrelli F, Viale G, Cabiddu M, Barni S (2015) Prognostic value of different cut-off levels of Ki-67 in breast cancer: a systematic review and meta-analysis of 64,196 patients. *Breast Cancer Res Treat* 153:477-491. https://doi.org/10.1007/s10549-015-3559-0

21. Thomssen C, Balic M, Harbeck N, Gnant M (2021) St. Gallen/Vienna 2021: a brief summary of the consensus discussion on customizing therapies for women with early breast cancer. *Breast Care (Basel)* 16:135-143. https://doi.org/10.1159/000516114

22. Polley MYC, Leung SC, McShane LM, *et al.* (2013) An international Ki67 reproducibility study. *J Natl Cancer Inst* 105:1897-1906. https://doi.org/10.1093/jnci/djt306

23. Polley MYC, Leung SC, Gao D, *et al*. (2015) An international study to increase concordance in Ki67 scoring. *Mod Pathol* 28:778-786. https://doi.org/10.1038/modpathol.2015.38

24. Røge R, Nielsen S, Riber-Hansen R, Vyberg M. (2019) Impact of primary antibody clone, format, and stainer platform on Ki67 proliferation indices in breast carcinomas. *Appl Immunohistochem Mol Morphol* 27:732-739. https://doi.org/10.1097/PAI.0000000000000799

25. Skjervold AH, Pettersen HS, Valla M, Opdahl S, Bofin AM (2022) Visual and digital assessment of Ki-67 in breast cancer tissue - a comparison of methods. *Diagn Pathol* 17:1-14. https://doi.org/10.1186/s13000-022-01225-4

26. Acs B, Pelekanou V, Bai Y, *et al*. (2019) Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. *Lab Invest* 99:107-117. https://doi.org/10.1038/s41374-018-0123-7

27. Kwon A-Y, Park HY, Hyeon J, *et al*. (2019) Practical approaches to automated digital image analysis of Ki-67 labeling index in 997 breast carcinomas and causes of discordance with visual assessment. *PLoS One* 14:e0212309. https://doi.org/10.1371/journal.pone.0212309

28. Stålhammar G, Martinez NF, Lippert M, *et al.* (2016) Digital image analysis outperforms manual biomarker assessment in breast cancer. *Mod Pathol* 29:318-329.

29. Stålhammar G, Robertson S, Wedlund L, *et al*. (2018) Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. *Histopathology* 72:974-989. https://doi.org/10.1038/modpathol.2016.34

30. Gudlaugsson E, Skaland I, Janssen EA, *et al*. (2012) Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology* 61:1134-1144. https://doi.org/10.1111/j.1365-2559.2012.04329.x

31. Egeland NG, Austdal M, van Diermen-Hidle B, *et al*. (2019) Validation study of MARCKSL1 as a prognostic factor in lymph node-negative breast cancer patients. *PLoS One* 14:e0212527. https://doi.org/10.1371/journal.pone.0212527

32. International Ki67 in Breast Cancer Working Group. https://www.ki67inbreastcancerwg.org/ Published 2009. Accessed February 2, 2024

33. Bankhead P, Loughrey MB, Fernández JA, *et al*. (2017) QuPath: Open source software for digital pathology image analysis. *Sci Rep* 7:16878. https://doi.org/10.1038/s41598-017-17204-5

34. Acs B, Leung SC, Kidwell KM, *et al*. (2022) Systematically higher Ki67 scores on core biopsy samples compared to corresponding resection specimen in breast cancer: a multi-operator and multi-institutional study. *Mod Path.* 35:1362-1369. https://doi.org/10.1038/s41379-022-01104-9

35. Norwegian Directorate of Health (2023) Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft (National Guidelines for Diagnosis, Treatment and Follow-up of Patients with Breast Cancer). https://nbcgblog.files.wordpress.com/2023/02/11.01.2023-nasjonalt-handlingsprogram-for-brystkreft-19.-utgave-publisert-11.01.23.pdf. Accessed 12 December 2023.

