Aus dem Institut für Pathologie
(Direktor: Professor Dr. med. P. Ströbel)
der Medizinischen Fakultät der Universität Göttingen

Image analysis of immunohistochemistry-based biomarkers in breast cancer

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades
für Zahnmedizin
der Medizinischen Fakultät der
Georg-August-Universität zu Göttingen

vorgelegt von
Judith Burchhardt
aus Warburg

Göttingen 2021

Dekan: Prof. Dr. med. W. Brück

Betreuungsausschuss:

Betreuer: PD Dr. med. P. Middel

Ko-Betreuer: Prof. Dr. med. G. Emons

Prüfungskommission:

Referent: PD Dr. med. P. Middel

Ko-Referent: Prof. Dr. med. G. Emons

Drittreferentin: PD Dr. Sabine Sennhenn-Kirchner

Datum der mündlichen Prüfung: 21.11.2022

Hiermit erkläre ich, die Dissertation mit dem Titel "Image analysis of immunohistochemistry-based biomarkers in breast cancer" eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Bonn, den

Teile der Daten, auf denen die vorliegende Arbeit basiert, wurden nach dem Peer-Review Verfahren in der Fachzeitschrift *Diagnostic Pathology* veröffentlicht:

Scheel AH, Penault-Llorca F, Hanna W, Baretton G, Middel P, **Burchhardt J**, Hofmann M, Jasani B, Rüschoff J (2018): *Physical basis of the 'magnification rule' for standardized Immunohistochemical scoring of HER2 in breast and gastric cancer.* Diagn Pathol 13(1),19

Table of contents

Index of figures

Index of tables

Index of supplementary information

Index of abbreviations

ADCC	Antibody dependent cytotoxicity
ASCO/CAP	American Society of Clinical Oncology / College of American Pathologists
AWMF	*Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften* (English: German association of the scientific medical societies)
CIS	Carcinoma in situ
csv	Comma seperated values
DAB	3,3'-diaminobenzidine
DAkkS	*Deutsche Akkreditierungsstelle GmbH* (English: German national institute of accreditation)
DDISH	Dual colour in situ hybridisation
DIA	Digital image analysis
ER	Estrogen receptor
Fc	Fragment crystalisable
FISH	Fluorescence in situ hybridisation
FOV	Field-of-view
H	Haematoxylin
H&E	Haematoxylin and eosin stain
h-score	Histoscore
Her2	The human epidermal growth factor receptor 2 protein
HER2/neu	The *HER2/neu* gene
IHC	Immunohistochemistry
ISH	In situ hybridisation
κ	Cohen's kappa coefficient of concordance
Ki-67	Antigen Ki-67
MR	Magnification rule
NA	Numerical aperture
PR	Progesterone receptor
r	Pearson's coefficient of correlation
RKI	*Robert Koch Institut*
ROI	Region-of-interest
SERM	Selective estrogen receptor modulator
SERD	Selective estrogen receptor degrader
T-DM1	Trastuzumab emtansine

TDLU	Terminal duct lobular unit
TIL	Tumour infiltrating lymphocyte
TMA	Tissue microarray
TNM	Tumour, [lymph] nodes, metastases classification system
WHO	World Health Organization

1 Introduction

1.1 Breast cancer

1.1.1 Summary

Breast cancer is an umbrella term for malignant neoplasms arising from the mammary gland. In western developed countries it is the most frequent malignant disease in women with a lifetime risk of 12.9% (RKI 2015). The most common type are carcinomas arising from the terminal duct lobular unit (TDLU) of the mammary gland (Fletcher 2020). Early detection of breast cancer and management of breast cancer treatment are important, interdisciplinary medical tasks. Definite diagnosis is based on biopsies and histopathology (*Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften* (AWMF) 2020). The analysis of biomarkers facilitates clinical decision making by providing information on the expected course of disease and on the response to systemic treatment.

1.1.2 Classification

There are several systems for the classification of lesions of the breast. The employed system depends on performed clinical diagnostic procedure. As recommended by current clinical guidelines, the following classification systems are used:

Radiology of breast, i.e. mammography, is classified by the Breast Imaging Reporting and Data System (BI-RADS). The system was created by the American College of Radiology and classifies the findings into six categories. The categories represent different likelihoods for malignancy. Categories 4 and 5 require biopsy and histopathology. Mammographies with a BI-RADS 5 finding have a > 95% likelihood for malignancy (Table 1.1) (AWMF 2020).

Biopsies of the breast are reported by a similar classification system that features five categories (Table 1.1) (Royal College of Pathologists Working Group 1991). Category 5 are malignant neoplasms that can be non-invasive (B5a) or invasive (B5b).

Staging is classified by the TNM-system that incorporates the extent of the primary tumour (T), the presence or absence of lymph node metastases (N) and of metastases (M). Clinical staging is tagged with the prefix c, pathologic staging of resection specimens with the prefix p. Staging after neoadjuvant treatment has the additional prefix y (Wittekind 2020).

Histologically, carcinomas of the breast can be subdivided into invasive carcinoma and pre-invasive carcinoma in situ (CIS). Invasive carcinoma has breached the basement membrane and infiltrates the surrounding tissue. It may cause metastasis to local lymph nodes and distant organs including the lungs, bones and brain. CIS is regarded as a precursor lesion to invasive carcinoma.

Table 1.1: Classification systems of mammography and breast biopsies

Category	BI-RADS (Mammography)	B-Classification (Biopsy)
B0	Incomplete	
B1	Negative	Non-diagnostic; exclusive normal tissue
B2	Benign	Benign lesion
B3	Probably benign	Lesion with unkown biological potential
B4	Suspicious	Suspicious for malignancy but limited validity
B5	Highly suggestiv of malignancy	Malignant, sub-categories: B5a: non-invasive carcinoma B5b: invasive carcinoma
B6	Known biopsy, proven malignancy	

Table 1.1: Mammographies are classified into seven categories (BI-RADS) that reflects the likelihood of malignancy. Breast biopsies use a five-step system with related categories (B-Classification).

It respects the basement membrane and does not infiltrate into the surrounding tissue. CIS can be found in association with invasive carcinoma or on its own.

Invasive carcinoma and CIS can be subclassified based on morphology. There are specific subtypes and the common type, which is designated non-specific type (NST). In older nomenclatures, invasive carcinoma NST was termed invasive ductal carcinoma. Among the specific subtypes, lobular neoplasms that lack expression of the cell-adhesion molecule E-cadherin are the most frequent type. Other types include tubular, medullary, mucinous, apocrine and metaplastic carcinoma as well as neuroendocrine carcinoma and several other, rare subtypes (World Health Organization (WHO) 2019). Most carcinomas arise from the TDLU, the functional unit of the mammary gland. The morphologic appearance is related to differences in the genetic alterations driving the disease; as such it likely does not reflect different cells of origin (Kumar et al. 2015).

Malignant carcinomas are the most common type of breast cancer, but tumours of mesenchymal cells of the associated stroma, connective tissue and vessels also occur. Phylloides tumours arise from periductal stromal cells and may be benign, borderline or malignant. Angiosarcoma is a malignant soft tissue tumour of endothelial cells of blood or lymphatic vessels. While generally infrequent, the risk increases after radiotherapy of the breast. Various other forms of rare soft tissue malignancies exist including liposarcomas, myosarcomas and even osteosarcomas (WHO 2019). Besides malignant and pre-malignant lesions, different forms of benign mammary changes and lesions exist. These include fibrocystic changes, inflammatory and sclerosing lesions, hyperplasias and benign neoplasms such as fibroadenoma and papilloma. Benign epithelial lesions

usually do not cause symptoms, but are frequently detected in mammography and biopsies (Kumar et al. 2015).

For biomarker analyses, it is essential to distinguish invasive and pre-invasive neoplasms as well as benign lesions. In analyses of human epidermal growth factor receptor 2 (Her2), pre-invasive DCIS often shows strong staining intensity and amplification of the *HER2/neu* gene, but unlike invasive carcinoma it does not have predictive value for anti-Her2 treatment.

The expression of estrogen receptor (ER) and Her2 receptor divides carcinomas of the breast into distinct groups that differ in biological behaviour, response to clinical treatment and prognosis. The receptors are most commonly tested by immunohistochemistry and summarised as receptor status (cf. 1.3). The hormone receptor positive group expresses ER but not Her2, the Her2 positive group expresses Her2 with or without ER and the triple negative group neither expresses ER, Her2 nor progesterone receptor (PR).

1.1.3 Pathogenesis

Cancer is a disease of the genes; changes to the DNA cause changes in gene expression patterns that mediate the hallmarks of malignant diseases: Uncontrolled and self-sustained cell divisions, evasion of apoptosis, invasive growth into surrounding tissues, invasion into lymphatic and blood vessels and the formation of metastases in lymph nodes and distant organs. (Kumar et al. 2015; Weinberg 2014). Malignant tumours are proliferations that arise from cells harbouring such alterations. They can be acquired or inherited by susceptibility genes. The mammary gland is a hormone-responsive organ and pathogenesis of its tumours is influenced by hormonal exposures: Hormonal therapy of menopausal symptoms is associated with increased breast cancer risk. Resection of the ovaries, as major organs of estrogen production, strongly reduces breast cancer risk. Clinical treatments that either block ER signalling or the production of estrogen decrease breast cancer risk. On the other hand, oral contraceptives do not seem to increase breast cancer risk (Kumar et al. 2015).

The major risk factors for breast cancer are female sex, hereditary susceptibility genes, lifetime estrogen exposure and age. Various other factors related to environment and lifestyle have been identified but their impact seems limited. About 99% of cases occur in females. Breast cancer in males may occur and is often linked to hereditary factors. Breast cancer risk increases with age and peaks at 70 to 80 years. Germline mutations in susceptibility genes cause 5-10% of cases. They are tumour suppressor genes and mutations cause hereditary syndromes that strongly increase the risk for cancer in different organs. Among the most relevant affect genes are *BRCA1* and *BRCA2* (familial breast and ovary cancer), *CHEK2* and *PALB2* (AWMF 2020; Kumar et al. 2015),

which are involved in DNA repair. BRCA-deficient tumours are amenable to specific treatment with PARP-inhibitors, which also modulate DNA repair (Lord and Ashworth 2017).

The genetic background of breast cancer is heterogeneous and may involve acquired mutations in different types of tumour suppressor genes and oncogenes. The acquisition of genetic alterations is thought to be a stepwise process leading to the formation of precursor lesions that give rise to invasive carcinoma. Indeed, most mutations in commonly affected genes are already found in CIS of the mammary gland. The progression of CIS into invasive carcinoma is not yet fully understood (Kumar et al. 2015). Genetic heterogeneity may cause resistance to clinical treatment. Non-responsive subclones may outlast treatment and give raise to recurring disease.

1.2 Treatment

Clinical treatment of breast cancer is an interdisciplinary task. The treatment modalities are selected on the extent of the disease, i.e. clinical staging, the histopathology of the biopsy, the receptor status as well as patients' characteristics according to current clinical guidelines (AWMF 2020). Surgical excision with the goal of complete tumour removal is the principle treatment in early breast cancer that has not spread to other organs. Sampling of local lymph nodes may be conducted by the sentinel node procedure: The hypothetical first node that drains lymph from the tumour region is tested during surgery by frozen section histology. Systemic cytotoxic chemotherapy may be used to reduce the risk of recurrent disease after surgery in adjuvant treatment, to reduce tumour burden prior to surgery in neoadjuvant treatment, or to improve overall survival in metastatic disease. Radiation therapy may be used as adjuvant treatment with or without systemic therapy. HR-positive tumours are sensitive to anti-estrogen treatments, while Her2-positive tumours are sensitive to specific Her2 antagonists. The palette of treatment options is further expanded by treatments with molecular-defined targets such as inhibitors of cyclin-dependent kinase and treatments restoring the anti-tumoural immune response such as PD-L1 inhibitors.

An effective measure is early breast cancer detection: Regular self-examination based on established procedures and clinical screening by mammography increase the chance to detect a tumour in an early stage that is curable by surgery and reduces breast cancer mortality (AWMF 2020).

1.3 Biomarkers

1.3.1 Definitions

Biomarkers are indicators that can be measured and provide clinically relevant information on a certain disease. Biomarkers may be measured in blood samples, during the work-up of tissue

samples, by radiologic methods or by other means. There are three kinds of biomarkers that are clinically employed (Badve and Kumar 2018):

Diagnostic biomarkers provide information if a patient is afflicted by a certain disease and/or on the classification of a disease. In histopathology, the investigation of proteins by immunohistochemistry (IHC) can provide information that may supplement histomorphology. Certain proteins are expressed in malignant neoplasm but not their benign counterparts. Other proteins are tissue-specific and may give clues on the origins of a metastatic malignant tumour.

Prognostic biomarkers are indicative of the expected course of disease. In malignant tumours, the expression of certain proteins or the presence of specific DNA mutations is associated with better or worse prognosis.

Predictive biomarkers provide information about the expected clinical benefit from a certain treatment. The best example is targeted therapy of malignant diseases: Kinase inhibitors are used as clinical treatment that are specific to a certain DNA mutation and do not affect the physiological, wild type version of the respective kinase. The DNA mutation is therefore predictive of treatment benefit with such an inhibitor.

1.3.2 Immunohistochemistry

A technique commonly used to quantify biomarkers in histopathology is immunohistochemistry (IHC) (Dabbs 2018). Specific antibodies are used to bind a target protein in a histological tissue section. The bound antibodies are visualised by the enzymatic reaction that creates a dye (Figure 1.1 A).

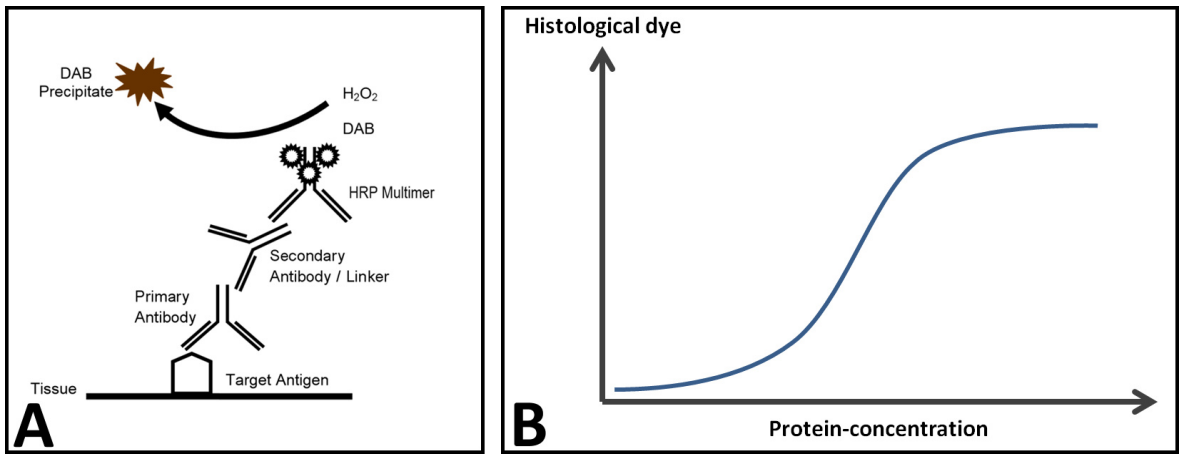


Figure 1.1: Immunohistochemistry is a histopathological technique to detect a specific protein in a tissue section. (A) schematic overview, the protein is bound by a specific antibody, the bound antibody is detected by an enzymatic reaction that creates a dye such as DAB. (B) Illustration of the relation between protein concentration and resulting dye. Due to the enzymatic detection system, the relation is non-linear.

The antibody that binds the target protein is called primary antibody. It can be a monoclonal antibody that binds just one epitope or a polyclonal mix of antibodies that bind several epitopes. In clinical biomarker testing, only monoclonal antibodies are used. They are usually produced by the hybridoma method in which a plasma cell is fused with a neoplastic myeloma cell. The bound epitope can be localised anywhere in the cell. Common subcellular locations include the cell membrane, the cytoplasm and the nucleus.

Most IHC protocols include several steps to amplify the signal, which increases sensitivity. The bound primary antibody can be detected by a secondary antibody that binds to the Fc fragment. Secondary antibodies are usually directed against all primary antibodies derived from one type of animal, e.g. anti-mouse or anti-rabbit and do not bind human antibodies. The secondary antibody may be detected by a third molecule or it may be conjugated to a polymer. The polymer is a large multi-molecular structure that includes several secondary antibodies and several enzymes which produce the dye. Polymer based detection system have become standard for most diagnostic purposes (Dabbs 2018). A commonly used enzyme is horseradish peroxidase (HRP). It can catalyse several chromogenic reactions that produce the dye. The currently most popular dye is 3,3'-diaminobenzidine (DAB); when oxidised by HRP, it forms a brownish, insoluble dye that precipitates at the site of the bound antibodies. Larger precipitates turn to a dark, black appearing colour (Figures 1.1, 2.2, 2.3, 3.8).

It should be noted that the relation between the protein concentration in the tissue and the resulting dye is not linear. Due to the enzymatic reaction involved, the relation follows a sigmoidal curve (Figure 1.1 B). Once a certain threshold is exceeded, visible dye is created. The curve features a linear section, before saturation is reached. In the protein concentration range that corresponds to the linear section, meaningful differences in the resulting dye can be observed. Thus, important characteristics of an IHC assay are the lower threshold, i.e. the lower limit of detection, the slope of the linear section, and the upper limit before saturation occurs.

1.3.3 Biomarkers in breast cancer

1.3.3.1 Overview

All three classes of biomarkers are employed in the routine pathological work-up of breast cancer samples:

Diagnostic markers include the transcription factor Gata3 and cell adhesion molecule E-cadherin which can be detected by IHC. Gata3 is frequently overexpressed in carcinomas of the breast and of the urinary bladder. In combination with other markers, Gata3 can be used to confirm tissue of origin of a carcinoma. E-cadherin is used in combination with histomorphology to identify the

lobular subtype of breast cancer since loss of E-cadherin expression is common in lobular neoplasms and is associated with the typical dispersed growth of the malignant cells.