36. Pons L, Hernández-León L, Altaleb A, *et al.* (2022) Conventional and digital Ki67 evaluation and their correlation with molecular prognosis and morphological parameters in luminal breast cancer. *Sci Rep* 12:8176. https://doi.org/10.1038/s41598-022-11411-5

37. Paik S, Kwon Y, Lee MH, *et al*. (2021) Systematic evaluation of scoring methods for Ki67 as a surrogate for 21-gene recurrence score. *NPJ breast cancer* 7:1-8. https://doi.org/10.1038/s41523-021-00221-z

38. Nielsen TO, Leung SCY, Rimm DL, *et al*. (2021) Assessment of Ki67 in breast cancer: updated recommendations from the international Ki67 in breast cancer working group. *J Natl Cancer* 113:808-819. https://doi.org/10.1093/jnci/djaa201

39. Jang MH, Kim HJ, Chung YR, Lee Y, Park SY (2017) A comparison of Ki-67 counting methods in luminal breast cancer: the average method vs. the hot spot method. *PLoS One* 12:e0172031. https://doi.org/10.1371/journal.pone.0172031

40. Thakur SS, Li H, Chan AM, *et al.* (2018) The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. *PLoS One* 13:e0188983. https://doi.org/10.1371/journal.pone.0188983

41. Røge R, Riber-Hansen R, Nielsen S, Vyberg MJB. (2016) Proliferation assessment in breast carcinomas using digital image analysis based on virtual Ki67/cytokeratin double staining. *Breast Cancer Res Treat* 158:11-19. https://doi.org/10.1007/s10549-016-3852-6

42. Shui R, Yu B, Bi R, Yang F, Yang W (2015) An interobserver reproducibility analysis of Ki67 visual assessment in breast cancer. *PLoS One* 10:e0125131. https://doi.org/10.1371/journal.pone.0125131

43. Regional Cancer Centres in Sweden. Bröstcancer vårdprogram - Kvalitetsbilaga för bröstpatologi (KVAST-bilaga) (Breast Cancer Care Program - Quality supplement for breast pathology (KVAST-supplement)). https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardprogram/kvalitetsdokument-for--patologi/https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardprogram/kvalitetsdokument-for--patologi/ Accessed 17 September 2023.

44. Danish Multidisciplinary Cancer Groups. Patologiprocedurer og molekylærpatologiske analyser ved brystkræft (Pathology Procedures and Molecular Pathology Analyses for Breast

Cancer), v.1.3. https://www.dmcg.dk/siteassets/forside/kliniske-retningslinjer/godkendte-kr/dbcg/dbcg_patologiprocedure-v1.3_admgodk040422.pdf. Accessed 17 September 2023

45. Stuart-Harris R, Caldas C, Pinder S, Pharoah P (2008) Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients. *Breast* 17:323-334. https://doi.org/10.1016/j.breast.2008.02.002

46. Arihiro K, Oda M, Ohara M, *et al*. (2016) Comparison of visual assessment and image analysis in the evaluation of Ki-67 expression and their prognostic significance in immunohistochemically defined luminal breast carcinoma. *Jpn J Clin Oncol* 46:1081-1087. https://doi.org/10.1093/jjco/hyw107

47. Boyaci C, Sun W, Robertson S, Acs B, Hartman J (2021) Independent clinical validation of the automated Ki67 scoring guideline from the International Ki67 in Breast Cancer Working Group. *Biomolecules* 11:1612. https://doi.org/10.3390/biom11111612

48. Robertson S, Acs B, Lippert M, Hartman J (2020) Prognostic potential of automated Ki67 evaluation in breast cancer: different hot spot definitions versus true global score. *Breast Cancer Res Treat*. 183:161-175. https://doi.org/10.1007/s10549-020-05752-w

49. Dowsett M, Nielsen TO, A'Hern R, *et al*. (2011) Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst* 103:1656-1664. https://doi.org/10.1093/jnci/djr393

50. Paik S, Shak S, Tang G, *et al.* (2005)Expression of the 21 genes in the Recurrence Score assay and tamoxifen clinical benefit in the NSABP study B-14 of node negative, estrogen receptor positive breast cancer. 2005 ASCO Annual Meeting Abstract. *J Clin Oncol* 23:510. https://doi.org/10.1200/jco.2005.23.16_suppl.510

51. Sparano JA, Gray RJ, Ravdin PM, *et al*. (2019) Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. *N Engl J Med* 380:2395-2405. https://doi:10.1056/NEJMoa1904819

52. Piccart M, van't Veer LJ, Poncet C, *et al.* (2021) 70-gene signature as an aid for treatment decisions in early breast cancer: updated results of the phase 3 randomised MINDACT trial with an exploratory analysis by age. *Lancet Oncol* 22:476-488. https://doi.org/10.1016/S1470-2045(21)00007-3