The expression of the hormone receptors (HR) ER and PR as well as the Her2 receptor are both prognostic and predictive. Both hormone receptors and Her2 are most commonly quantified by immunohistochemistry. Her2 can also be tested by hybridisation of the *Her2/neu* gene. Carcinomas that are HR positive, carcinomas that are positive for Her2 and carcinomas negative for all of these receptors, so-called triple negative carcinomas, differ in their prognosis. HR positive carcinomas have a more favourable prognosis, while Her2 positive and especially triple negative carcinomas have adverse prognoses with more rapid progression and higher risk of metastases. HR positive carcinomas are amenable to antihormone therapy while Her2 positive carcinomas are responsive to treatments that specifically target and inactivate the Her2 receptor. Thus, the receptor status is indicative of the expected course of disease and of the potential benefit of different systemic treatments.

The most important prognostic factor is the pN status, i.e. the presence or absence of lymph node metastasis (Cianfrocca and Goldstein 2004). The number of lymph nodes affected is directly correlated with the disease free interval and the overall survival rate (Tumorzentrum München 2017). The pN status is included in clinical decision-making concerning adjuvant chemotherapy. In nodal negative patients, additional prognostic factors are employed: Ki-67 is a nuclear protein associated with proliferation. It may provide additional information about the risk of recurrent disease. Other prognostic biomarkers are also surrogate parameters for cellular proliferation. The most traditional form is the counting of mitotic figures on hematoxylin and eosin (H&E) stained histological sections. Newer approaches are gene expression analysis assays, which analyse the RNA levels of several genes and interpret the result into a clinical meaningful score (Paik et al. 2021; Robertson et al. 2020).

1.3.3.2 Hormone receptors

Hormone receptors are cellular proteins that bind sex hormones and act as transcription factors. There are three types specific for estrogen, progesterone and testosterone. Currently, only ER and PR and clinically relevant in breast cancer. Androgen receptors that bind testosterone may be expressed in some subtypes of breast cancer, but are currently not part of routine histopathologic work-up. Upon ligand binding, HRs form dimers, translocate into the nucleus and act as transcription factors by binding to the DNA and activating gene expression.

ER and PR are integral part of sexual maturation, the menstrual cycle and gestation. ER has two isoforms called α and β which have similar structures and functions and are preferentially

expressed in different tissue: ER α is expressed in the mammary gland, endometrium, ovary as well as certain brain regions, ER β is expressed in several organs including the bones.

Increased levels of estrogen signalling are associated with increased risk of breast and endometrial cancer. Adipose tissue may express the enzyme aromatase, a key enzyme in the production of estrogen which catalyses the aromatisation of androgens. Correspondingly, obesity is a risk factor for breast and endometrial cancer (Keum et al. 2015). Chemicals that act as endocrine disruptors such as Bisphenol-A may increase the risk for HR-dependent malignancies.

Antihormonal therapy (AHT) is based on substances that disrupt estrogen receptor signalling in cancer cells. It is used in ER positive, Her2 negative carcinomas. In early breast cancer AHT, can be used as adjuvant treatment after surgery to reduce the risk of recurrent disease and lower the risk of death (Davies et al. 2011). In advanced breast cancer, AHT increases progression free survival and overall survival (Robertson et al. 2021). Pharmacological strategies include selective estrogen receptor modulators (SERMs), selective estrogen receptor degraders (SERD) and aromatase inhibitors. SERMs such as Tamoxifen and Raloxifen compete with estrogen for receptor binding and have mixed agonistic and antagonistic effects that depend on the type of tissue (Gottardis et al. 1988). In the breast, they act antagonistic and inhibit tumour cell growth; in other tissues including bones they act agonistic which avoids osteoporosis. SERDs bind to ER and cause increased protein degradation. Aromatase inhibitors such as Anastrozole and Letrozole block the enzyme aromatase that produces estrogen and thus reduce ER activation in all tissues.

HRs are usually quantified by IHC during histopathologic work-up of biopsies and resection specimens. Given the subcellular localisation of the receptors, they show a nuclear staining. The staining is interpreted by applying a semiquantitative score. Frequently used systems are the Allred-Score (Harvey et al. 1999) and the Immunoreactivity-Score (Dietel and Klöppel 2013). Both scores use a combination of staining intensity and proportion of stained cells.

1.3.3.3 Her2

Her2 is a membranous growth factor receptor and belongs to the ErbB-family of receptors which also includes EGFR, Her3 and Her4. ErbB proteins are receptor tyrosine kinases and activate cellular proteins by transferring phosphor-groups. Upon ligand binding, Her2 forms a dimer, autophosphorylates and activates a variety of intracellular signalling pathways. The ErbB-family is involved in several types of cancer, e.g. activating mutations of EGFR are common in non-small cell lung cancer. In a proportion of breast cancers, Her2 is aberrantly activated by gene amplification and subsequent protein overexpression. Consequently, Her2 positive cancers can be identified by detection of the gene amplification which is most commonly done by using molecular, sequence-specific probes that are hybridised to the DNA (in situ hybridisation, ISH).

The probe can be detected by fluorescent dyes (FISH) or brightfield techniques (DISH). Alternatively, Her2 positivity can be identified by quantifying the Her2 protein using IHC (Dietel and Klöppel 2013).

Anti-Her2 treatment of Her2 positive breast cancer as adjuvant treatment after surgery reduces the risk of recurrent disease and improves survival (Cameron et al. 2017). Anti-Her2 treatment can also be used in combination with cytotoxic chemotherapy as neoadjuvant treatment prior to surgery. The most common type of treatment are monoclonal antibodies that bind to the extracellular domain of Her2. The first clinically approved antibody is Trastuzumab. It is a monoclonal IgG1 antibody that binds Her2, blocks ligand receptor interaction and inhibits downstream signalling. The bound antibody also mediates antibody-dependent cell-mediated cytotoxicity (ADCC). ADCC is a type of immune defence that may contribute to tumour cell killing: The bound antibodies on the cell surface attract macrophages and natural killer cells. A second approved IgG1 antibody is Pertuzumab. It binds to the dimerisation domain of Her2 and stops dimer-formation, which is required for Her2 activation. Since this is a different mode of action, it can be combined with Trastuzumab. A third antibody-based strategy is the compound Trastuzumab emtansine (T-DM1). It is a combined molecule: Trastuzumab is linked to the cytotoxic agent DM1. The compound has a dual mode of action by blocking Her2 receptor signalling and by intracellular release of DM1 which interrupts cell division by inhibiting the formation of microtubules (Lewis Phillips et al. 2008). Besides antibody-based strategies, small-molecule inhibitors of Her2 such as Lapatinib exists. Lapatinib is a tyrosine kinase inhibitor that blocks Her2 and EGFR (Nelson and Dolder 2006).

The Her2 status of a breast carcinoma is determined by an algorithmic combination of Her2 IHC and Her2 ISH. Respective guidelines have been published by the American Society of Clinical Oncology / College of American Pathologists (ASCO/CAP) (Wolff et al. 2007; 2013; 2018). IHC is performed first. The interpretation is based on the histomorphological pattern and staining intensity. In breast cancer, cancer cells with a membranous, circular staining are included in Her2 scoring. The staining intensity is quantified by a four-step score (0, 1+, 2+, 3+) (Wolff et al. 2007; 2013; 2018). Score 3+ cases are considered as positive and are eligible for anti-Her2-therapy. 0 and 1+ cases are negative and ineligible. Cases with IHC-score 2+ are considered as equivocal and undergo ISH testing. The standard mode of ISH testing uses two probes that bind the *HER2/neu* gene and the centromeric region of the respective chromosome 17. The probes are labelled with different dyes, i.e. dual colour ISH. ISH is interpreted by counting the gene copy number per cell and the ratio of *HER2/neu* signals to chromosome 17 signals. The majority of ISH cases can be classified as either positive (+) or negative (-), a small group of cases remain equivocal.

Interobserver variability was noticed as a potential issue in Her2 IHC interpretation. It can be addressed by specific training of pathologists and external quality assessment of laboratory procedures (Rüschoff et al. 2017).

1.3.3.4 Ki-67

Ki-67 is a nuclear protein that is expressed in proliferating cells. It was discovered in the German city Kiel in 1983 during studies that subtyped lymphomas by generating new antibodies; the respective antibody was clone number 67, hence the name: Ki(el)-67 (Gerdes et al. 1983). Ki-67 is a DNA-binding protein that spatially organises heterochromatin and controls gene expression. Despite its association with proliferation, it is not required for cell cycle control and even cells lacking Ki-67 may divide (Sobecki et al. 2016). Given its constitutive expression in proliferating cells it is widely used as proliferation marker in immunohistochemistry and as prognostic biomarker for malignant neoplasms. A literature search on PubMed for "Ki-67" yields over 29 000 hits. The general utility of Ki-67 IHC is widely accepted but standardisation of Ki-67 IHC staining and interpretation has been challenging. IHC is affected by various pre-analytical issues, including the type and duration of sample fixation (Downsett et al. 2011). Multiple clones are available for IHC that have different binding characteristics (Hida et al. 2020), the overall design of the IHC protocol influences the staining and the interpretation was shown to have limited interobserver reproducibility. For breast cancer, an internal working group for Ki-67 IHC standardisation was formed, that published a validated scoring protocol (Leung et al. 2016) and recommendation for Ki-67 assessment (Dowsett et al. 2011; Nielsen et al. 2020). The prognostic value was disputed by some authors but ultimately confirmed in a large study that re-analysed 8088 cases (Abubakar et al. 2016). Interestingly, Ki-67 has prognostic value only in ER positive carcinomas, but not in Her2 and triple negative carcinomas. Interpretation of Ki-67 is based on the percentage of Ki-67 positive carcinoma cells, which is called Ki-67 index. The staining intensity is not included in the scoring. A minimum of 500 malignant cells should be counted. Ki-67 exhibits spatial heterogeneity in histological specimens and the most common approach is to score the area of highest Ki-67 expression. Quality control by using validated methodology, specific training of pathologists and external quality assessment are highly recommended (Nielsen et al. 2020).

1.4 Digital pathology

1.4.1 Summary

The term digital pathology (DP) commonly refers to digital microscopy, i.e. the interpretation of digitised histopathology glass slides on a computer screen rather than using an optical microscope. Usually, whole slide scanning is employed to create digital images of tissue sections

stained with bright-field dyes including H&E and IHC. DP allows manual interpretation of IHC-based biomarkers and digital image analysis (DIA). Some techniques and procedures also use fluorescent dyes or probes.

During the 2010s the availability of DP became wide-spread since powerful computers and large data storage devices became cheap and ubiquitous and the performance of whole slide histology scanners increased. In parallel, new and versatile software approaches to DIA were developed for various purposes ranging from industrial applications to consumer electronics. While most pathologist still review their cases using optical microscopes, it is now possible to run a pathology laboratory completely digital. Guidelines on the correct application of the new methodology have been published by the ASCO/CAP (Pantanowitz et al. 2013) and by the German *Berufsverband deutscher Pathologen e.V.* (Berufsverband deutscher Pathologen 2018). DP adds a new layer of complexity to clinical histopathology and may require the support of computer scientists. On the other hand, it offers new approaches to standardisation and documentation. Semi-automated or fully-automated analyses of biomarkers may derive information that cannot be determined by manual inspection because the respective features are too subtle or prone to intra- and inter-observer discordance (Stålhammar et al. 2016; 2018).

1.4.2 Image acquisition

The most common way of image acquisition in DP is whole slide scanning. Respective scanners are available from several companies including 3DHistech, Hamamatsu, Leica-Aperio, Philips and Roche Diagnostics. The devices have capacities to load several hundred glass slides and perform the digitisation fully automated. Some devices such as the Pannoramic P1000 scanner (3DHistech) allow continuous scanning, i.e. slides may be loaded and unloaded while the scanner is running.

Most devices rely on tile-stitching: Common microscope optics in combination with a digital camera are used to take microscopic images of the slide that is mounted a moving stage. Depending on the size of the scanned region, several hundred images may be taken. The individual images are subsequently combined by a stitching algorithm into one file that contains the digital slide. An alternative approach are line-scanners which use continuous image acquisition similar to photocopying machines.

Digital slides use common image format systems that may be uncompressed (bitmap, tagged image file) or compressed (JPEG). Uncompressed formats are lossless but produce very large files. JPEG with a quality factor of ≥ 80 is generally considered as sufficient for digital pathology and constitutes a compromise of quality and file size. JPEG-compressed digital slides typically have file sizes of 0.5-4 Gigabyte. Scanning times range from less than a minute for small biopsies to several minutes per slide for large resection specimens.

1.4.3 Image analysis

Digital microscopy has equal sensitivity and specificity as optical microscopy. Validation studies that compared histopathological diagnostics by both methods achieved excellent concordance rates (Campbell et al. 2012; 2014). Digital microscopy may facilitate several aspects of pathology: Digital slides can be archived on network storage devices and are quickly available. They can be reviewed from outside the pathology laboratory and allow working from home office for pathologists. It is easy to share digital slides via a network or the internet for teaching or for reference diagnostic purposes. Digital slides enable DIA, i.e. the usage of software to identify, quantify and/or interpret elements within the slide. Subjectivity and interobserver variability can be issues in the interpretation of IHC-based biomarkers. DIA has the potential for more objective IHC interpretation.

DIA is a branch of the field of computer vision (Foryth and Ponce 2015). Many software algorithms have been developed to automatically identify objects in digital image (Nixon and Aguado 2019). In digital pathology, they can be applied to identify different types of tissue, cells and subcellular structures such as nuclei. Both commercial and free open-source software are available for biomedical DIA. Open-source means that the program source code is available to the general public which enables customisation and advancement of the program. Among the most widely used programs is *ImageJ*, an image analysis environment based on the Java programming language. It can be used for various biomedical purposes from counting bacterial colonies to quantitation of protein blots and histological image analysis (Schneider, Rasband and Eliceiri 2012). An open-source software dedicated specifically to pathology and quantitation of histopathologic biomarkers is *QuPath* (Bankhead et al. 2017). It was developed by an academic consortium at the university of Belfast, Northern Ireland, that includes renowned pathologists. Several studies have shown that *QuPath* can produce equal results compared to commercial DIA software (Bankhead et al. 2018; Acs et al. 2019; Robertson et al. 2020). Since *QuPath* may be used free-of-charge, it has made a worldwide impact on digital pathology (Humphries, Maxwell and Salto-Tellez 2021). Commercial software includes Definiens *TissueStudio*, 3DHistech *QuantCenter* and *Visiopharm*. *TissueStudio* was the first software to rely on image segmentation. Besides impressive results in concordance studies it has not found wide application, possible because of its high costs (Healey et al. 2017). *QuantCenter* and *Visiopharm* are possibly the most commonly used software packages for digital pathology. Both have in vitro diagnostics certifications for specific applications such as automated Ki-67 quantification. They feature a working space in which a collection of software applications can be used to perform specific tasks such as the quantification of stained nuclei in Ki-67 and HR, membrane quantification in Her2 and PD-L1 as well as histomorphological analyses (Hartage et al. 2020; Jeon, Kim and Kim 2021).

DIA may not only mimic manual scoring of IHC-based biomarkers, but has also been demonstrated to perform complex analyses that are impossible by conventional microscopy. In breast cancer, quantification of tumour infiltrating lymphocytes (TILs) can be used as predictive and prognostic biomarker (Ingold et al. 2016). In the neoadjuvant setting, the number of TILs can be predictive for pathological complete response. As prognostic marker, the TIL-number is correlated with overall survival. Structured protocols for manual TIL quantification have been validated and published (Dieci et al. 2018). In principle, TIL-quantification could be performed on conventional H&E stained histological sections and provide additional information. However, such analysis has not found wide application, possible due to limited reproducibility. DIA might overcome such difficulties and enable standardised TIL quantification (Klauschen et al. 2018). Another prognostic biomarker in triple negative breast cancer could be ratio between carcinoma cells and tumour-associated stroma as quantified by the *QuPath* software (Millar et al. 2020).

1.5 Hypothesis

Immunohistochemistry is a frequently used technique to analyse biomarkers in histopathology. It is available in most pathology laboratories and relatively easy to use, however, standardisation is known to be challenging. In breast cancer, IHC of ER, PR, Her2 and Ki-67 is part of the routine histopathological work-up. The derived receptor status has direct implications for the clinical management. DIA has the potential to objectify the interpretation of IHC based stainings. Our hypothesis was:

DIA can be used to increase the standardisation of IHC interpretation in breast cancer biopsies.

To test the hypothesis, we sought to investigate the concordance of manual IHC interpretation and DIA on a large collective of breast cancer biopsy specimens. The following questions were to be addressed:

1.) How is the concordance of manual interpretation and DIA for ER, PR, Her2 and Ki-67?
2.) Which parameters influence concordance rates?
3.) How can DIA be integrated into the diagnostic work flow of a histopathological laboratory?

This work describes the retrospective re-analysis of n = 613 breast cancer biopsy specimens, the optimisation of DIA using the software *QuantCenter*, concordance analyses and practical aspects of DIA and of computer vision in general.

2 Materials and methods

2.1 Samples and ethical approval

Histological slides of biopsy specimens of n = 613 breast cancer patients were retrospectively analysed. The patients were diagnosed within twelve months at one pathological institution (*Pathologie Nordhessen*, Kassel, Germany) (Table 3.1). Laboratory procedures were performed according to quality-certified procedures (*DAkkS Akkreditierung*) that were not changed during that period, ensuring comparable sample treatment.

The study was evaluated by the responsible ethical committee (*Ethik-Kommittee der Landesärztekammer Hessen*, Frankfurt, Germany) and was granted approval (file number FF 135/2013). Given the retrospective character of the study and type of the planned investigations (image analysis of existing histological specimens), individual consent of the involved patients was not considered necessary by the committee. Data were processed using pseudonyms and all published data are anonymised (Scheel et al. 2018).

2.2 Histology and reporting

Biopsy specimens were fixed in 4% phosphate buffered, neutral formalin dehydrated using alcohols of ascending concentrations and embedded in paraffin overnight in an automated device (Shandon Excelsior ES Tissue Processor). On the second day, the tissue was cast into paraffin blocks using an embedding centre (TES 99) and histological sections of 3 µm thickness were cut using a rotation microtome (HM 355S). The sections were placed in a heated water bath to remove wrinkles and mounted on negatively charged glass slides with strong adhesion (StarFrost slides). H&E stainings were performed on an automated staining device (HMS 760X).

The specimens were reviewed by a group of board-certified pathologists at one institution (*Pathologie Nordhessen*, Kassel, Germany) during the routine work-up of the cases. Diagnostic interpretation and reporting was done according to the AWMF S3-guidelines in the 2012 version (the current version was updated in 2020 (AWMF 2020)). In particular, the biopsy-classification was used to classify the specimen and lesion (cf. table 1.1). Reports were created, managed and stored using the software *DC Pathos*.

2.3 Immunohistochemistry

Immunohistochemistry was performed on an automated staining platform that conducts all involved steps (Ventana Benchmark XT): De-paraffinisation, heat-induced antigen retrieval, incubation with primary and secondary antibodies and enzymatic staining reaction.

The tissue is de-paraffinised using a one-step solution. However, the antigens have lost their normal, three-dimensional configuration due to the fixation and embedding process. Since the antigenicity depends on the correct three-dimensional form, it has to be restored in a process called antigen retrieval. This is archived by incubating the tissue in a heated solution for a fixed period of time. Depending on the type of antigen to be restored, the solution can be basic or acid. After antigen retrieval, the antigenicity is restored and the primary antibody can bind its epitope. A polymer-based system was used for detection of the bound primary antibodies. Multiple secondary antibodies and peroxidase-molecules are attached on a tiny filament. The binding of one of the secondary antibodies thus recruits not just one but several peroxidase-molecules to the site of the primary antibody yielding an amplification of the resulting signal. In the present study, the Ventana OptiView assay was used for all IHC stainings. The employed primary antibodies are listed at section 2.9.

2.4 Manual biomarker scoring

Manual scoring of the biomarkers was done by board-certified pathologists as part of routine diagnostic work-up of the biopsy specimens. The IHC stained glass slides were reviewed using standard diagnostic microscopes and the staining patterns were evaluated based on the current clinical guidelines (Dietel and Klöppel 2013).

Expression of the estrogen and progesterone receptors were quantified based on the scoring system according to Harvey and Allred (Harvey et al. 1999) (also called Allred-Score) which combines the number of positive cells and the staining intensity. The staining intensity is classified into four categories (0, negative; 1, weak intensity; 2, moderate intensity; 3, strong intensity). The percentage of positive cells is classified into five categories (0, no staining; 1, < 1%; 2, 1-10%; 3, 11-33%; 4, 34-66%; 5, 67-100%). Intensity and proportion are summed to yield the score which ranges from 0 to 8 (Figure 2.1). A case is considered positive if the score is ≥ 3.

Her2 was quantified based on the ASCO/CAP guidelines (Wolff et al. 2007, 2013, 2018). The IHC staining pattern was classified into four categories (0, negative; 1+, weak; 2+, equivocal; 3+, positive). Categories 0 and 1+ are considered as negative, category 3+ as positive. Equivocal Her2 IHC stainings (category 2+) were supplemented by in situ hybridisation of the *HER2/neu* gene as recommended in the guidelines.

Harvey / Allred Score

Staining intensity	
Pattern	**Points**
No staining	0
weak	1
moderate	2
strong	3

Staining proportion	
Percentage positive	**Points**
No staining	0
< 1%	1
1 - 10%	2
11 - 33%	3
34 - 66%	4
67 - 100%	5

Figure 2.1: The Harvey / Allred-Score for the manual interpretation of ER and PR IHC. Left panel: The staining intensity is classified by a four-step system. Right panel: The percentage of stained cells by a six-step system. The Allred-score is calculated by adding the points of intensity and percentage, i.e. 0-8.

Ki-67 was scored according to Dowsett et al. 2011 and reported as percentage of positive carcinoma cells. The results were stored in the clinical database program *DC Pathos*. For statistical analysis, the results were transferred into a comma separated values (csv) table that can be processed both by *Excel* and by *R* statistical programming language (cf. 2.7).

2.5 Whole slide scanning

Glass slides with tissue sections stained by H&E or IHC were digitised by whole slide scanning. A Pannoramic P250 Flash II scanner (3DHistech, Budapest, Hungary) was used and all slides were scanned on the same device to ensure comparability. The P250 is a tile scanner, i.e. hundreds of individual photos that cover the whole tissue area are captured and are subsequently stitched together to yield one single, seamless image. The area covered by each photo is called field-of-view (FOV). Stroboscopic illumination of the tissue by a xenon flash tube allows for a continuous movement of the slide during scanning. Triggering of the flash tube and image capturing are synchronised which results in a stream of evenly illuminated, sharp pictures. The FOVs slightly overlap to enable photo-stitching, thus, the total number of pixels of the completed scan is slightly less than the sum of pixels of all FOVs. In the P250 Flash II scanner, pictures are captured by a two mega-pixel camera and a Zeiss plan-apochromat microscope objective which yields a resolution of 5.1 pixels per micrometer, i.e. an erythrocyte with a diameter of 7 µm is represented by approximately 36 pixels across. The digital slides are thus true to scale and allow quantifications of physical dimensions. Focussing is achieved by an automated method that relies on five steps: A coarse but very fast pre-image is captured by a secondary macro-camera. The tissue area is automatically identified in the pre-image based on saturation and shape. Focus points are evenly distributed over the tissue area. The focus is automatically determined at each

focus point by an algorithm that browses through the Z-axis. The obtained focus map is used for the high-resolution scan of the slide.

The scanning speed in the investigated collective was 28 ± 1.3 (SD) FOVs per second (Table 3.2). The scanner loads cartridges containing 25 slides. Up to 10 cartridges can be loaded per run, i.e. up to 250 slides.

File names for the digital slides can be manually entered, read from bar codes printed on the glass slides or taken from a predefined csv table. Respective tables can be prepared by common spreadsheet software such as *Excel* (Microsoft, Redmond, USA) or by database software and programming languages. In this study, predefined csv tables were used for slide management. The P250 takes a macro-photo of the label area of the glass slides which can be used to verify slide identity and to make sure that file name and glass slide match.

Scans are saved in the mirax file format (.mrxs). Mirax files may use different storage methods depending on the respective settings, including JPEG, PNG, BMP and TIFF. In this study, the lossy method JPEG at quality factor 85 (0-100) was used as a compromise of image quality and file size. BMP and TIFF are lossless methods but require more storage space.

2.6 Image analysis

2.6.1 Overview of image analysis

Digital slides in the mirax file format can be reviewed with the free-of-charge software *Pannoramic Viewer* (3DHistech). The software offers simple slide management functions, displays the digital slides and shows annotations and results of image analyses. The digital slides can also be annotated; in particular, areas for image analysis can be manually drawn and saved. Image analysis can be performed by two different modes:

The first mode applys a so-called profile on annotated regions or slides. The profile contains all settings and values to run a certain analysis such as Ki-67 quantification. The analysis of multiple slides can be started simultaneously, which is referred to as batch processing.

The second mode is to start the software *QuantCenter* which provides access to all quantification modules and their settings. It can be used to customise an analysis or to set up a new profile.

QuantCenter was used to set up one profile for each analysed biomarker (ER, PR, Her2 and Ki-67). Each profile consisted of a general analysis module called *PatternQuant*, that identifies the neoplastic cells based on their nuclear features and a specific module that quantifies either the nuclear staining (*NuclearQuant*: ER, PR, Ki-67) or the membranous staining (*MembraneQuant*: Her2). Example images of ER analysis (Figure 2.2) and Her2 analysis (Figure 2.3) illustrate the functioning of *NuclearQuant* and *MembraneQuant*.

For image analysis, a suitable tumour region was manually drawn and indicated for quantification. Since the different paraffin sections of each case are not perfectly congruent, the manual drawing was repeated for every digital slide. To ensure representative results, a minimum of n = 500 cells was sought. If less than 500 cells were recognised, a second region was drawn and the results of both regions were combined.

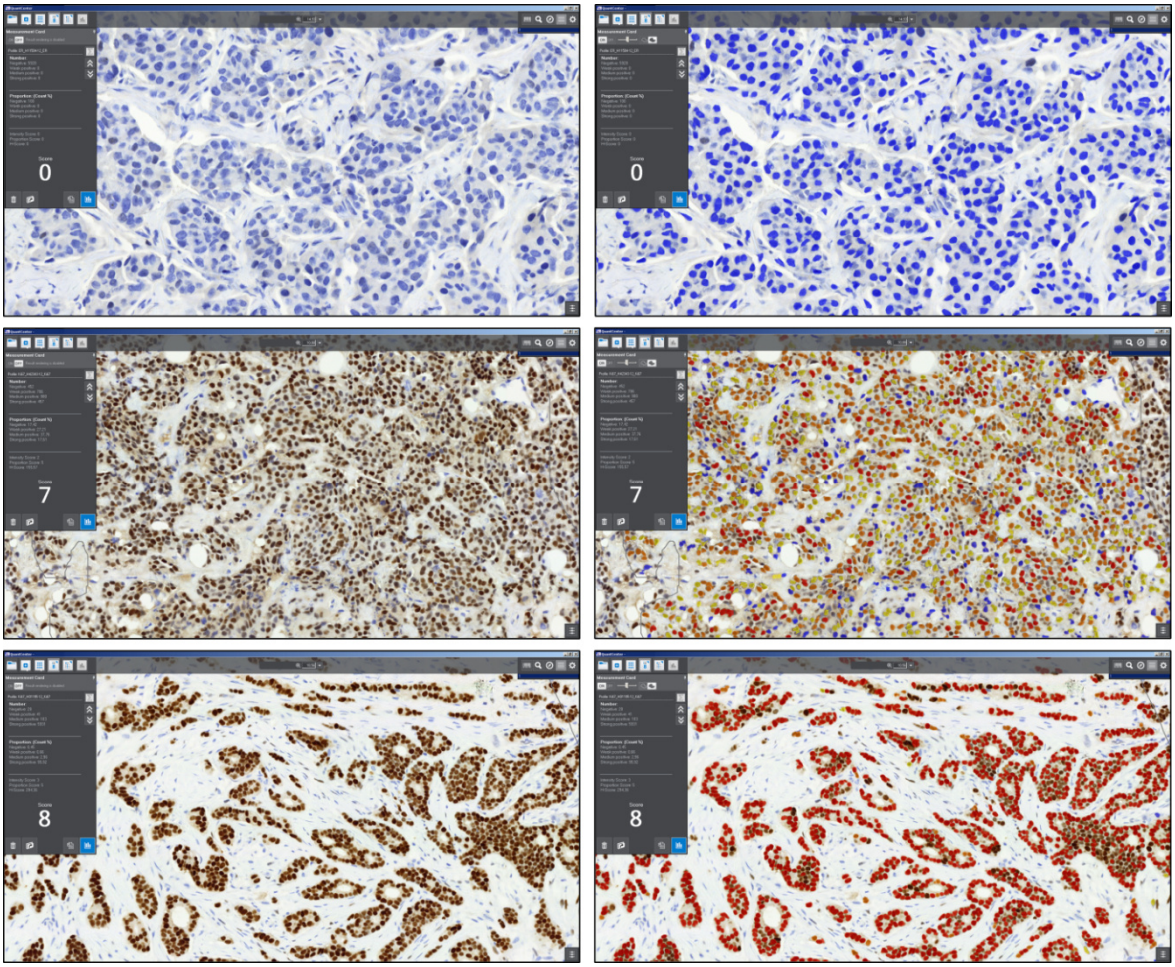


Figure 2.2: Examples of ER IHC (left column) and image analysis (right column) showing different expression levels: Negative (upper row), moderate expression (middle row) and high expression (lower row). ER is visualised by the dark-brownish DAB, the cells are counterstained by the blue hematoxylin. The cells detected by image analysis are highlighted by colours representing the different expression levels.

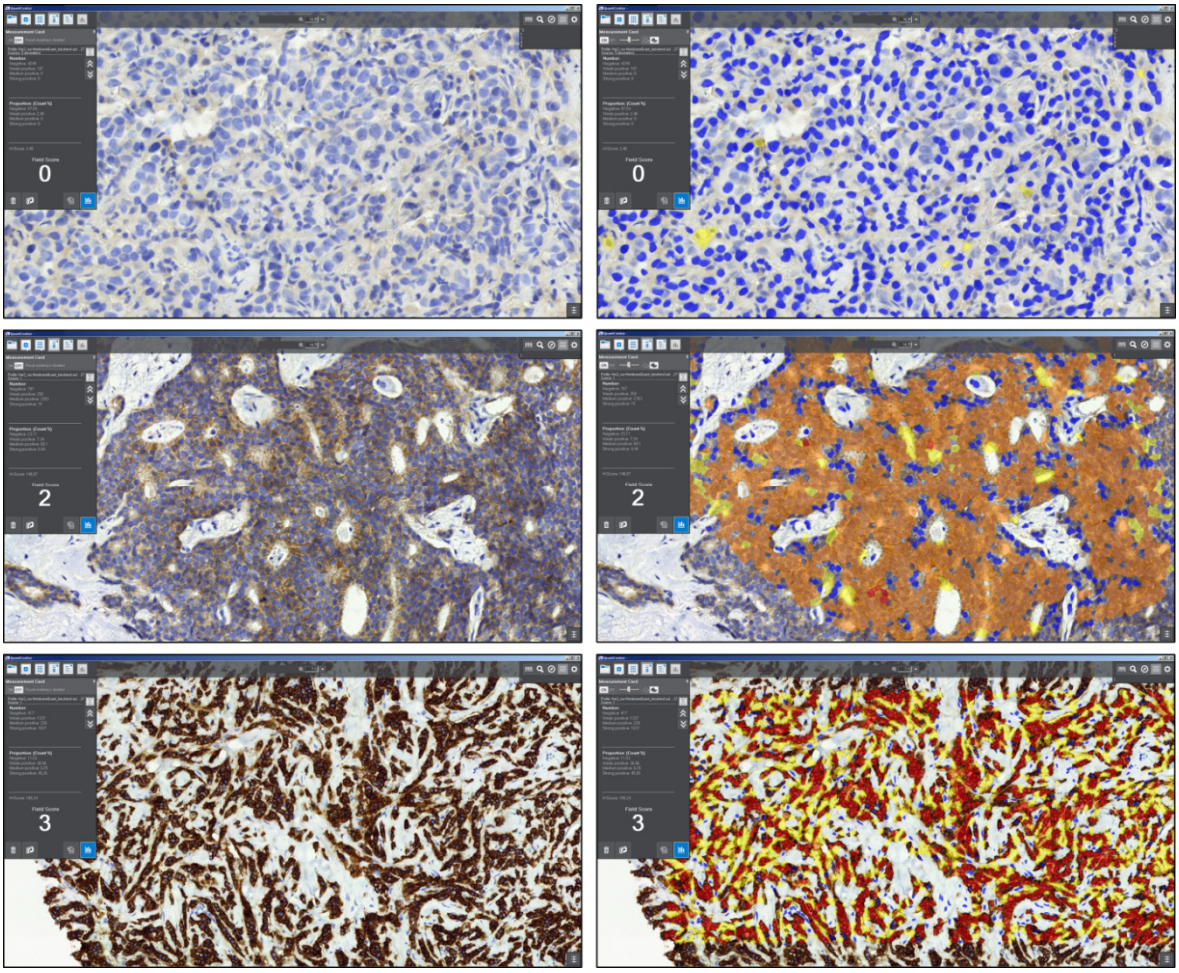


Figure 2.3: Examples of Her2 IHC (left column) and image analysis (right column) showing different expression levels: Negative (upper row), 2+ / equivocal (middle row) and 3+ / positive (lower row). Her2: brown, dark-brown coloured DAB; counterstain: hematoxylin.

The *Quant* software was configured to output scores similar to validated scoring-systems in manual scoring, i.e. the Allred-Score (0-8) for ER and PR and the IHC-Score (0-3+) for Her2. Besides, the software may also output other metrics as quantitation of a respective biomarker. For more detailed analyses, the histoscore (h-score) was used: Positive cells are classified by their expression intensity as 0 (negative), 1+ (weak), 2+ (moderate) and 3+ (strong); each intensity category is multiplied by the respective proportion of cells (0-100%) and the results of each intensity group are summed up (1 * [proportion 1+] + 2 * [proportion 2+] + 3 * [proportion 3+]), yielding an interval scale that ranges from 0 (negative) to 300 (100% of cells are intensity 3+).

2.6.2 Optimisation of image analysis with *QuantCenter*

QuantCenter and its modules for the various biomarkers come with a set of predefined standard parameters. The parameters affect how the neoplastic cells are recognised and how the biomarkers are quantified and categorised. For testing and optimisation, subsets of cases were

selected that are representative for the different biomarkers and show the full spectrum of expression levels. For ER, n = 16 cases were selected that were either negative (n = 5), strongly positive (n = 5) or weakly to moderately positive (n = 6) in manual scoring (Table 2.1). Using the standard settings of the module *NuclearQuant* resulted in good detection of the strongly positive cases, while the moderately positive and negative cases were not reliably identified. Cohen's kappa coefficient (κ) for the overall interobserver concordance was κ = 0.4. Manual optimisation of the parameters improved the recognition of moderate positive and negative cases. The optimised software settings yielded near perfect concordance with manual ER-scoring, κ = 0.86 (Table 2.1). Besides the improved concordance, the number of correctly identified cells also greatly increased (Table 2.1).

Using the ER-settings as starting point, a similar *NuclearQuant* profile was created and optimised for PR-IHC. For PR, n = 15 representative cases were selected that were negative (n = 5), weakly to moderately positive (n = 6) or strongly positive (n = 4) (Table 2.2). Optimisation improved Cohen's kappa from κ = 0.32 to κ = 1.0. Similar to the optimisation of ER, the number of correctly identified cells also improved (Table 2.2).

Likewise n = 16 cases were used to set up a Ki-67 *NuclearQuant* profile (Table 2.3). Using the standard settings did not yield any concordance between manual Ki-67 scoring and image analysis as the majority of values deviated categorically. The mean difference in Ki-67 indices was 30% with a standard deviation of ± 20%. Using the ≥ 20% cut-off produced a Cohen's κ < 0.1. Optimising the software settings strongly improved the concordance; the mean difference was reduced to 17% (SD ± 17%) and Cohen's κ = 0.48. Further cut-off optimisation for the interpretation of Ki-67 is described in the results section (cf. 3.3.2).

For Her2, n = 19 cases were used to set up an optimised *MembraneQuant* profile (Table 2.4). However, this biomarker was more challenging to optimise compared to the other three. Standard settings yielded 14/19 concordant cases, κ = 0.74. Most importantly, one case was manually scored 2+ but 0 by image analysis. Optimised settings yielded greatly increased numbers of identified cells. The 2+ case previously classified 0 was now classified 1+. However, overall 13/19 cases were concordant with a similar κ = 0.74. More detailed analyses into Her2 are described in the results section (cf. 3.3.3). The optimised profiles were used for subsequent analysis of the complete patient cohort.