53. Ohnstad HO, Borgen E, Falk RS, *et al.* (2017) Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. *Breast Cancer Res*. 19:1-12. https://doi.org/10.1186/s13058-017-0911-9

54. Sestak I, Buus R, Cuzick J, *et al.* (2018) Comparison of the performance of 6 prognostic signatures for estrogen receptor–positive breast cancer: a secondary analysis of a randomized clinical trial. *JAMA Oncol* 4:545-553. https://doi:10.1001/jamaoncol.2017.5524

55. Ohnstad H, Borgen E, Mortensen E, *et al.* (2023) 103P Impact of Prosigna test on treatment decision in lymph node-negative early breast cancer: A prospective multicenter study (EMIT1). ESMO Open Abstract. *ESMO Open* 8:7. https://doi.org/10.1016/j.esmoop.2023.101327

The Ki67 Dilemma: Investigating Prognostic Cut-Offs and Inter-Platform Reproducibility for Automated Ki67 Scoring in Breast Cancer.
Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

Supplementary Information

Supplementary Table S1

Supplementary table S1: Identification and version number for each Visiopharm® AI Ki67 Application

Number	Identification	Version
01	#10182 – IHC Tissue Detection	2022.09.0.12236
02	#10180 – Invasive Tumor Detection	2022.07.0.12224
03	#10180 – Invasive Tumor Postprocessing	2022.07.0.11960
04	#10173 – Ki-67 Nuclei APP, Breast Cancer, AI	2022.07.0.11960
05	#10114 Hot Spot Detection	2022.09.0.12417
06	#10114 Hot Spot Quantification	2022.09.0.12417

Supplementary Table S2

Supplementary table S2: The Ki67 nuclei classifier script developed in QuPath.

```
setImageType('BRIGHTFIELD_H_DAB');
setColorDeconvolutionStains('{"Name" : "H-DAB default", "Stain 1" : "Hematoxylin", "Values 1" : "0.65111 0.70119 0.29049 ", "Stain 2" : "DAB", "Values 2" : "0.26917 0.56824 0.77759 ", "Background" : " 255 255 255 "}');
runPlugin('qupath.imagej.detect.cells.PositiveCellDetection', '{"detectionImageBrightfield": "Optical density sum",  "requestedPixelSizeMicrons": 1.0,  "backgroundRadiusMicrons": 10.0,  "medianRadiusMicrons": 0.1,  "sigmaMicrons": 1.5,  "minAreaMicrons": 15.0,  "maxAreaMicrons": 400.0,  "threshold": 0.1,  "maxBackground": 2.0,  "watershedPostProcess": true,  "excludeDAB": false,  "cellExpansionMicrons": 5.0,  "includeNuclei": true,  "smoothBoundaries": true,  "makeMeasurements": true,  "thresholdCompartment": "Nucleus: DAB OD mean",  "thresholdPositive1": 0.2,  "thresholdPositive2": 0.4,  "thresholdPositive3": 0.6000000000000001,  "singleThreshold": true}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": false}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": true}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": false}');
runPlugin('qupath.lib.plugins.objects.SmoothFeaturesPlugin', '{"fwhmMicrons": 25.0,  "smoothWithinClasses": false}');
```

Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

Supplementary Table S3

Supplementary Table S3: Overview of patient characteristics including age, follow-up time, tumor size, Nottingham Grade, treatment, and operation type for the 294 cases eligible for conventional and digital analysis of Ki67.

Characteristics	**N=**				**All**	**No DM**	**DM**
	All	*No DM*	*DM*				
Age Mean (Range)	294	254	40		56 (29-70)	56 (32-70)	53 (29-70)
Follow-up time (months)* Mean (range)	294	254	40		161 (8-345)	173 (9-345)	88 (8-302)
Tumor size**	292	252	40		1.55 (0.2-6.5)	1.49 (0.2-4.2)	1.92 (0.7-6.5)
Nottingham Grade***	292	252	40	I	102 (35%)	96 (38%)	6 (15%)
				II	138 (47%)	118 (47%)	20 (50%)
				III	52 (18%)	38 (15%)	14 (35%)
Adjuvant Systemic Therapy	202	167	35	NTR[a]	187 (93%)	158 (95%)	29 (83%)
				TR[b]	15 (7%)	9 (26%)	6 (17%)
Operation Type****	231	195	36	Conservative/ Lumpectomy	139 (60%)	124 (64%)	15 (42%)
				Mastectomy	92 (40%)	71 (36%)	21 (57%)