Table 2.1: Optimisation of *NuclearQuant* for estrogen receptor

Index number	Manual Scoring			Automated scoring, standard settings							Automated scoring, optimised settings						
	ER-Proportion	ER-Intensity	ER-Score	Cells detected	ER, negative	ER, Intensity 1+	ER, Intensity 2+	ER, Intensity 3+	ER-Proportion	ER-Score	Cells detected	ER, negative	ER, Intensity 1+	ER, Intensity 2+	ER, Intensity 3+	ER-Proportion	ER-Score
44	5	3	8	2149	81	26	1184	858	96.23%	8	2336	0	3	164	2169	100.00%	8
45	5	3	8	1137	40	14	822	261	96.48%	8	1214	6	6	153	1049	99.51%	8
46	5	3	8	401	54	56	270	21	86.53%	7	532	51	41	204	236	90.41%	8
48	5	3	8	323	36	44	223	20	88.85%	7	499	24	38	216	221	95.19%	8
50	5	3	8	1179	348	176	574	81	70.48%	7	1124	75	76	394	579	93.33%	8
160	3	2	5	222	156	65	3	0	30.63%	4	952	691	207	54	0	27.42%	4
95	2	2	4	37	31	6	0	0	16.22%	3	1802	1786	16	0	0	0.89%	2
64	2	1	3	0	0	0	0	0			63	49	12	2	0	22.22%	4
133	1	2	3	6	5	1	0	0	16.67%	3	1552	1512	28	12	0	2.58%	3
182	1	2	3	5	5	0	0	0	0.00%	0	873	858	12	0	0	1.37%	3
211	2	1	3	47	32	14	1	0	31.91%	4	1861	1737	103	21	0	6.66%	3
60	0	0	0	175	131	44	0	0	25.14%	4	394	394	0	0	0	0.00%	0
69	0	0	0	0							138	138	0	0	0	0.00%	0
72	0	0	0	0							252	252	0	0	0	0.00%	0
100	0	0	0	0							1164	1164	0	0	0	0.00%	0
105	0	0	0	4	4	0	0	0	0.00%	0	1654	1653	1	0	0	0.06%	2
				Cohen's kappa: **κ = 0.40**							Cohen's kappa: **κ = 0.86**						

Table 2.1: 16 representative cases were used to optimise the settings of the *NuclearQuant* image analysis software for ER. In manual scoring, n = 5 cases were strongly positive (Allred-Score 8), n = 6 cases were weak to moderate (scores 3-5) and n = 5 cases were negative. Using the standard settings for image analysis yielded only a fair concordance (κ = 0.4) and 5685 cells were detected in total. Using the optimised settings, an almost perfect concordance was achieved (κ = 0.86) and 16410 cells were detected.

Table 2.2: Optimisation of *NuclearQuant* for progesterone receptor

Index number	Manual Scoring			Automated scoring, standard settings							Automated scoring, optimised settings						
	PR-Proportion	PR-Intensity	PR-Score	Cells detected	PR, negative	PR, Intensity 1+	PR, Intensity 2+	PR, Intensity 3+	PR-Proportion	PR-Score	Cells detected	PR, negative	PR, Intensity 1+	PR, Intensity 2+	PR, Intensity 3+	PR-Proportion	PR-Score
48	5	3	8	1697	1347	86	86	178	20.62%	6	296	28	16	113	139	90.54%	8
61	5	3	8	3060	1106	386	1198	272	60.65%	7	1729	137	82	351	1159	92.08%	8
62	5	3	8	1703	1329	126	127	121	21.96%	6	2198	1	0	395	1802	99.95%	8
160	5	3	8	504	426	24	29	25	15.48%	5	638	77	32	197	332	87.93%	8
45	4	2	6	4566	4560	4	1	1	0.13%	2	200	32	66	90	12	84.00%	7
44	2	3	5	1206	717	67	180	242	40.55%	7	223	119	7	50	47	46.64%	6
46	2	2	4	381	353	7	10	11	7.35%	5	130	59	17	27	27	54.62%	7
133	1	2	3	1665	1664	1	0	0	0.06%	2	985	958	14	13	0	2.74%	4
50	1	1	2	915	908	5	2	0	0.77%	2	16	16	0	0	0	0.00%	0
182	1	1	2	2978	2965	11	2	0	0.44%	2	1427	1415	11	1	0	0.84%	2
69	0	0	0	3057	3050	7	0	0	0.23%	2	76	76	0	0	0	0.00%	0
72	0	0	0	3481	3453	25	3	0	0.80%	2	99	99	0	0	0	0.00%	0
95	0	0	0	2588	2509	37	33	9	3.05%	3	678	677	0	0	0	0.00%	0
100	0	0	0	724	388	65	105	166	46.41%	7	753	753	0	0	0	0.00%	0
105	0	0	0	3028	2578	359	91	0	14.86%	4	1592	1592	0	0	0	0.00%	0
				Cohen's kappa: **κ = 0.32**							Cohen's kappa: **κ = 1.0**						

Table 2.2: 15 representative cases were used to optimise the settings of the *NuclearQuant* image analysis software for PR, similar as in ER. In manual scoring, n = 4 cases were highly positive (Allred-Score 8), n = 6 cases showed low or moderate expression (scores 2-6) and 5 cases were negative. The standard settings for image analysis yielded a fair concordance (κ = 0.32) while the optimised settings yielded an optimal concordance (κ = 1.0). The detected cells decreased from 31553 to 11040.

Table 2.3: Optimisation of *NuclearQuant* for Ki-67

Index number	Manual	Automated scoring, standard settings		Automated scoring, optimised settings	
	Ki67-Index	Cells detected	Ki67-Index	Cells detected	Ki67-Index
2	80%	2450	100%	3106	100%
13	70%	1548	70%	2356	50%
16	90%	2075	100%	2227	90%
44	20%	1112	60%	1646	40%
45	20%	184	90%	945	60%
46	25%	556	70%	996	60%
50	5%	708	50%	1405	30%
62	10%	13	60%	185	20%
69	60%	764	70%	824	60%
72	60%	394	90%	546	80%
95	50%	2036	60%	2278	50%
100	70%	2078	90%	2359	80%
105	80%	2612	100%	2716	90%
133	40%	2192	90%	2295	80%
160	10%	912	30%	1098	10%
182	40%	1913	80%	1903	70%
Mean difference (± SD):		**29% (± 19.7%)**		**14% (± 17%)**	
Cut-off = 20% :		**κ < 0.1**		**κ = 0.48**	

Table 2.3: 16 representative cases were used showing a spectrum of different Ki-67 indices in manual scoring. Using the standard settings did not yield concordance with manual scoring. Optimised settings for *NuclearQuant* yielded a moderate concordance (κ = 0.48).

Table 2.4: Optimisation of *MembraneQuant* for Her2

Index number	Manual	Automated scoring, standard settings						Automated scoring, optimised settings					
	Her2-IHC-Score	Cells detected	Her2, negative	Her2, Intensity 1+	Her2, Intensity 2+	Her2, Intensity 3+	Her2-Score	Cells detected	Her2, negative	Her2, Intensity 1+	Her2, Intensity 2+	Her2, Intensity 3+	Her2-Score
7	3	1470	146	0	48	1076	3	3231	1027	620	197	1387	3
12	3	580	61	11	118	390	3	2892	1911	378	103	500	3
95	3	1755	43	10	39	1663	3	4083	2055	123	47	1858	3
133	3	3128	353	25	461	2289	3	3408	449	175	369	2415	3
1	2	583	583	0	0	0	0	963	863	68	32	0	1
5	2	1235	1201	0	21	13	2	2793	1663	176	953	1	2
6	2	624	541	1	62	20	2	2584	1846	384	335	19	2
31	2	659	487	17	146	9	2	333	122	131	69	11	2
14	1	367	365	0	2	0	0	1465	913	446	106	0	1
16	1	1675	1673	0	1	1	0	3133	2189	762	182	0	1
40	1	1898	1869	0	18	11	2	3416	1844	794	777	1	2
44	1	2326	2326	0	0	0	0	2730	2518	188	24	0	0
45	1	871	871	0	0	0	0	1692	1672	18	2	0	0
46	1	1205	271	21	812	85	2	3842	336	1977	1458	71	2
2	0	2551	2551	0	0	0	0	2975	2975	0	0	0	0
3	0	687	687	0	0	0	0	1424	1241	135	48	0	1
13	0	2051	2051	0	0	0	0	3286	3029	256	1	0	0
20	0	346	343	0	1	2	0	172	164	7	1	0	0
37	0	2872	2872	0	0	0	0	3250	3250	0	0	0	0
		Cohen's kappa: **κ = 0.74**						Cohen's kappa: **κ = 0.74**					

Table 2.4: 19 representative cases were used, including n = 4 Her2 positive (3+) cases, n = 4 equivocal cases (2+), n = 6 weak cases (1+) and n = 5 cases showing no Her2 staining (0). Optimised settings increased the number of detected cells from 26883 to 47672 while the concordance remained at a good Cohen's kappa of κ = 0.74.

2.6.3 Explorative image analysis with *ImageJ*

ImageJ is an open-source software suit for scientific image analysis based on the Java programming language (Schneider, Rasband and Eliceiri 2012). It is developed by the United States Institutes of Health in Bethesda and can be used for a variety of purposes. In this study, *ImageJ* was used to quantify the membrane thickness and staining intensity of Her2 IHC. A subgroup of n = 120 representative Her2 IHCs was selected from patient cohort, including n = 40 per Her2 staining category 1+, 2+ and 3+. For these cases, the manual Her2-score was created by consensus of three pathologists. The digital slides in the mirax file format were exported as JPEG files to allow quantitation by *ImageJ*. The cell membrane was manually indicated by drawing 4 regions-of-interest (ROIs) perpendicular to the linear DAB precipitates of 10 cells per case, yielding 40 ROIs per case, 1600 per category and 4800 in total. *ImageJ* was used to calculate the width of each ROI in µm and the colour-intensity of the DAB as 8-bit grey scale value, i.e. $2^8 \triangleq$ values ranging from 0 (black, maximum intensity) to 255 (white, minimum intensity). To facilitate interpretation and simplify the figures, intensity value were inverted and displayed as relative values, i.e. 0% (white) to 100% (black).

2.7 Statistics

Statistics were performed using *R* statistical programming language (version 3.1; http://www.r-project.org/) which is available under the GNU General Public License Version 2. Custom scripts were created to process and analyse the image analysis data. To import the data into *R*, the results of each analysed digital slide were exported as *Excel* table by the *QuantCenter* (SI table 1). The *Excel* tables were converted into csv tables by running a batch process that started a visual basic script for each table:

```
Dim oExcel
Set oExcel = CreateObject("Excel.Application")
Dim oBook
Set oBook = oExcel.Workbooks.Open(Wscript.Arguments.Item(0))
oBook.SaveAs WScript.Arguments.Item(1), 6
oBook.Close False
oExcel.Quit
```

e.g. `u:\convert.vbs u:\tabs\H00138-12_ER.xls u:\tabs\H00138-12_ER.csv`

The batch process itself was created by an *R* script that identified each *Excel* table in a directory and created the commands to call the visual basic script by using a `for()` loop.

The converted tables were imported into an *R* working space using `file()`, deconstructed into single lines using `scan()` and the contained values were extracted using `strsplit()`. The obtained values were merged into one comprehensive `dataframe` that was stored as txt file.

Subsequent statistical analyses were performed on the comprehensive table. Subgroups were created using `split()`. The major commands for statistical analysis were: `length()` (number of elements in a respective subgroup), `mean()` (arithmetic mean), `sd()` (standard deviation), `sum()` (sum) and `summary()` (frequency analysis).

Interobserver concordance was analysed by calculating Cohen's kappa coefficient using the function `ckappa()` from the open-source *R* package *psy* version 1.0 by Bruno Falissard. The kappa coefficients were interpreted according to the recommendations by Landis & Koch (Landies and Koch 1977). The values range from 0 to 1.0 and are into five categories, slight ($\kappa < 0.2$), fair (0.2-0.4), moderate (0.4-0.6), substantial (0.6-0.8) and almost perfect ($\kappa \geq 0.8$).

Diagrams were drawn with *R* and Microsoft *Excel*.

2.8 Lists of devices, software and primary antibodies

Table 2.5: List of devices

Description	Name	Manufacturer
External hard disc drives	My Book 2 Terabyte	Western Digital, Irvine, USA
Computer for image analysis	Celsius i7-4770	Fujitsu Technology Solutions GmbH, Munich, Germany
IHC staining device, automated	Benchmark XT	Ventana, Tucson, USA
Whole slide scanner	Pannoramic P250 Flash II	3DHistech, Budapest, Hungary
Cold plate	OTS 40	Medite GmbH, Burgdorf, Germany
Dehydration machine	Shandon Excelsior ES Tissue Processor	Thermo Fisher Scientific GmbH, Schwerte, Germany
Drying cabinet	UNE 400	Memmert GmbH, Schwabach, Germany
Embedding centre	TES 99	Medite GmbH, Burgdorf, Germany
Film Coverslipper	Tissue-Tek Film	Sakura Finetek Germany GmbH, Staufen, Germany
Freezer, -20°C	Liebherr Premium Product line	Liebherr Gruppe, Biberach an der Riss, Germany
Fridge, 4°C	Liebherr Premium Product line	Liebherr Gruppe, Biberach an der Riss, Germany
Magnetic stirrer and hot plate	MR Hei-Standard	Heidolph Instruments GmbH, Schwabach, Germany
Microscope	Ecliple 80i with Plan Fluor Objectives (1x, 4x, 10x, 20x, 60x)	Nikon, GmbH Germany, Düsseldorf, Germany
Pipettes	Eppendorf Research Plus	Eppendorf AG, Hamburg, Germany

Rotation microtome	HM 355 S	MICROM International GmbH, Walldorf, Germany
Scale, digital, De = 0.1g	Kern-PCB6000-1	Satorius GmbH, Göttingen, Germany
Staining machine, HE	HMS 760X	MICROM International GmbH, Walldorf, Germany

Table 2.5: Employed devices including descriptions of the devices, name or product ID and manufacturers with manufacturers' addresses.

Table 2.6: List of software

Name, Version	**Manufacturer**
DC Pathos	DC Systeme, Heiligenhaus, Germany
ImageJ (1.48)	National Institutes of Health, Bethesda, USA; https://imagej.nih.gov/ij/
Microsoft *Office 2010*	Microsoft, Redmond, USA
Pannoramic Viewer (1.15.3)	3DHistech Ltd, Budapest, Hungary
QuantCenter (2.0.46136), containing modules *PatternQuant*, *MembraneQuant* and *NuclearQuant*	3DHistech Ltd, Budapest, Hungary
R language for statistical computing (3.1.0)	R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org
Software packge *psy* (1.0) for R	Bruno Falissard, falissard_b@wanadoo.fr; https://cran.r-project.org/package=psy
SlideScanner Software	3DHistech Ltd, Budapest, Hungary

Table 2.6: Employed software including name, version if available and manufacturers with manufacturers' addresses.

Table 2.7: List of primary antibodies

Antigene	**Clone**	**Manufacturer**
ER	SP1	Ventana, Tucson, USA
Her2	4B5	Ventana, Tucson, USA
Ki-67	30-9	Ventana, Tucson, USA
PR	1E2	Ventana, Tucson, USA

Table 2.7: Employed primary antibodies including bound antigene, clone ID and manufacturer.

3 Results

3.1 Patients' characteristics

The patient cohort included n = 612 female and one male patient. The average age was 62.8 years (standard deviation: ± 13.6 years) and 82.5% of the patients were ≥ 50 years old. A histogram of the age distribution indicates a peak between 60 and 65 years (Figure 3.1, Table 3.1). The majority of cases were diagnosed with ductal invasive carcinoma, i.e. NST (n = 523; 85%), the second most common histotype was lobular invasive carcinoma (n = 66; 10.8%), other histotypes were exceedingly rare (combined: n = 24; 3.8%). CIS was present in n = 76 (12.4%) of cases. Most cases were graded as medium grade (56%); the second most common grade was high grade (30.9%) while well-differentiated carcinoma was less common (6.5%). Concerning stage, complete pathological staging information were available for approximately 70% of cases. In the other 30% of cases, the resection specimens had been submitted to other pathological institutions and only information about the biopsy specimens were available. Most cases featured pT1, i.e. had a largest tumour diameter of ≤ 2cm (39.8%) or pT2, i.e. > 2cm and ≤ 5cm (24.3%) and featured no lymph node metastases (pN0 = 43.9%).

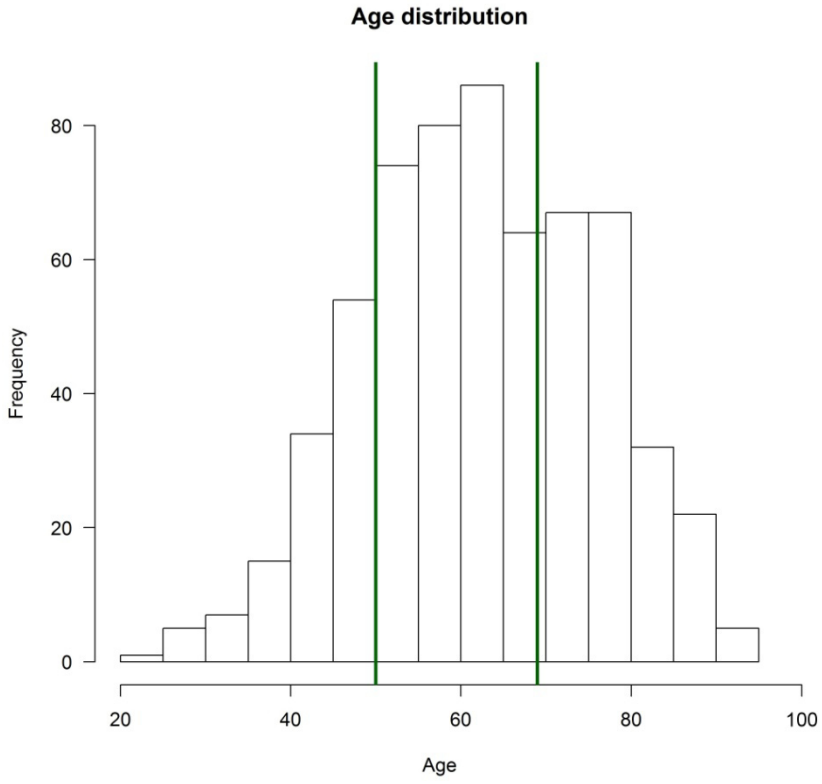


Figure 3.1: Distribution of the patients' age among the analysed cases.