**Significant difference between no DM and DM groups (Mann-Whitney U, p<0.001)*
***Significant difference between no DM and DM groups (Mann-Whitney U, p=0.007)*
****Significant difference between no DM and DM groups (Kruskall Wallis, p<0.001)*
*****Significant difference between no DM and DM groups (Kruskall Wallis, p=0.014)*
[a]NTR: No Treatment
[b]TR: Treatment: Treatment types included chemotherapy, HRT or chemotherapy combined with HRT

Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

Supplementary Table S4

Supplementary table S4: Overview of the ROI specifications tested for the VIS1-HS method and resulting mean total tumor cell count and Ki67 score.

Specifications (VIS1)			N=	Mean total tumor cell count	%Ki67		
Drawing radius	Size of ROI	Number of ROIs			Mean	Median	Range
175µm	0.2mm^2	1	98	705	18.0	13.8	0.8 – 87.8
400µm	0.2mm^2	1	98	704	16.7	10.7	1.0 – 76.9
175µm	1mm^2	1	98	2849	15.7	10.6	1.0 – 77.4
400µm	1mm^2	1	98	3360	16.0	10.9	0.4 – 81.1
400µm	0.2mm^2	**2**	27	782	22.3	16.1	4.7 – 68.2
400µm	0.2mm^2	**3**	27	752	22.0	17.1	4.8 – 64.6
400µm	0.2mm^2	**4**	26	753	20.2	16.4	4.4 – 62.4
400µm	0.2mm^2	**5**	26	753	20.3	15.6	4.5 – 60.1
400µm	**550 TC min**	1	27	692	23.2	17.5	4.2 – 74.9
400µm	**1000 TC min**	1	27	1119	23.0	16.2	3.7 – 76.2

These specifications were used for the main investigations of the study.

Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

Supplementary Table S5

Supplementary table S5: Performance of tested cutoffs for binary categorization of Ki67 score as assessed by manual and DIA scoring methods.

Method (Ki67 cut-off)	Log-rank (p-value)	N=					
		False negative	False positive	True negative	True positive	Sens.	Spec.
Conventional							
3.5%	0.004	4	153	75	35	88%	32%
6.5%	0.001	9	114	114	30	76%	49%
10.6%	0.001	14	83	145	25	63%	64%
14%	**0.026**	**20**	**67**	**161**	**19**	**49%**	**71%**
20%	0.023	25	46	182	14	36%	80%
30%	0.004	29	24	204	10	26%	89%
48%	0.253	35	11	217	4	10%	95%
Global UW							
2%	0.068	4	195	59	36	92%	23%
8%	0.031	12	134	120	28	73%	49%
10%	0.016	13	121	133	27	70%	53%
14%	**0.021**	**19**	**86**	**168**	**21**	**51%**	**68%**
20%	0.317	26	68	186	14	38%	74%
31%	0.087	29	40	214	11	30%	84%
40%	0.011	32	21	233	8	22%	92%
Global W							
2%	0.110	5	193	61	35	89%	24%
5%	0.069	10	153	101	30	76%	40%
10%	0.007	13	113	141	27	70%	56%
14%	**0.020**	**19**	**86**	**168**	**21**	**57%**	**67%**
20%	0.081	25	61	193	15	41%	77%
30%	0.198	30	41	213	10	27%	84%
40%	0.008	31	23	231	9	24%	90%
VIS1-HS							
4%	0.022	2	198	56	38	95%	22%
8%	0.001	9	126	128	31	78%	50%
10%	0.001	12	103	151	28	70%	59%
14%	**0.000**	**18**	**67**	**187**	**22**	**55%**	**74%**
20%	0.000	23	46	208	17	43%	82%
30%	0.002	31	21	233	9	23%	92%
41%	0.858	38	12	242	2	5%	95%
VIS2-HS							
3.4%	0.005	1	195	59	39	98%	23%
6%	0.001	5	149	105	35	88%	41%
10%	0.000	11	106	148	29	73%	58%
14%	**0.000**	**16**	**75**	**179**	**24**	**60%**	**70%**
20%	0.007	23	57	197	17	43%	78%
30%	0.001	29	27	227	11	28%	89%
41%	0.136	35	15	239	5	13%	94%
VIS2-G							
2.3%	0.027	5	178	76	35	81%	28%
7.6%	0.001	15	86	168	25	62%	68%
10%	0.001	19	67	187	21	46%	75%
14%	**0.000**	**22**	**48**	**206**	**18**	**45%**	**81%**
20%	0.010	30	28	226	10	25%	89%
42%	0.190	37	8	246	3	12%	98%
QuPath							
14%	**0.000**	**12**	**82**	**163**	**28**	**70%**	**67%**
20%	0.000	19	62	183	21	53%	75%
30%	0.000	24	33	212	16	40%	87%

Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

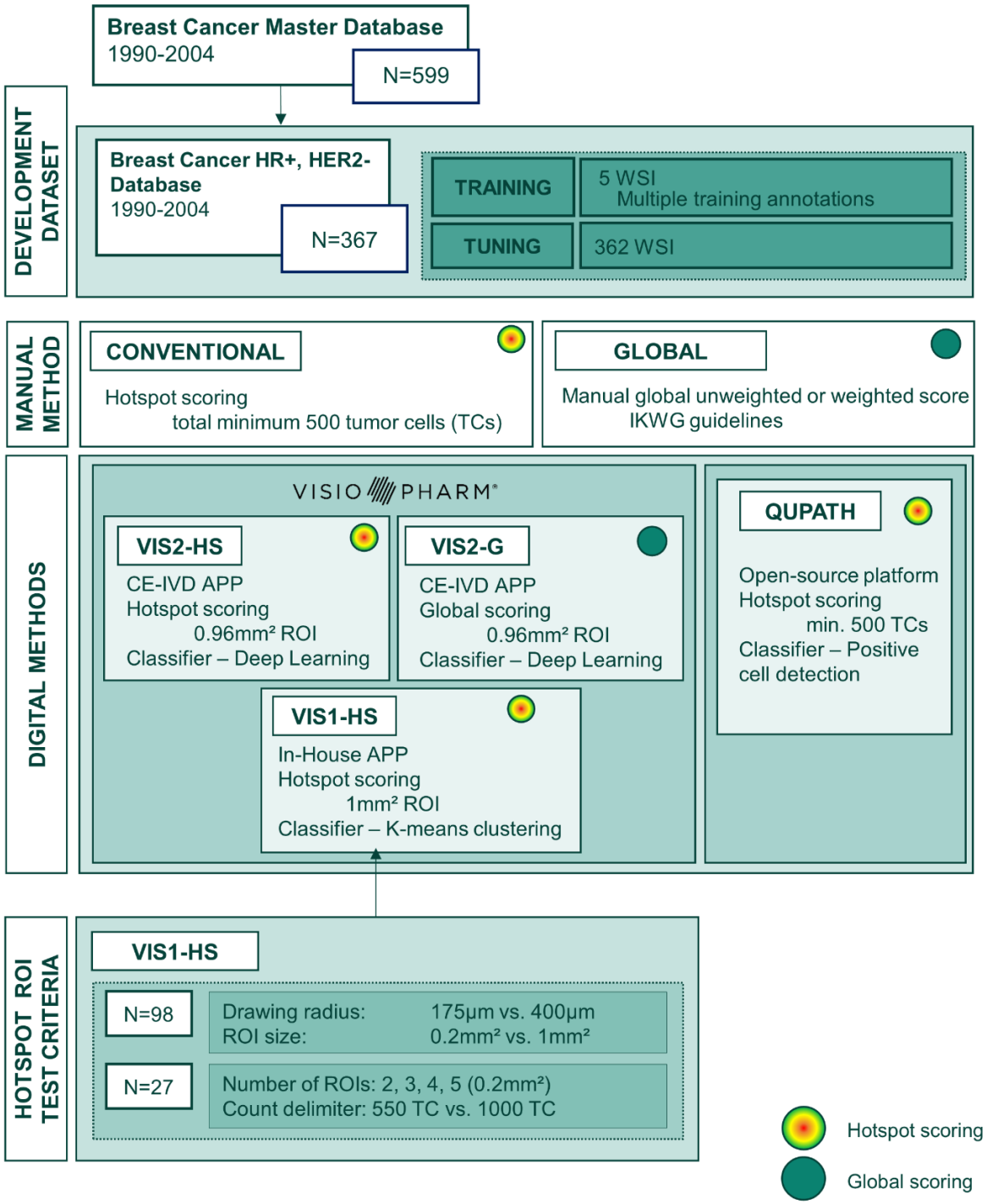
Supplementary Table S6

Supplementary table S6: Percentage agreement of Ki67 quantification methods using a 14% cutoff.