Table 3.1: Characteristics of patient cohort

Patients		n = 613
Age (mean ± SD)		62.8y ± 13.6y
Age ≥ 50y		505 (82.4%)
Sex		F: 612, M: 1
Histotype	Ductal	523 (85.3%)
	Lobular	66 (10.8%)
	Other	24 (3,9%)
Invasiveness	Invasive	611 (99.7%)
	Invasive with CIS	76 (12.4%)
	CIS	2 (0.3%)
Grade	Grade 1	40 (6.5%)
	Grade 2	344 (56.1%)
	Grade 3	190 (30.9%)
	n/a	39 (6.4%)
Staging, pT	pT1	244 (39.8%)
	pT2	149 (24.3%)
	pT3	20 (3.3%)
	pT4	14 (2.3%)
	n/a	186 (30.3%)
Staging, pN	pN0	269 (43.9%)
	pN1	102 (16.6%)
	pN2	29 (4.7%)
	pN3	24 (3.9%)
	n/a	189 (30.8%)
Biopsies/Block (mean ± SD)		3.82 ± 1.2

Table 3.1: Age, histotypes and staging information of the analysed patients.

3.2 Image acquisition

A total of n = 3065 glass slides were digitised for H&E staining, IHC for ER, PR, Her2 and Ki-67 and additional slides with IHC for E-cadherine, a diagnostic marker to distinguish lobular mammary carcinoma from NST carcinomas, in situ hybridisation of *HER2/neu* in cases that were Her2 IHC 2+ / equivocal and some slides that were repeated for quality issues (Table 3.2). On average, a glass slides featured n = 2256 FOVs. Since the size of the tissue per slide varied between the cases, the standard deviation for the FOVs per slide was relatively high, ± 1036. The scanning time per slide was 82.1 seconds with a standard deviation of 38 seconds; thus, the majority of slides was

digitised in less than two minutes. The scanning speed showed a mostly linear relation between the number of FOVs and required scanning time, with Spearman's rho = 0.96 (Figure 3.2). In cases with a relatively large tissue area (> 3000 FOVs) a slight increase in variance was noticed, i.e. some slides required slightly more scanning time. This might be attributed to an increased number of focus points or more time required for moving the glass slide inside the scanner. Overall, image acquisition worked adequately in nearly every slide.

Table 3.2: Performance of the P250 slide scanner

Scanner	
Camera	2 Megapixel
Resolution	5.1 px/µm
FOV/s	28 ± 1.3
mm²/s	0.215 ± 0.001
Scans	
Scans	n = 3065
FOVs	2256 ± 1036
Scan-Time (s)	82.1 ± 38
rho (Time, FOVs)	0.96

Table 3.2: Properties of the employed P250 whole slide scanner (upper part, scanner) and performance during the conducted scans (lower part, scans).

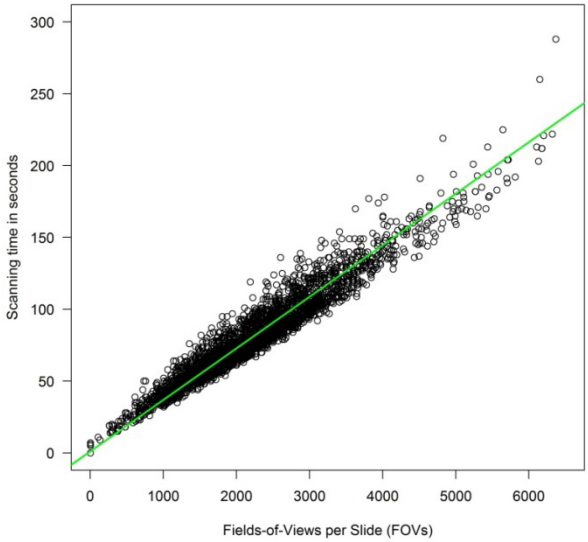


Figure 3.2: Scatterplot of Fields-of-view and scanning time for each scanned histological specimen. The two parameters show a mostly linear correlation, Spearman's rho = 0.96.

3.3 Manual scoring versus image analysis

3.3.1 Hormone receptors

Manual scoring of the IHCs for ER and PR was based on the Harvey/Allred-Score that combines the proportion of stained cells and the staining intensity (cf. section 2.4). According to the respective guidelines, ≥ 3 was used as cut-off value to define positive cases. The majority of cases was HR positive (ER: 85.4%, PR: 77.2%), most frequently ER-PR double positive (76.5%) (Table 3.3). Histograms of ER and PR showed a bimodal distribution since most cases were either

strongly positive for ER and PR, or negative. Only a small proportion of cases showed a weak or moderate expression (Figures 3.3 A, 3.4 A).

Image analysis of ER and PR was optimised by using a representative subset of cases as described in the methods sections (cf. 2.6.2). Using the optimised software profiles yielded near perfect concordance compared to manual scoring, with κ = 0.86 for ER and κ = 1.0 for PR (Tables 2.1, 2.2). Besides the increase in concordance rates, the numbers of identified cells also increased leading to more robust classification of the cases. Analysis of the entire cohort produced positivity frequencies and distributions of the respective scores similar to manual scoring (Figures 3.3 B, 3.4 B). The concordance to manual scoring was substantial based on Cohen's kappa coefficient for interoberserver concordance: κ = 0.77 for ER and κ = 0.71 for PR. The majority of discordant cases resulted from an increase of weak and moderate cases in image analysis, that were scored as strongly positive in manual scoring (Figures 3.3 B, 3.4 B). Besides the Allred-Score for ER/PR, the more detailed h-score was also evaluated (Figures 3.3 C, 3.4 C). The h-score also showed bimodal distributions for ER and PR. For ER, a distinct peak for score ≥ 270 was observed and few cases showed score 1-200 (Figure 3.3 C). For PR, the distribution was more even with no obvious peak for highly positive cases (Figure 3.4 C).

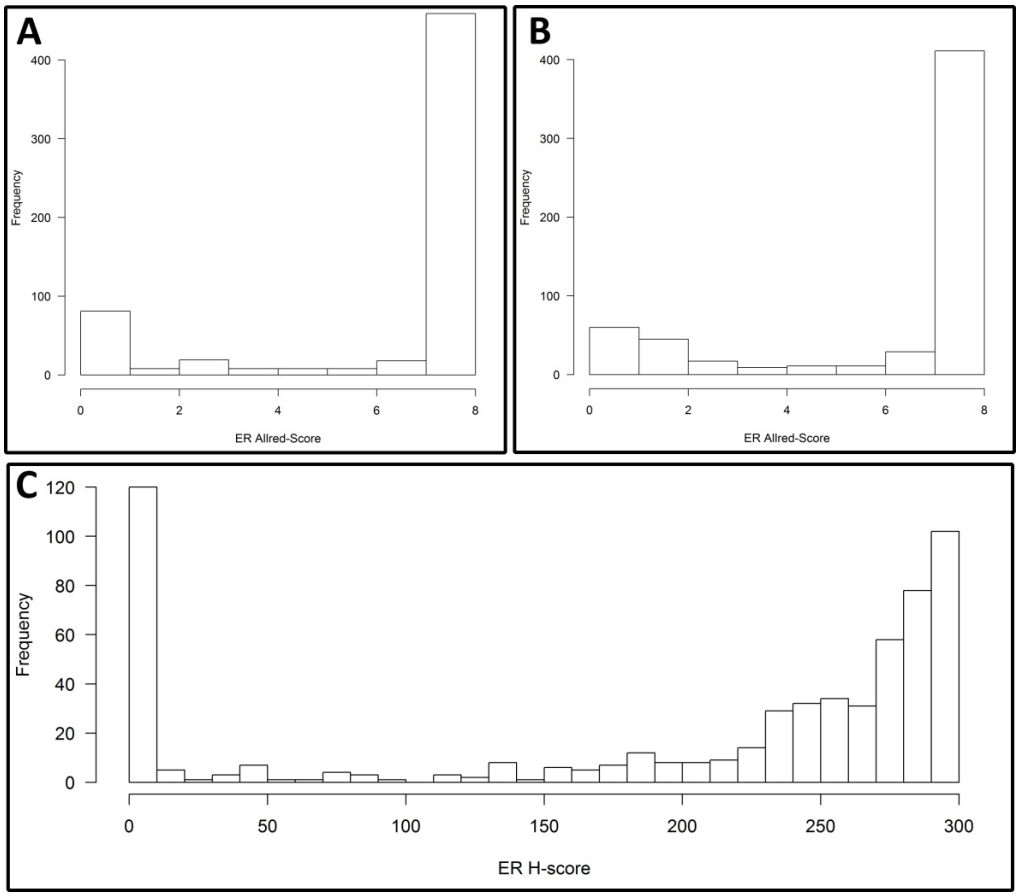


Figure 3.3: A: Distribution of the Allred-Scores for ER in manual scoring. B: Distribution of the ER scores in DIA. C: Alternative representation of ER in DIA by using the h-score.

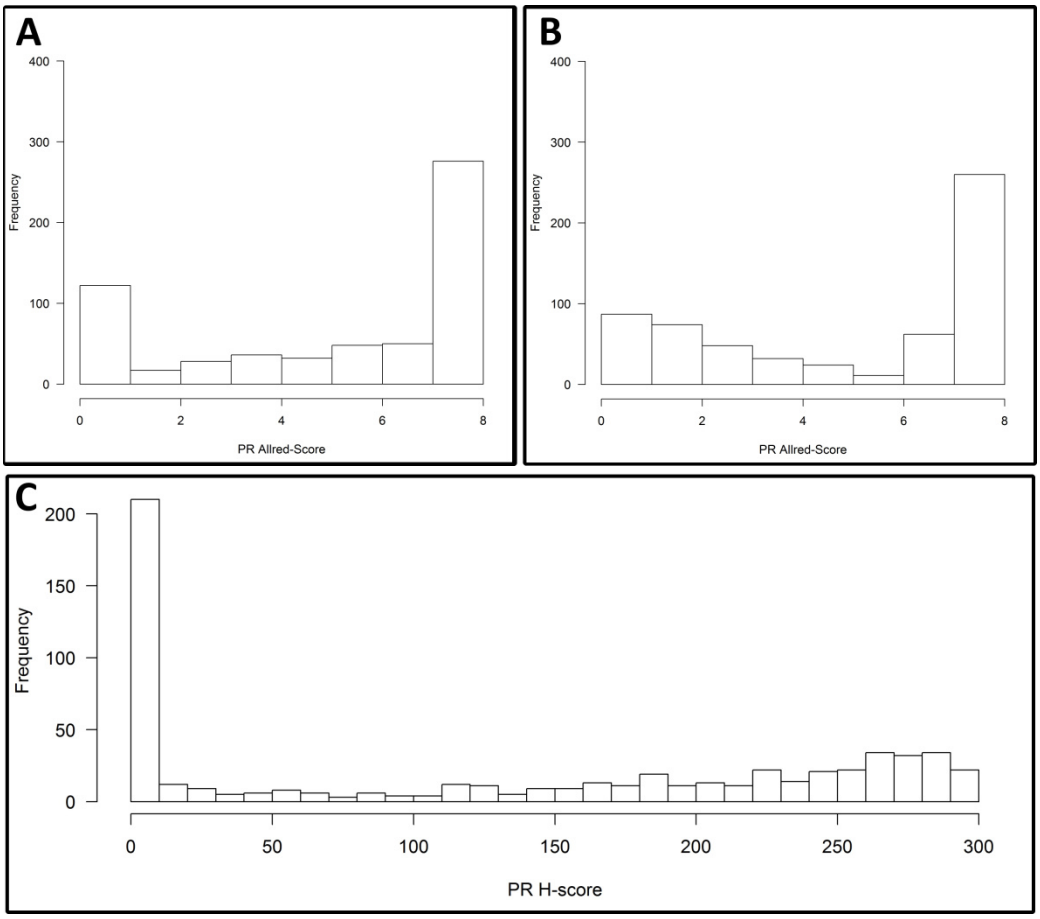


Figure 3.4: Distribution of the Allred-Scores for PR in manual scoring. B: Distribution of the ER scores in DIA. C: Alternative representation of ER in DIA by using the h-score.

Table 3.3: Manual scoring of ER and PR

	ER	PR
0	81 (13.2%)	122 (19.9%)
2	8 (1.3%)	17 (2.8%)
3	19 (3.1%)	28 (4.6%)
4	8 (1.3%)	36 (5.9%)
5	8 (1.3%)	32 (5.2%)
6	8 (1.3%)	48 (7.8%)
7	18 (2.9%)	50 (8.2%)
8	459 (74.9%)	276 (45%)
NA	4 (0.7%)	4 (0.7%)
Negative	89 (14.6%)	139 (22.8%)
Positive	520 (85.4%)	470 (77.2%)

Table 3.3: Distribution of the Allred-Scores for ER and PR in manual scoring.

3.3.2 Ki-67

The interpretation of Ki-67 is less rigorously established compared to the hormone receptors and Her2 and different cut-offs have been discussed in the literature. Many guidelines do not state numeric values for the interpretation of Ki-67 but recommend a classification into low and high based on local standards (Goldhirsch et al. 2013; AWMF 2020). Thus, calculation with different cut-offs were performed. Ki-67 is usually reported as Ki-67 labelling index, i.e. the percentage of positive carcinoma cells. The staining intensity is not incorporated into the Ki-67 index.

According to the ≥ 20% cut-off, manual scoring indicated 60.2% of cases as Ki-67 high and 39.8% of cases as Ki-67 low. The labelling index showed a right skewed distribution with a peak between 0-20% and a long slope ranging from 20% to 100%, indicating that the Ki-67 high cases showed a wide range of different percentages (Figure 3.5 A).

Image analysis did not produce relevant concordance if the standard settings were used. After optimisation of an subset of n = 16 cases, the number of detected cells increased, the mean difference of the Ki-67 indices produced by manual scoring and image analysis was reduced from 30% ± 20% to 14% ± 17% and a Cohen's kappa of κ = 0.48 was achieved (Table 2.3).

Analysis of all cases using the optimised Ki-67 profile produced different distribution of the Ki-67 indices compared to manual scoring (Figure 3.5 B). The cases were more evenly distributed. The peak of cases with a relatively small index was broader ranging from 0-30% and lower compared to manual scoring. Calculation of the h-score, i.e. including the staining intensity, show a peak between h-score 0 and 70 that was better delimited, but showed several more local minima and thus did not point at an obvious cut-off for interpretation (Figure 3.6 A). Overall, the results were not directly comparable to manual scoring.

Concordance analysis for manual scoring of Ki-67 and image analysis resulted in fair to moderate interobserver concordance if the same cut-off values were used. For example, using ≥ 14% as cut-off yielded 81% concordant and 19% discordant cases, κ = 0.44 (Figure 3.6 C).

Given the distribution of the h-score in image analysis of Ki-67, different cut-offs were employed for the classification of manual scoring and image analysis (Figures 3.6 B, C). The concordance improved. By systematic evaluation of three manual cut-offs (≥ 10%, ≥ 14%, ≥ 20%) with any possible cut-off for image analysis (1-99%), an optimal combination could be identified. For ≥ 20% (manual scoring) and ≥ 27% (image analysis), the interobserver concordance was κ = 0.68, indicating substantial concordance.

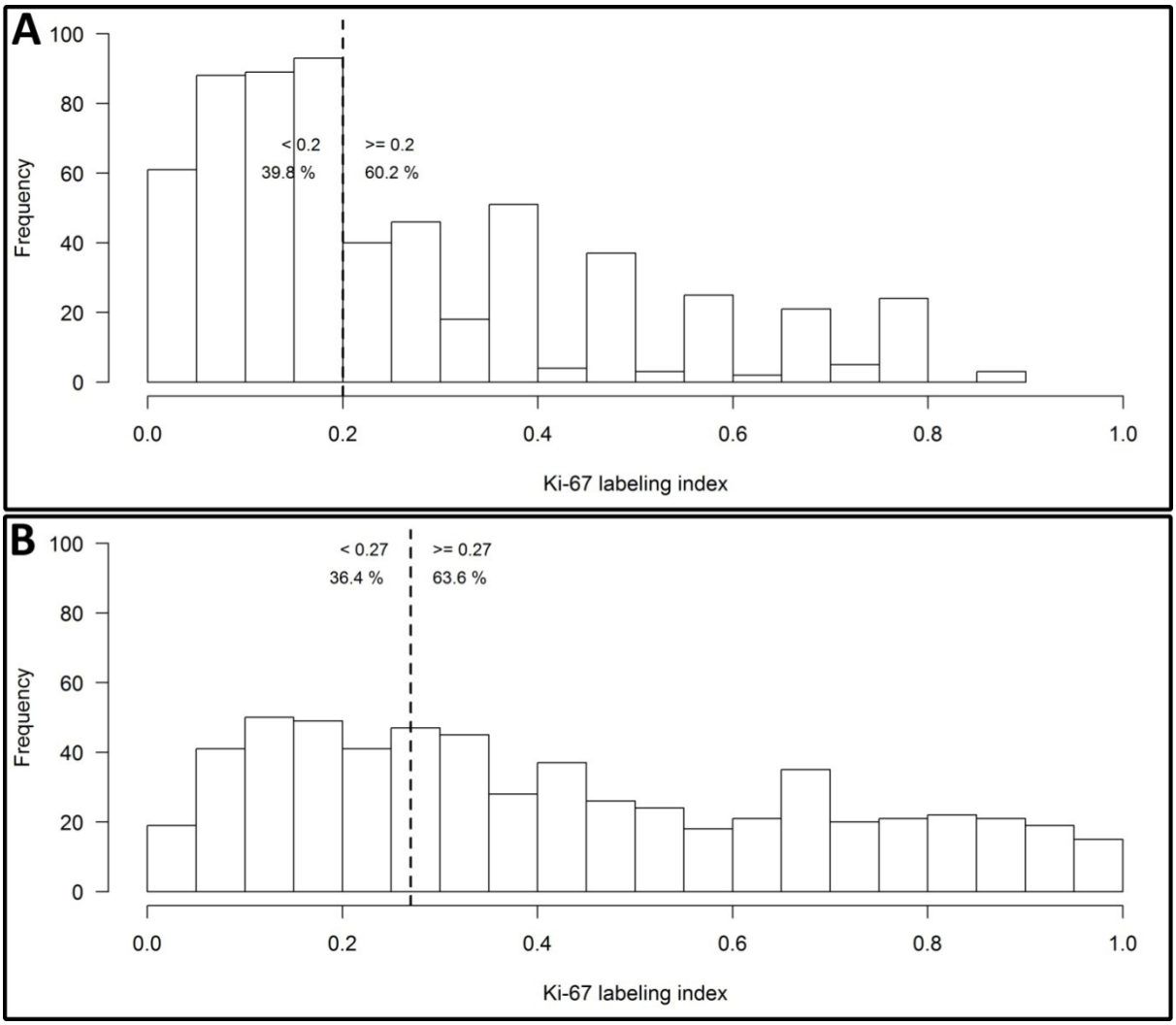
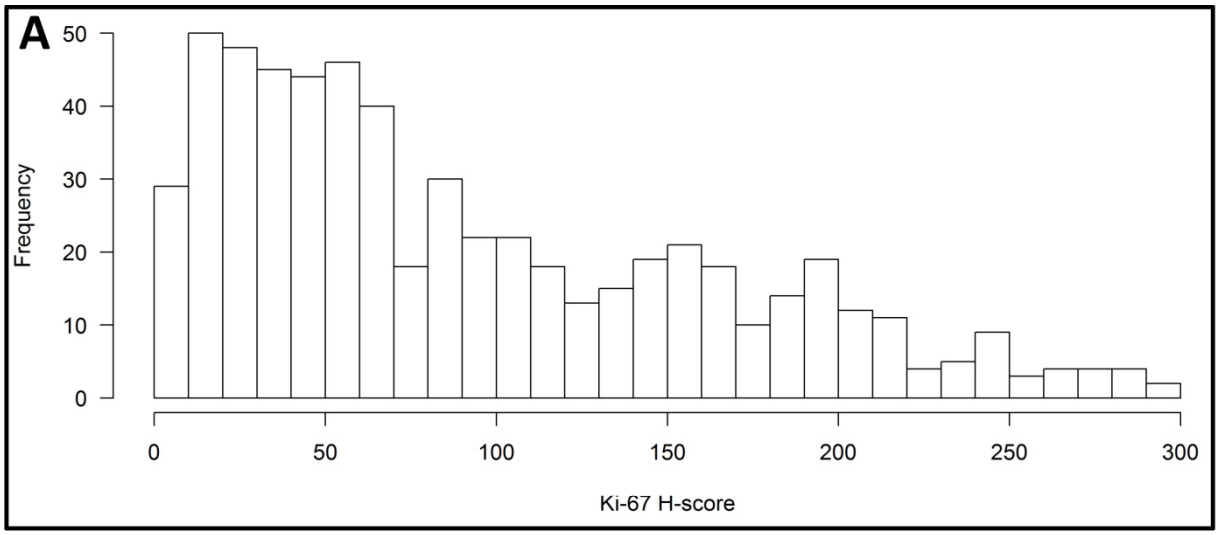
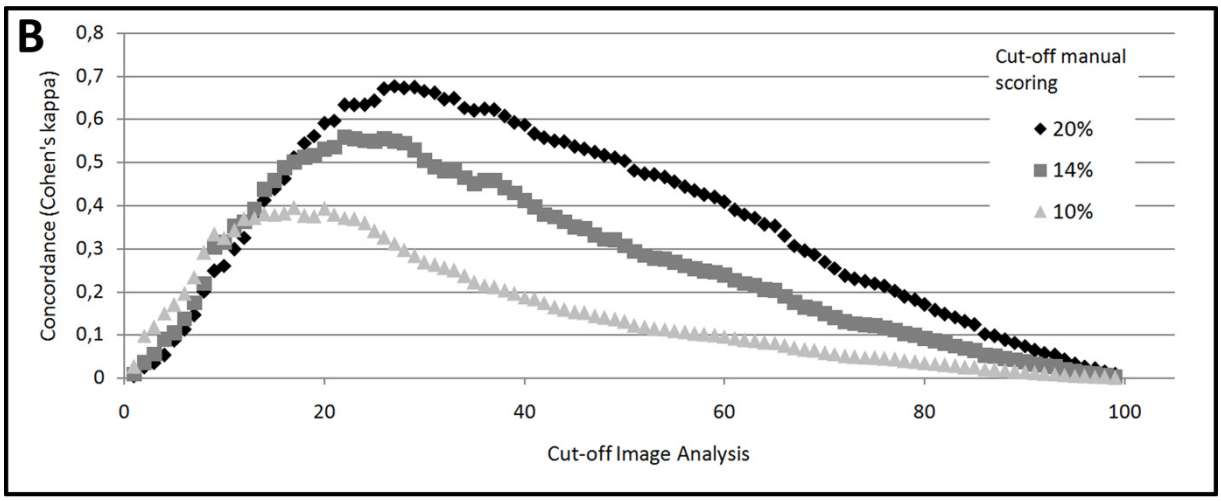


Figure 3.5: A: Distribution of the Ki-67 indices among the analysed cases by manual scoring. B: Distribution in DIA. The dashed lines indicate the optimised thresholds of 0.2 (A) and 0.27 (B) which yield the highest concordance.





C

		Image Analysis		
		Low (<14%)	**High (>=14%)**	Sum
Manual Score	**Low (<14%)**	**69**	79	148
	High (>=14%)	33	**416**	449
	Sum	102	495	597

D

		Image Analysis		
		Low (<27%)	**High (>=27%)**	Sum
Manual Score	**Low (<20%)**	**181**	54	235
	High (>=20%)	37	**325**	362
	Sum	218	379	597

Figure 3.6: Optimisation of Ki-67 interpretation. A: Distribution of Ki-67 scores in DIA represented as h-score. B: Concordance (y-axis) of combination of three different cut-off values for manual scoring (≥ 10%, ≥ 14%, ≥ 20%) and all possible cut-off values for DIA (x-axis). The highest concordance is achieved by a combination of ≥ 20% (manual scoring) and ≥ 27% (DIA). C: Classification of cases according to manual scoring and DIA by using a cut-off of ≥ 14% for both methods. D: Classification of cases according to manual scoring and DIA if optimised cut-offs are used.

3.3.3 Her2

While ER, PR and Ki-67 are nuclear IHC markers, Her2 is a transmembrane protein and shows a membranous IHC pattern. In breast cancer, manual Her2 IHC scoring is based on pattern and staining intensity as outlined in the ASCO/CAP recommendations (Wolff et al. 2007). Cases that are 2+ / equivocal in IHC are further analysed by in ISH of the *HER2/neu* gene to determine the amplification status. Manual IHC scoring showed that the majority of cases was negative (81%) showing either no IHC staining (category 0, 20.1%) or weak staining (category 1+, 39%) or showing a combination of IHC 2+ / equivocal but ISH negative (21.5%). 19% of the cases were classified as Her2 positive, most frequently due to a positive IHC-score (category 3+, 13.7%); only 5.5% of cases were IHC 2+ / equivocal and ISH positive (Table 3.4).

Image analysis of Her2 turned out to be more complicated compared to ER, PR and Ki-67. As described in the methods section, optimisation increased the number of identified cells per case and altered to classification of some cases. The likelihood of false-negative cases was reduced but several false-positive cases remained in the training set of n = 19 cases (Table 2.4). However, the overall concordance did not improve with $\kappa = 0.74$.

Analysis of all cases produced a moderate concordance of $\kappa = 0.43$ if the results of image analysis were interpreted according to the ASCO/CAP 2013 guidelines (Figure 3.7), i.e. were classified according to the four-step score 0, 1+, 2+ and 3+. 57.5% of cases showed concordant score. The majority of discordant cases were scored higher by image analysis compared to manual scoring, 74.5%. About one quarter of the discordant cases were scored lower, 25.4%. A histogram of the h-score for Her2 showed several local maxima and seemed to correspond to the different manual scoring categories (Figure 3.7 A). Thus, h-score cut-off values could be derived and the concordance between image analysis / h-score and manual scoring could be calculated, which yielded $\kappa = 0.48$ (Figure 3.7 C).

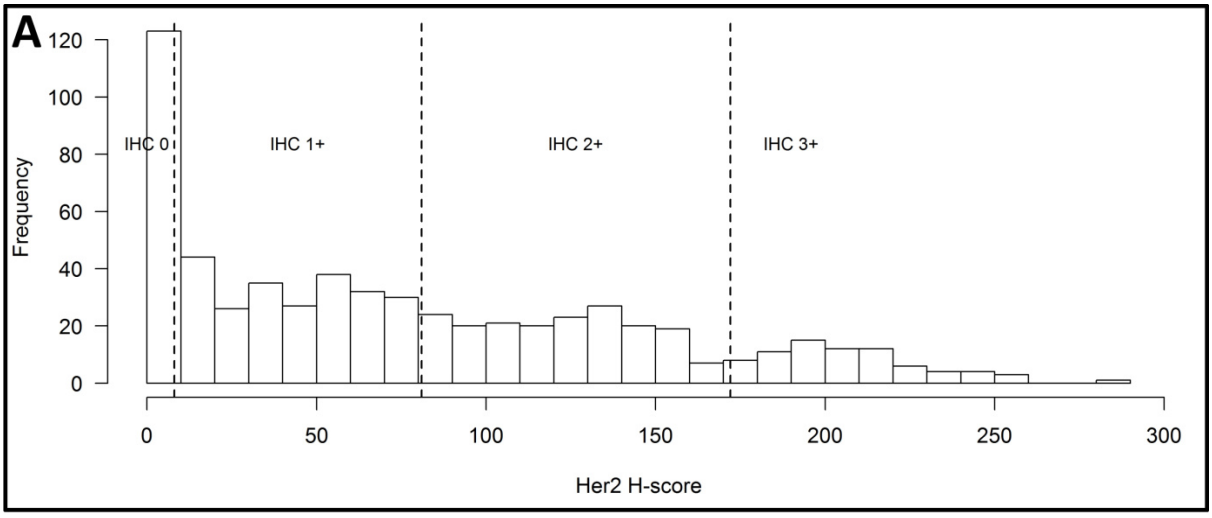
Since score 0 and 1+ cases are both considered as negative and do not require additional ISH testing, a discordance between 0 and 1+ is of less importance. Hence, the two groups 0 and 1+ can also be analysed combined. The concordance of the resulting three-step score (negative (0,1+); equivocal 2+ and positive 3+) yielded a moderate kappa of $\kappa = 0.55$.

Overall, the results of Her2 image analysis remained unsatisfactory in spite of extensive optimisation efforts. An exploratory analysis of manual Her2-scoring was supplemented by re-analysing the same digital slides (cf. 3.4).

Table 3.4: Manual scoring of Her2

Her2 IHC-Score	n (%)	
0	123 (20.1%)	
1+	239 (39.0%)	
2+	166 (27.1%)	
	FISH positive:	34 (5.5%)
	FISH negative:	132 (21.5%)
3+	84 (13.7%)	
NA	1 (0.2%)	

Table 3.4: Distribution of the Her2 scores in manual scoring and FISH status of the *Her2/neu* gene in IHC 2+ cases



B

		Image Analysis				
		Negative (0)	**Negative (1+)**	**Equivocal (2+)**	**Positive (3+)**	Sum
Manual Score	**Negative (0)**	79	26	11	0	116
	Negative (1+)	45	61	124	1	231
	Equivocal (2+)	2	12	123	23	160
	Positive (3+)	0	0	4	72	76
	Sum	126	99	262	96	583

C

		Image Analysis, H-Score				
		0-7	**8-80**	**81-172**	**>=172**	Sum
Manual Score	**Negative (0)**	77	39	6	1	123
	Negative (1+)	32	148	57	2	239
	Equivocal (2+)	1	56	98	11	166
	Positive (3+)	0	5	16	62	83
	Sum	110	248	177	76	611

Figure 3.7: Optimisation of Her2 image analysis interpretation. A: Distribution of Her2 scores in DIA represented as h-score. Dashed lines indicate the cut-offs to classify the four scoring categories. B: Classification of cases according to manual scoring and DIA by standard setting for *MembraneQuant*. C: Classification of cases by using optimised settings for *MembraneQuant* based on the h-score. The number of cases in B and C differes, because not all cases could be successfully processed in B.

3.4 Physical basis of manual Her2 scoring

The comprehensive collection of Her2 IHC slides was used for an explorative study into the physical basis of manual Her2 scoring. A subset of n = 120 cases of NST carcinoma was selected that best represents the different manual interpretation categories 1+, 2+ and 3+, n = 40 for each category. Negative cases of the 0 category were not included as they do not show any DAB precipitate. All IHC 2+ cases were supplemented by ISH of Her2 and n = 20 cases were ISH positive and n = 20 cases ISH negative. The width and the colour intensity of the linear DAB precipitates at the cell membranes were measured by the *ImageJ* software. The resulting values represent the physical appearance of the Her2 staining in an objective way. Correlation to the respective manual scoring category revealed the underlying physical relations.

3.4.1 Her2 scores are correlated with DAB width

According to the so-called magnification rule for the manual interpretation of Her2 IHC, the scoring category and microscope objective required to observe the DAB precipitate are interrelated. Figure 3.8 illustrates this relationship: In positive cases, the Her2 staining, i.e. the DAB colour precipitates in the tissue, are readily observable at low magnification. In moderate and weak cases the staining is not visible at low magnification. Only at high magnification can the linear, membranous staining be appreciated.

Based on the numerical aperture (NA) of the microscope objectives and the wavelength of employed light (λ), the microscope resolution can be calculated, i.e. the minimum distance required to distinguish two points. NA is a dimensionless measure that combines the index of refraction (n) and the opening angle (α) of the objective, i.e. NA = n * sin(α). The resolution (d) is calculated by the formula $d = \frac{\lambda}{NA}$. For a wavelength of λ = 600 nm, which corresponds to an orange hue, the resolution of common standard diagnostic objectives are: 5x (NA = 0.12-0.15): 2.0-2.5 µm, 10x (NA = 0.25-0.30): 1.0-1.2 µm, 20x (NA = 0.40-0.50): 0.60-0.75 µm, 40x (NA = 0.65-0.75): 0.40-0.46 µm (Figure 3.9).

The digital slides created in this project have a resolution of 5.1 pixels per µm, i.e. each pixel represents 0.2 µm and thus allows meaningful quantitation of objects in the range of one micrometer. The DAB precipitates of n = 120 Her2 IHCs were measured by image analysis using the software *ImageJ* and differences in the DAB width were found (Figure 3.9).

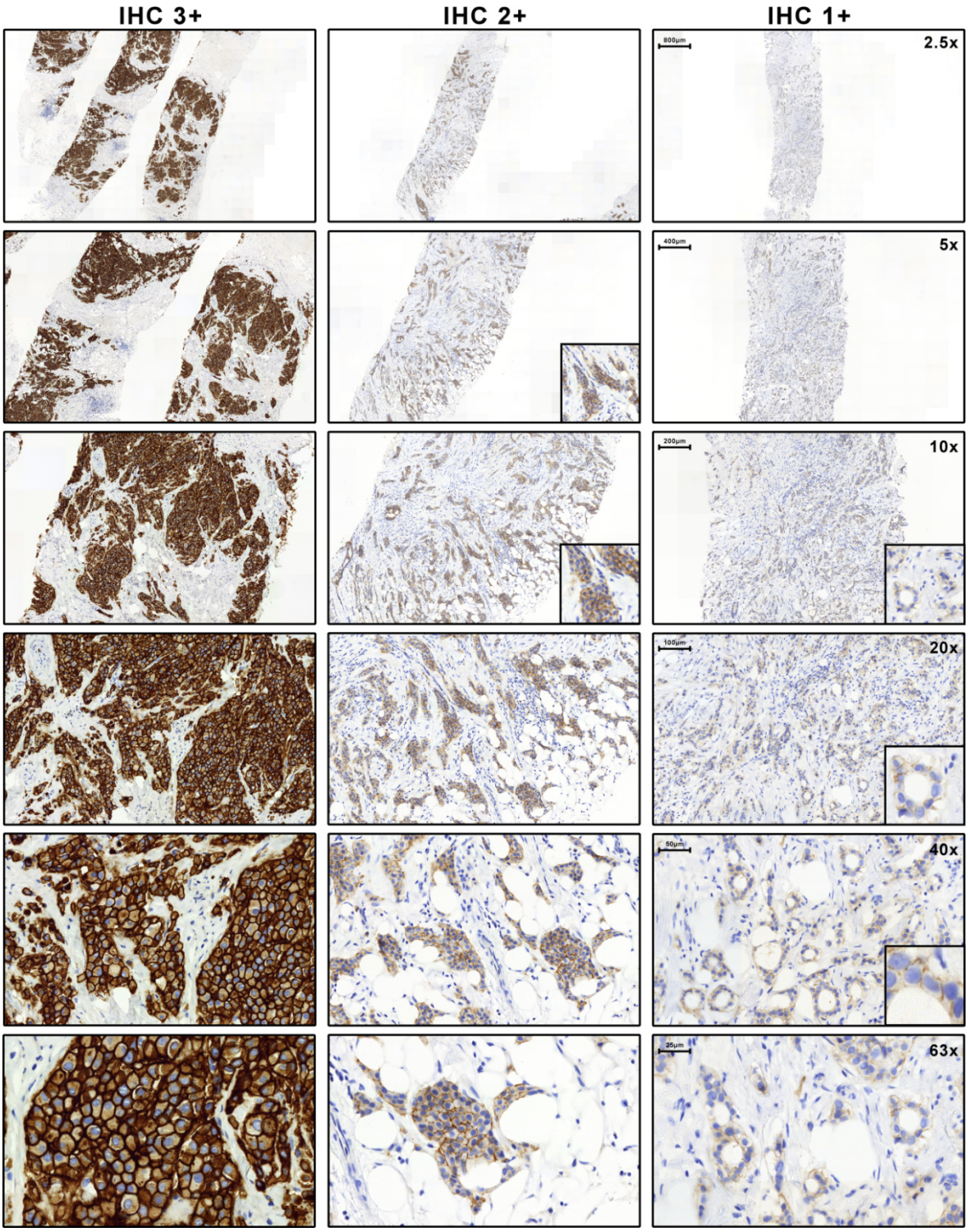


Figure 3.8: Example images of Her2-IHC illustrating the magnification rule. The three cases (1+, 2+ and 3+) show typical cases of the different scoring categories. If the different magnification levels (2.5x, 5x, 10x, 20x, 40x and 63x) are compared, 3+ staining is readily observed at low magnification while 2+ and 1+ stainings are only visible at moderate or high magnification.