Method	**Measure**	**Method**					
		Conventional	*Global uW*	*Global W*	*VIS1-HS*	*VIS2-HS*	*VIS2-G*
Conventional	% Agreement						
	% Disagreement						
Global uW	% Agreement	87%					
	% Disagreement	13%					
Global W	% Agreement	85%					
	% Disagreement	15%					
VIS1-HS	% Agreement	91%	86%	84%			
	% Disagreement	9%	14%	16%			
VIS2-HS	% Agreement	91%	89%	88%	95%		
	% Disagreement	9%	11%	12%	5%		
VIS2-G	% Agreement	89%	84%	84%	89%	88%	
	% Disagreement	11%	16%	16%	11%	12%	
QuPath	% Agreement	87%	87%	86%	89%	92%	83%
	% Disagreement	13%	13%	14%	11%	8%	17%

Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

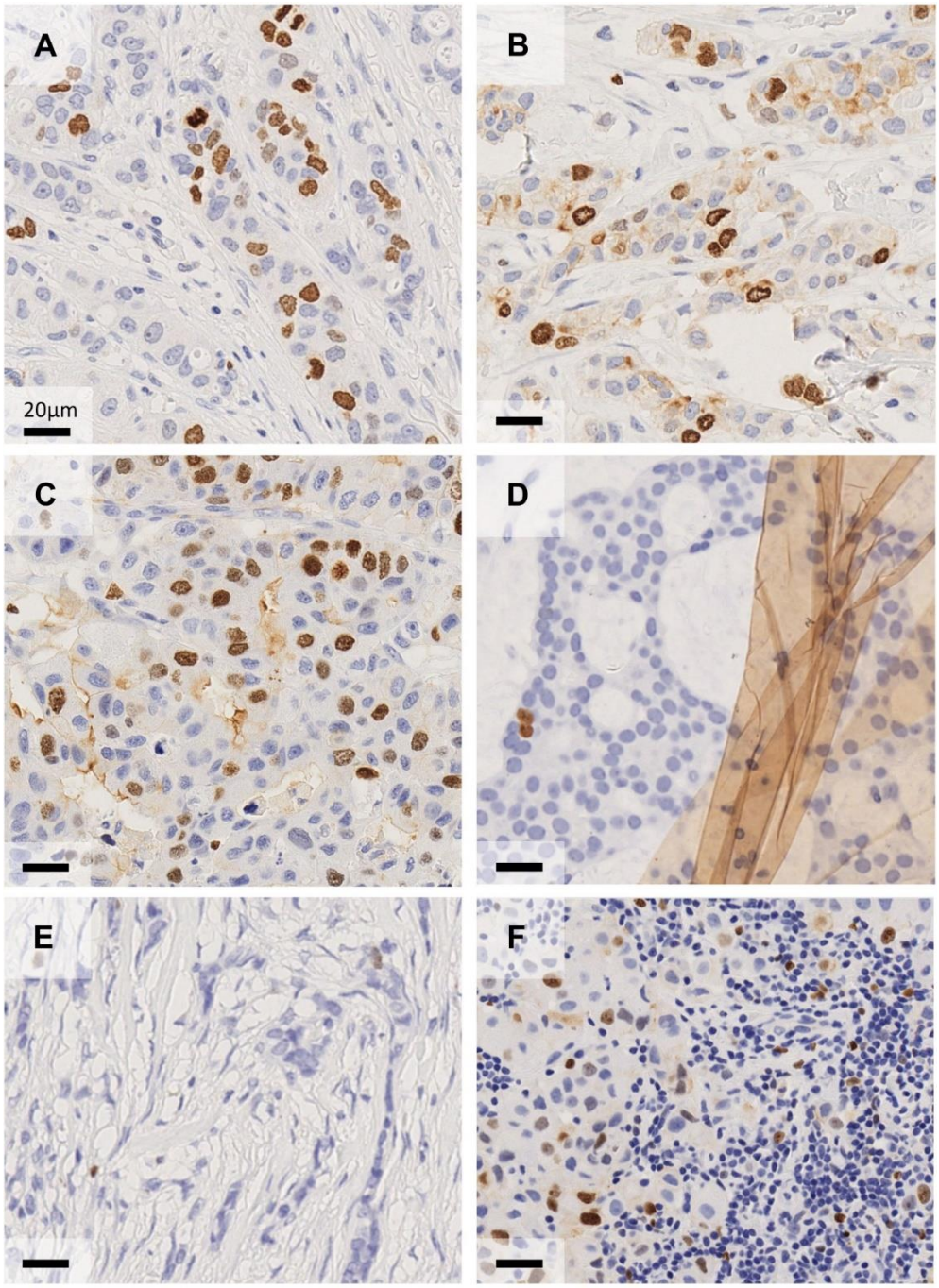
Supplementary Figure S1



Supplementary Figure S1: Case summary of the study cohort. A development dataset was used to train and tune two in-house Ki67 scoring algorithms using a commercial platform (VIS1-HS) and an open-source platform (QuPath). Cases from the development dataset were also evaluated using a using two manual methods (conventional hotspot Ki67 scoring and global unweighted and weighted scoring) and a commercial CE-IVD application (VIS2-HS/G). Specifications for each method is outlined. A variety of parameters was investigated for specification of hotspot ROI criteria.

Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

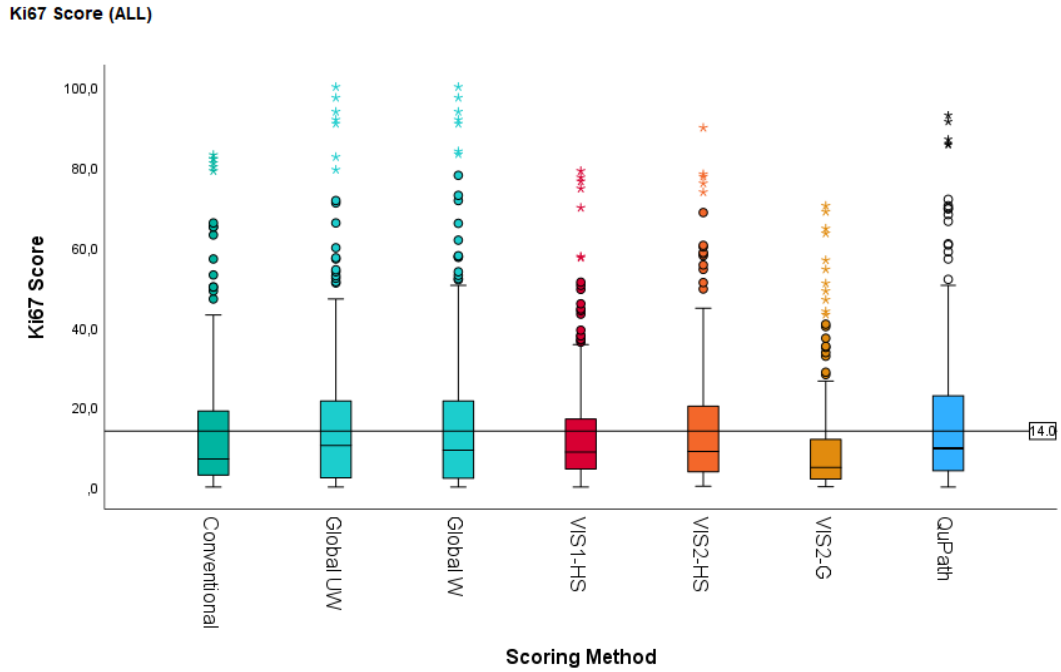
Supplementary Figure S2



Supplementary figure S2: Examples of optimal and poor quality Ki67 staining patterns and artefacts. A) Optimal Ki67 staining for clear distinction of negative (blue) nuclei and positive (brown) nuclei. B-C) Background staining observed in the cytoplasm and a staining edge effect observed between tumor cell clusters. This affected analysis to varying degrees depending on the intensity of the background stain, software used and area coverage (smaller areas could be excluded from analysis). A case was rejected when background staining resulted in an unacceptable number of false positive nuclei detections. D) Example of an artefact that required manual removal due to the negative (blue nuclei) beneath the fragment being labelled as positive by all digital methods. E) An example of a poor quality tissue section likely due to poor fixation. F) A region where a high number of tumor infiltrating lymphocytes are present amongst tumor cells. These regions could cause the analysis to be rejected or were manually excluded from analysis.