Frequency analysis of the obtained values across all cases showed two peaks at 0.65 µm and 2.2 µm as well as local minimum at 1.5 µm. This correlates with the microscope objectives since the 2.2 µm peak is in the range of the 5x objective, the 0.65 µm peak is in the range of the 20x and 40x objectives and the values around 1.0 µm can be observed with the 10x objective. The correlation is highlighted further by calculating the mean DAB width per scoring category. The categories 1+, 2+ and 3+ show DAB precipitates, while category 0 is negative and does not contain DAB. In three groups that have DAB, the width significantly differs: 1+, 0.64 ± 0.1 µm; 2+, 1.0 ± 0.23 µm; 3+, 2.14 ± 0.4 µm (arithmetic mean ± standard deviation). The groups show increasingly wide precipitates and only group 3+ is wide enough to be observable with common 5x objectives while 2+ and 1+ require 10x or 20x/40x objectives respectively.

A

Magnification	40x	20x	10x	5x
Consensus Intensity-Score	'1+'	'2+'		'3+'
Num. Aperture	0.65 - 0.75	0.40 - 0.50	0.25 - 0.30	0.12 - 0.15
Resolution [µm]	0.40 - 0.46	0.60 - 0.75	1.0 - 1.20	2.0 - 2.50
DAB-Precipitate width ± SD [µm]	0.64 ± 0.1	1.0 ± 0.23		2.14 ± 0.4

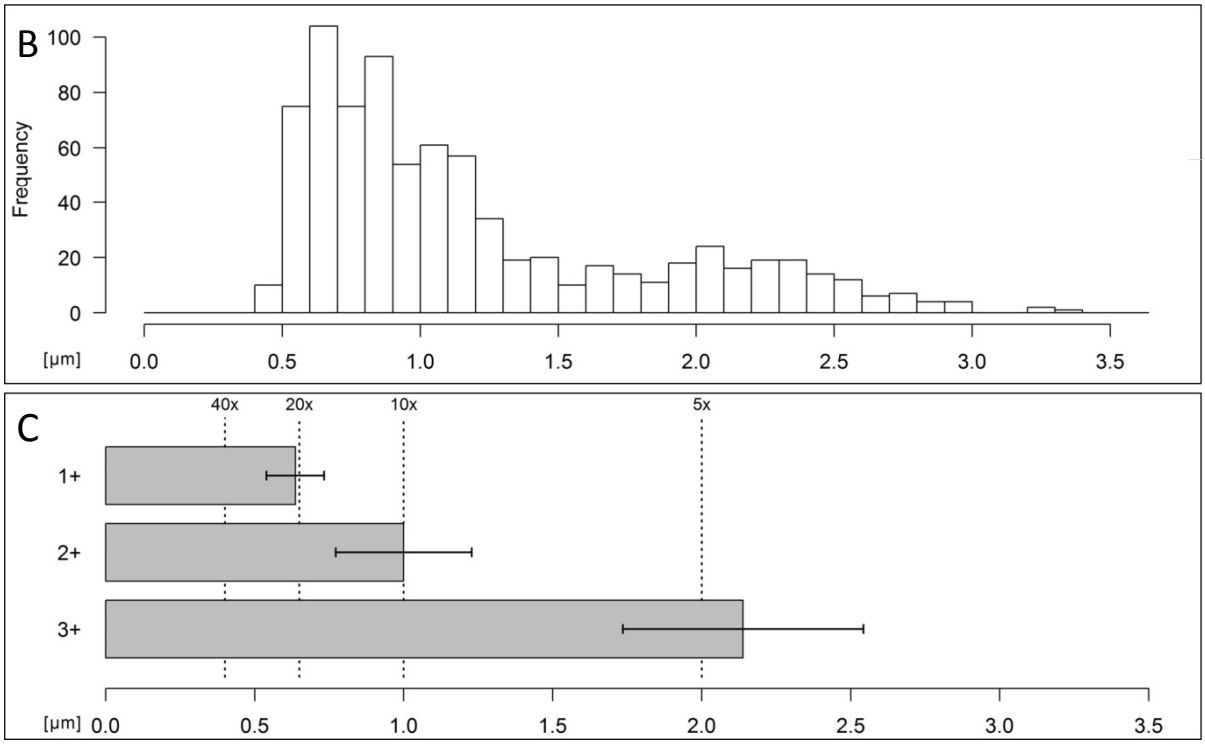


Figure 3.9: Physical background of magnification rule for the interpretation of Her2 IHC. A: Numerical aperture and optical resolution of common microscope objectives; Her2 scoring categories and measured width of the DAB precipitates. B: Frequency distribution of DAB width in µm. C: Bar plots of the average DAB widths of the different Her2 scoring categories (antennae: SD). Dashed lines indicate the average optical resolution of the indicated microscope objectives

3.4.2 DAB width and staining intensity are correlated

Manual Her2 scoring includes both the staining pattern and the staining intensity. Correspondingly, the Her2 DAB staining intensity was also investigated in the subset of n = 120 cases. Figure 3.10 shows a scatter plot of DAB width as detailed in figure 3.9 against the staining intensity. Interestingly, the scoring categories 1+, 2+ and 3+ form well separated clouds. The clouds show little overlaps and differ both in staining intensity and DAB width. In the 1+ and 2+ cases, a linear relation is obvious and Pearson's correlation coefficient is r = 0.73. In the 3+ category, the staining intensity is nearly 100%, i.e. is saturated and cannot increase further. This matches the subjective observation and Her2 1+ and 2+ cases appear brownish while 3+ cases are dark brown to black. However, the cases within the 3+ category do show differences in DAB width that range from 1.5 µm to 3.5 µm. Overall, the shape of the scatter plot resembles the typical staining behaviour of a DAB-based IHC assay, which shows a sigmoid curve (Figure 1B). A certain protein concentration is required for a staining to form, increasing protein concentrations show a linear correlation with the resulting staining for a certain range until saturation sets in (Dabbs 2018).

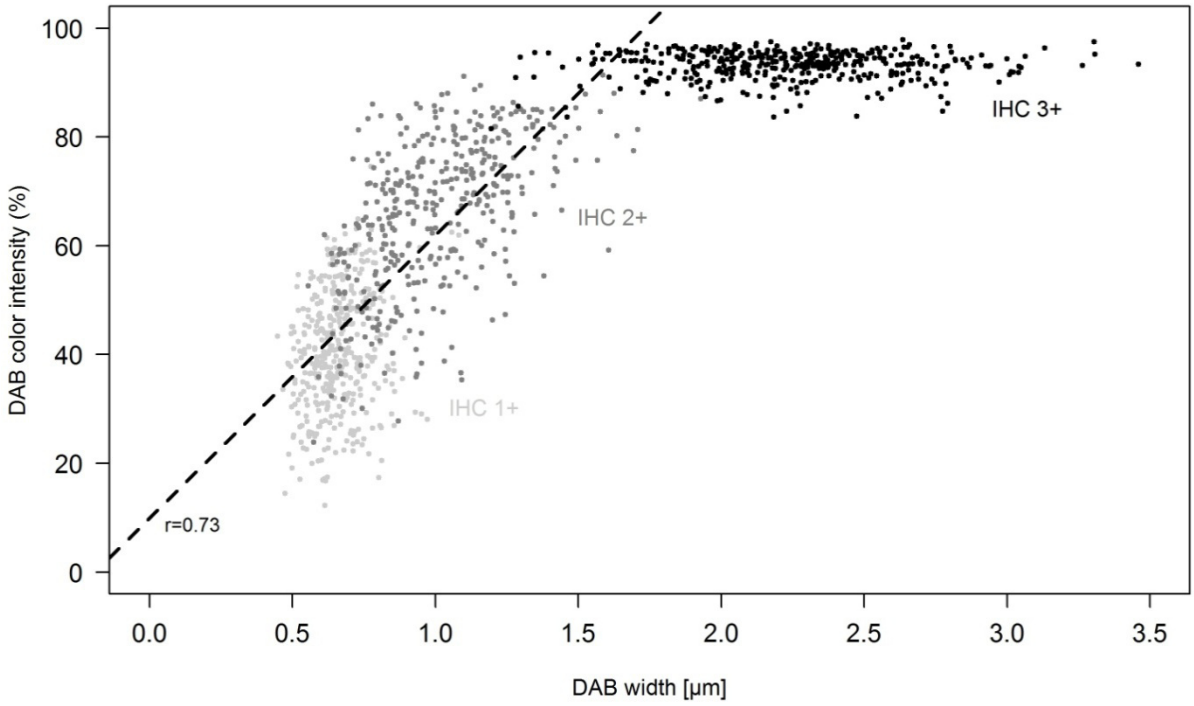


Figure 3.10: Scatterplot of DAB width in µm and DAB staining intensity (relative staining intensity 0 = white, no staining to 100% = black, maximal staining). For 1+ and 2+ cases, width and intensity show a strong linear correlation with Pearson's r = 0.73. For 3+, the intensity is saturated and cannot increase further. However, differences in DAB width among the 3+ cases are evident.

3.4.3 No correlation between DAB width and amplification status

Figure 3.10 shows differences in DAB width and staining intensity within the IHC 2+ category. This category is considered as equivocal and requires additional ISH testing of the cases to determine the *HER2/neu* gene amplification status. Thus, it was investigated if the detailed subdivision of the cases by image analysis could be used to predict the ISH status:

Figures 3.11 and 3.12 show subanalyses, in which the IHC 2+ cases are divided into ISH positive (+) and ISH negative (-). In figure 3.11, the DAB width and staining intensity are represented as boxplots, which summarise the distributions. The boxes represent the middle 50% of all data points; the lower and the upper ends of the boxes correspond to the lower of upper quartiles of the respective distributions, while the black bar inside the boxes is the median value. The boxes for categories 1+, 2+ (ISH+ and ISH- combined) and 3+ differ significantly and do not overlap for both the DAB width and the DAB staining intensity. However, the two boxes for 2+ (ISH-) and 2+ (ISH+) have very similar shapes and do not differ significantly neither for the width nor for the colour intensity. Figure 3.12 shows a more detailed look at the distributions of the DAB width by histograms and density estimations. The distributions of IHC 2+ (ISH-) and IHC 2+ (ISH+) overlap but show a slight difference in shape. For the ISH- group, the distribution is more right-skewed and peaks around 0.8 µm. The ISH+ group shows a more symmetric distribution with a broader peak around 1 µm. The difference is not significant and cannot be used for meaningful prediction of the ISH status. Overall, analysis of the DAB width and colour intensity corresponds well to the manual IHC scoring categories, but does not allow prediction which IHC 2+ / equivocal cases show *HER2/neu* amplification in ISH.

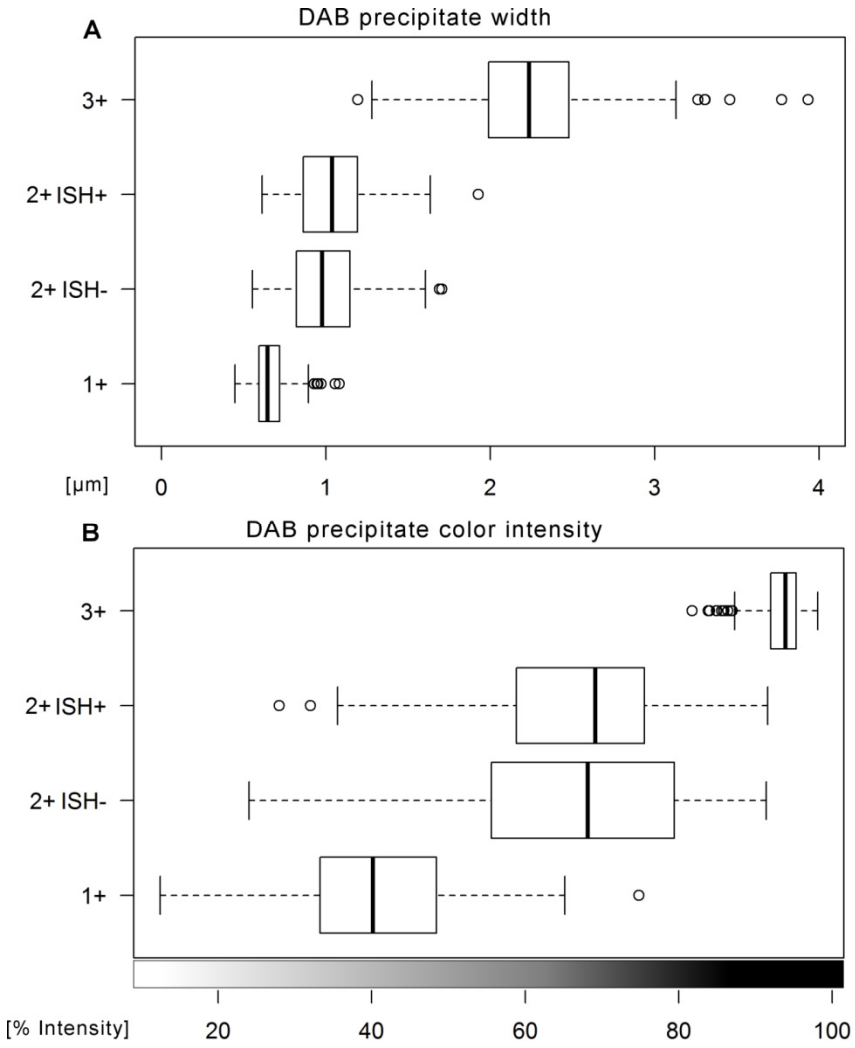


Figure 3.11: Subanalysis of DAB width and DAB intensity in IHC 2+ cases. No difference between IHC 2+ ISH- and IHC 2+ ISH+ cases.

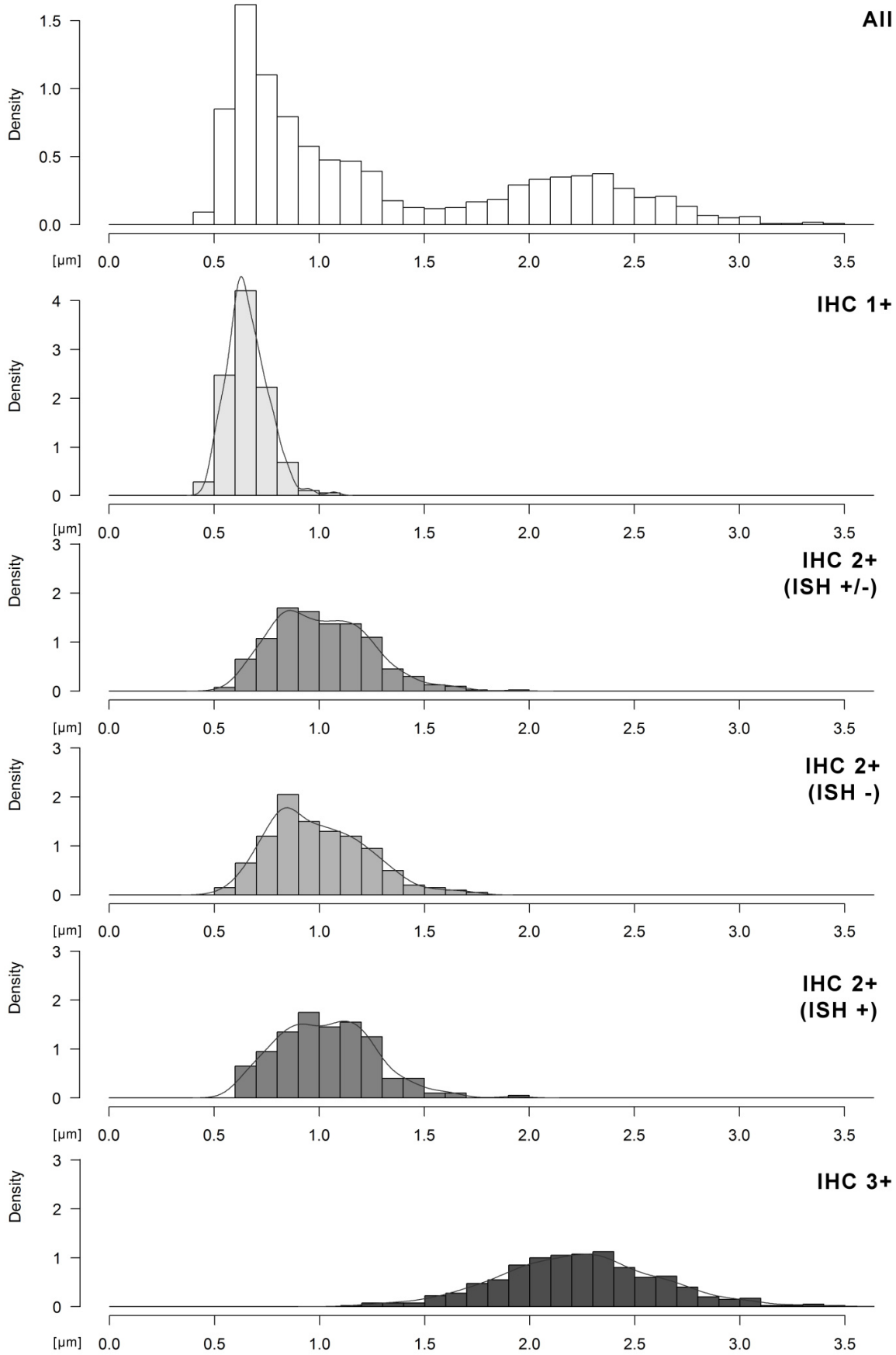


Figure 3.12: Detailed subanalysis of DAB width and DAB intensity in IHC 2+ cases. Diagrams included the respective distributions. Still no difference between IHC2+ ISH- and IHC 2+ ISH+ cases is apparent.

4 Discussion

4.1 Overview

In this study, n = 613 core needle biopsies were reanalysed by whole slide scanning and semi-automated DIA. The data give an indication of the concordance rates between the employed software, *QuantCenter* by 3DHistech, and manual scoring. The study also addressed more general aspects of DIA that are important in histopathology but also other branches of biomedical research. Furthermore, the digital slides were used to explore the physical basis of manual Her2 interpretation.