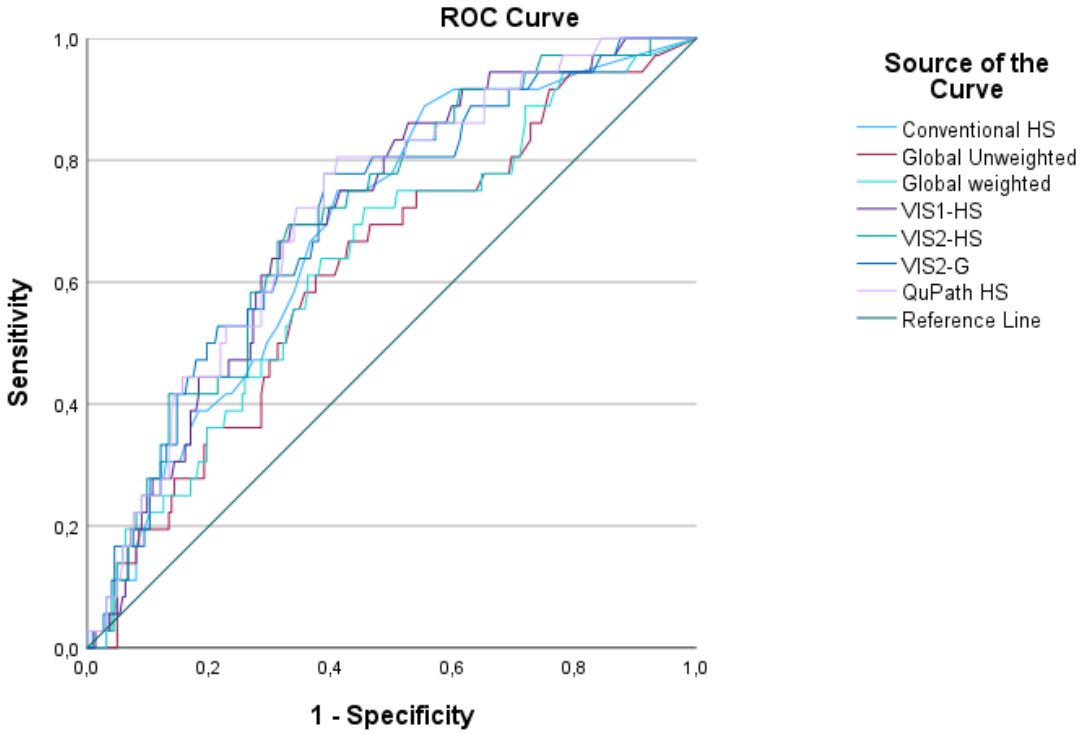
Emma Rewcastle[a,b], Ivar Skaland[b], Einar Gudlaugsson[b], Silja Kavlie Fykse[b], Jan P.A. Baak[b], Emiel A.M. Janssen[a,b]*
[a]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway. [b]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.
**Corresponding author. E-mail address: emma.rewcastle@sus.no.*

Supplementary Figure S3



Supplementary Figure S3: The distribution of Ki67 score for all manual (conventional-HS, global unweighted-UW and weighted-W) and automated (VIS1-HS, VIS2-HS, VIS2-G, QuPath) scoring methods. The line intersecting all box plots indicates 14%.

Supplementary Figure S4



Supplementary figure S4: ROC curves for manual (conventional HS, global weighted, global unweighted) and digital (VIS1-HS, VIS2-HS, VIS2-G, QuPath HS) Ki67 scoring methods.

PAPER IV

Applicability of mitotic figure counting by deep learning: a development and pan-cancer validation study

Tarjei S. Hveem*, Maria X. Isaksen*, Joakim Kalsnes*, Frida Julbø, Manohar Pradhan, Andreas Kleppe, Sepp De Raedt, Ole-Johan Skrede, Turid Torheim, John Arne Nesheim, Hans Martin Mohn, Hanne A. Askautrud, Karolina Cyll, Wanja Kildal, Emma Rewcastle, Melinda Lillesand, Vebjørn Kvikstad, Emiel Janssen, Robert Jones, Odd Terje Brustugun, Bjørn Brennhovd, Erik Skaaheim Haug, Lill-Tove Rasmussen Busund, Elin Richardsen, Sigve Andersen, Tom Dønnem, Kristina Lindemann, Gunnar Kristensen, Neil A. Shepherd, Marco Novelli, Knut Liestøl, David Kerr, Håvard E. Danielsen

*Contributed equally

Corresponding author: Tarjei S. Hveem

Manuscript submitted. This paper is not included in the repository because it is not yet published.