Computer vision is a fast-growing, interdisciplinary field with applications in industrial processes, consumer products and biomedical research and diagnostics. The automated interpretation of digital images is used in a multitude of settings including the automated monitoring of manufacturing processes, facial recognition for identity verification, self-driving cars but also assisted interpretation of x-ray images and histological specimens (Forsyth and Ponce 2015). Digital cameras and powerful computers have become widely spread and affordable which facilitates practical application. In dentistry, computer vision is clinically used for some applications and investigated for various purposes:

In prostodontics, computer assisted design / computer assisted manufacturing (CAD/CAM) augments clinical procedure by image analysis. Dental restoration may be achieved by capturing the intraoral situation via optical scanning and either using a 3D-printed analogue model on which the dental technician creates an analogue denture or by using digital 3D models and digital manufacturing processes in combination with a dental milling unit. Either way the digitalisation of these processes reduces possible mistakes caused by analogue dental impressions, plaster models, casting compound etc. Multi-unit bridges and other implant-supported solutions would not be possible at today's fitting precision and high quality without this new technology.

Digital photos on the other hand facilitate documentation for medical and forensic purposes. Documentation systems to monitor suspicious intraoral or cutaneous lesions may improve the clinical follow-up and can be used to create a time lapse. The images can also be augmented by automated image interpretation.

In the interpretation of dental radiographs, software solutions were validated for the detection of early caries and may be more accurate than manual interpretation (Cantu et al. 2020). Image analysis software using convolutional neural networks have been adopted successfully in research settings (Schwendicke et al. 2019).

As clinical applications of image analysis increase, the importance of understanding basic concepts, advantages, limitation and potential sources of errors grows.

4.2 Patients and samples

The average age of the re-analysed patients was 62.8 years, which is very close to the national average of 64 years for female patients (RKI 2015). The likelihood of breast cancer increases with age and is highest in the seventh decade. In Germany, the mammography screening program is recommended for patients aged 50-69 years, which covers the peak in incidence. Most cases were UICC-stage I, i.e. the primary tumours' largest diameter was below 2 cm (pT1) and no lymph node metastases were present (pN0). The preponderance of early-stage breast cancer is in agreement with national reference data (RKI 2015). The majority of cases were histologically categorised as NST. Only 11% were categorised as lobular subtype and less than 5% as subtypes other than NST and lobular. Overall, the patient cohort seems representative of breast cancer in Germany.

The re-analysed samples were processed at one pathological institution within one year by using standardised procedures. Immunohistochemistry was performed on an automated staining platform (Ventana Benchmark XT). This is of importance since inter-laboratory variability is a well-known phenomenon. The resulting staining in immunohistochemistry depends on a variety of parameters. Subtle differences in tissue processing, cutting thickness, treatment times, room temperature and humidity may influence the enzymatic staining reaction and cause differences between laboratories. Using samples from only one lab minimises such influences.

However, not all parameters that may influence staining results could be analysed. The biopsy samples were submitted by several hospitals and medical institutes. Neutral buffered formalin was used for all samples but the fixation times were not recorded at the time the study was performed. Thus, differences caused by pre-analytical factors including sample logistics and fixation times were not analysed.

4.3 Practical aspects of image analysis

Over n = 3000 histology slides were digitised during this study using a Pannoramic P250 whole slide scanner. The average scanning time per slide was 82 seconds. The auto-focus worked very reliably and only few re-scans were required. In summary, image acquisition seems technically mature. Since the beginning of this study, new generations of whole slide scanners were introduced. Scanning speed and sample-handling times further improved. Devices such as the Pannoramic P1000 scanner (3DHistech) allow continuous operation, i.e. slides may be loaded and unloaded while the scanner is running. On the other, the practical implementation of digital pathology into routine diagnostics has several requirements that were also encountered during this study:

The files of scanned slides have to be stored using unique IDs and linked to the clinical cases. The slide-IDs have to be generated in a standardised way. Many databases commonly used in German pathology labs, including software *DC Pathos* and *PathoPro*, use case-specific IDs but not slide-specific IDs. Such software has to be updated or be replaced with newer databases that assign slide-specific IDs. In our study, we used a custom, study-specific table to create and organise the slide-specific IDs. During scanning, the IDs were transferred in csv tables to the scanning software. While this solution minimised preparation efforts of the slides, it is also prone to errors because the scanner does not verify the slide. A minor sample mix-up could cause severe misinterpretations by allocating slides to wrong clinical cases. It would be much safer to label the slides with a code that can be interpreted by the scanner, such as a bar code or data matrix code. The scanner would than verify the slide and generate the file-name from the code. Given the retrospective design, this was not an option for this study because the slides were already labelled.

The digital slides are quite large in terms of storage space requirements. An average biopsy required 0.5-2 Gigabytes (GB, 10^9 bytes) storage; a case of six slides (HE, ER, PR, Her2, Ki-67, E-cadherin) may require more than 10 GB. Storage space has become cost effective with one Terabyte (TB, 10^{12} bytes) costing less than 30 Euros. On the other hand, histopathology slides have to be archived at least 10 years and pathology labs commonly process several ten thousand cases every year. A full digital pathology institute would require a very large storage solution. An alternative would be to save the glass slides for long-term storage and keep the digital slides only for a limited period of time. In this study, external USB 3.0 hard disk drives with 2 TB each were employed. This approach seems appropriate for projects with limited storage requirements. If digital pathology was to be used on a routine basis, a professional storage solution has to be set up.

The results of DIA have to be stored in a local database. Since most pathology laboratories have databases to store diagnostic reports, the DIA data have to be integrated. Depending on the type of existing database, this can be challenging. The *QuantCenter* software outputs one *Excel* table per analysed slide (SI table 1). In this study, a sequence of two scripts was used to automatically transfer the results per case into one table with all results (cf. section 2.7). While this worked satisfactorily for the purposes of this study, it would be optimal if the DIA software would directly upload its results into the database.

The slides have to be inserted into plastic cassettes that are placed in the scanner. Manual handling of hundreds of slides can be time-consuming and may slow down the diagnostic process. New scanners can be adjusted to accommodate a range of different slide-racks. This way, the slides can be inserted into the scanner in the same racks used for H&E or IHC staining. Still,

manual handling of the slides is required at some steps which has to be considered in the overall work-flow.

Similar issues tend to arise whenever DIA is used in a clinical setting (*Berufsverband deutscher Pathologen* 2018). Currently, most hospitals, medical institutions or dental clinics do not have one software but use several different systems for specific tasks. Data integration and allocation can be a considerable obstacle when a new device is set up or if data are shared between institutions. Depending on the type of image recorded and the number of cases, the required storage can rapidly increase. In dentistry, using optical scanners or cameras during routine dental practice can be time-consuming, especially for colleagues who are new to these methods. All of these issues have to be taken into consideration when a new DIA application or device is adopted and comprehensive solutions should be set up in the beginning.

4.4 Human visual perception

A major advantage of DIA is a possible increase in objectivity. The human visual system perceives colours in a context-dependent manner: It is easy to tell if an object is brighter or darker relative to its surroundings. It is impossible to quantify colour in an absolute way. A playful way to illustrate the circumstances are optical illusions. Figure 4.1 demonstrates how perceived brightness depends on the adjacent objects (Figure 4.1). Human colour perception is further complicated by subjective interpretation. Known objects are associated with their usual colours. This may cause misinterpretation of the actual, as is exemplified by figure 4.2.

Our visual system is optimised for everyday life, while many medical settings require robust colour interpretation. Aesthetics are major issue in dentistry and the selection of appropriate colours is of great importance for the patients' and dentists' satisfaction. Traditionally, colour selection is a highly subjective process based on the dentist's and/or dental technician's experience. For whole mouth restorations that usually leads to brilliant results. Yet, the most difficult restorations are single front tooth's crowns. Many factors are known to influence colour-selection, including the surrounding light, colours, time of mouth opening, daytime and even a possible lack of sleep of the practitioner[1] (Paravina Pérez and Ghinea 2019). Here, an objective DIA would be a great advancement to current procedures. The clinical interpretation of colour can also be crucial, e.g. whether a whitish oral lesion is an innocent hyperkeratosis, a fungal overgrowth or a malignant squamous cell carcinoma.

[1] Own expert knowledge

In immunohistochemistry, the staining intensity is of key importance for the correct interpretation. Colour quantification is an area in which DIA can readily contribute important information that are difficult or even impossible to estimate by looking at it.
Measuring colour intensity and hue is also affected by the type and strength of illumination. Slide scanners used for digital pathology are closed devices that shut out external light to ensure equal brightness. The Pannoramic P250 employed in this study uses a flash tube for illumination that emits full-spectrum white light.

4.5 Image analysis of immunohistochemistry

Quantitation of immunohistochemistry by DIA encompasses several steps that are quite universal and apply to any software employed: Determination of the area or cells to quantify, quantitation of the colour precipitate and interpretation of the measurement.

Determining where to measure or which cells to measure is possibly the most difficult step and is subject of ongoing research by computer scientists. Two opposing strategies are possible: 1.) a semi-automated approach that requires an operator / pathologist to manually indicate the area-to-be-analysed, 2.) fully-automated approaches that use advanced algorithms for so-called feature extraction to identify the area and/or cells of interest. The software used in this study contains the module *PatternQuant* that detects cells based on classic algorithm, i.e. methods that search for cell-shaped, round objects based on their geometric shape-descriptors including size and roundness. The values have to be manually entered and can be saved as profiles. This was sufficient to distinguish between most malignant and non-neoplastic bystander cells within the tumour area. It was not sufficient to distinguish between invasive breast carcinoma, pre-invasive lesions / CIS, associated inflammation and adjacent breast tissue. This was not surprising, given the morphologic variability of breast cancer, even within the invasive carcinoma NST group. Thus, we used a combination of both strategies by manually drawing the tumour area and using the algorithms to detect the correct cell type. Optimal settings for *PatternQuant* were established by analysing representative test cases (cf. section 2.6).

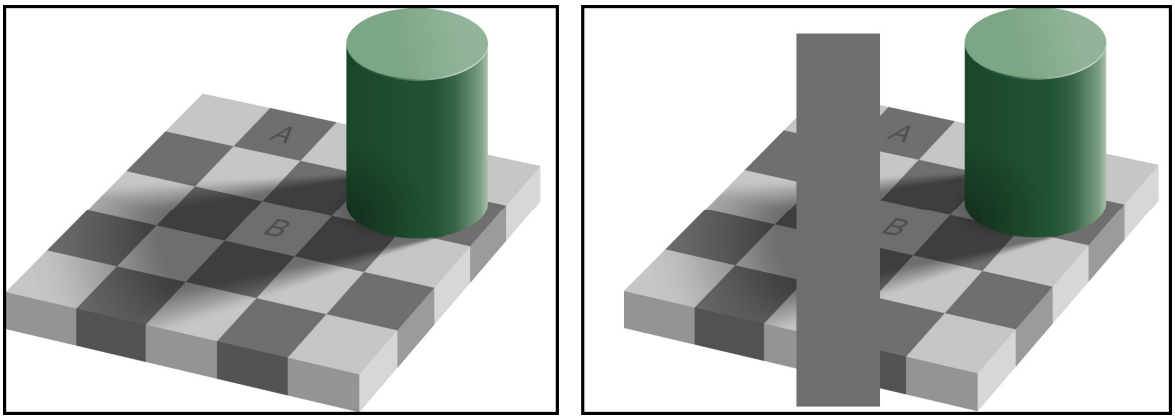
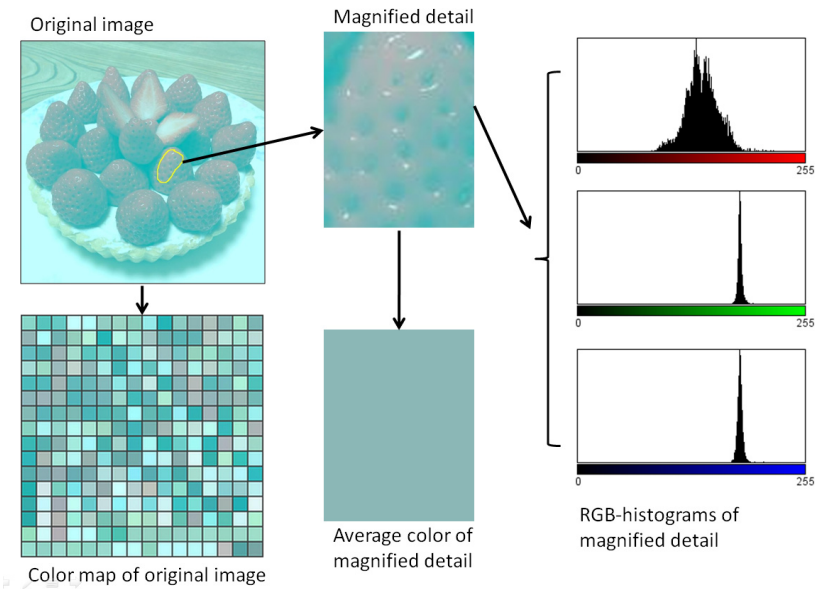


Figure 4.1: The checker shadow illusion. Tiles A and B have the same shade of grey. Left: Tile A appears darker than B. Right: The connecting bar is of the same shade and proves A and B as equal (modified image, original image by Edward H. Adelson, Massachusetts Institute of Technology, USA; used with permission).

Figure 4.2: Turquoise strawberries. The depicted strawberry cake is of turquoise colour and does not contain red hues. While this seems contrary to the perceived colour, image analysis can be used to compute objective results (below). The colour map shows every colour contained in the original image. Whitish, grey and turquoise colours are present, but no red. Calculation of histograms for the different colour channels shows that the image contains a mixture of mostly green and blue. The average colour is a difficult to tell grayish turquoise. (Original image by Akiyoshi Kitaoka (Ritsumaikan University, Kyoto Japan), custom analysis).



Other software that feature fully-automated approaches are Visiopharm (Hida et al. 2020) and *QuPath* (Bankhead et al. 2017). *Visiopharm* relies on the combination of two or three IHC stainings to separate neoplastic and non-neoplastic cells. The technique is called virtual double- / triple-stains, since the stains are on separate slides and are only combined during software analysis. *QuPath* uses a morphology-based technique.

Current approaches to digital pathology use algorithms taken from the field of artificial intelligence (AI), such as machine learning and neural networks. In machine learning, a training set of test data is used that allows an algorithm to learn a certain classification. Unlike the *QuantCenter* approach, the algorithm is not explicitly programmed for the specific task but infers a model from the test data that is used for subsequent classification. Machine learning may be based on statistical techniques or rely on artificial neural networks. Neural networks are simulated neural circuits that are trained to perform a specific task. Algorithms based on neural networks have been demonstrated as powerful tools for DIA including histopathology (Niazi et al. 2019). An impressive demonstration are so-called virtual stains (Rivenson et al. 2020): Unstained tissue section are digitised on a fluorescence scanner. Based only on the autofluorescence of the tissue, AI algorithms can interpolate common chemical stains such as H&E and PAS. Virtual and actual H&E stains are indistinguishable.

Quantitation of the colour precipitate is relatively simple once the correct cells of interest have been identified. ER, PR and Ki-67 are nuclear markers, meaning the entire nucleus of the cell is measured. Her2 creates a membranous pattern and interpretation is based on a combination of colour intensity and shape. In breast cancer, the staining has to be membranous-circular. In gastric cancer, a baso-lateral staining of the cells is also considered as positive.

More challenging is the interpretation of the measurements. On its own, a DIA software will only output numerical values indicating how many cells show which staining intensity and distribution of the IHC-chromogen. To make these results clinically relevant, they have to be translated into meaningful categories. In our study, the results were translated into the scoring categories used for manual scoring, i.e. the Allred-Score for ER and PR, the four-step IHC-score for Her2 and the Ki-67 index (cf. section 2.4). This translation was set up using small sets of test cases for each biomarker. The test cases were selected to represent the dynamic range of each marker, i.e. negative cases, weakly and strongly positive cases. The calculated concordance coefficients thus reflect the technical inter-method agreement of manual scoring and DIA. In other words, they reflect how closely DIA can mimic manual interpretation.

4.6 Manual scoring versus image analysis

To quantify the level of agreement between DIA and manual scoring, Cohen's kappa coefficient was calculated (cf. section 2.7). The kappa coefficient ranges from 0 to 1. The values were interpreted according to Landis & Koch (Landies and Koch 1977), who recommend five categories: Slight ($\kappa < 0.2$), fair (0.2-0.4), moderate (0.4-0.6), substantial (0.6-0.8) and almost perfect ($\kappa \geq 0.8$). The results for the training data and the entire cohort are summarised in table 4.1 (Table 4.1).

Table 4.1: Summary of concordance analyses

Biomarker:	ER			PR			Ki-67			Her2		
Cases:	Training cohort, n=16		All cases, n=613	Training cohort, n=15		All cases, n=613	Training cohort, n=16		All cases, n=613	Training cohort, n=19		All cases, n=613
Software-Settings:	Standard	Optimised	Optimised	Standard	Optimised	Optimised	Standard	Optimised	Optimised	Standard	Optimised	Optimised
Cohen's Kappa:	0.40	0.86	0.77	0.32	1.00	0.71	< 0.1	0.48	0.68	0.74	0.74	0.55

Table 4.1: The table summarises the concordance between manual scoring and DIA, calculated by Cohen's kappa coefficient. The results for ER, PR, Ki-67 and Her2 are shown. For each marker, a training set of cases was used to optimise the software settings. The concordance greatly increased compared to the standard settings. Subsequently, all cases were analysed using the optimised settings.

For ER and PR, the standard software settings yielded fair concordance rates. Optimisation greatly increased the concordance and the analysis of all cases indicated substantial concordance. The obtained distributions are in-line with published data (Jeon, Kim and Kim 2021). For ER, the observed bimodal distribution of the IHC scoring results is typical (Collins, Botero and Schnitt 2005). DIA of hormone receptors seems ready to use.

For Ki-67, additional adjustments to the interpretation of DIA were required. It turned out that DIA systematically identified more cells as positive compared to manual scoring. Accordingly, the concordance increased if different cut-offs were applied to manual scoring and to DIA. The combination ≥ 20% (manual) / ≥ 27% (digital) showed a substantial concordance of $\kappa = 0.68$ (cf. section 3.3.2). The systematic difference is not surprising, given the long pre-history of Ki-67 standardisation (Nielsen et al. 2020, Dowsett et al. 2011). It has been a matter of debate for decades how to interpret Ki-67 and how to maximise interobserver-reproducibility. Different cut-offs to define breast cancer cases as Ki-67 high and low have been proposed but it is not just a matter of interpretation: Ki-67 IHC stainings show variability between different laboratories (Polley et al. 2013). There are multiple factors that contribute to such interlaboratory variance, including the choice of primary antibody, differences in IHC staining protocols and different laboratory devices (Leung et al. 2016) as well as counting methods (Polley et al. 2013). While using the same cut-off at different institutions has limited reproducibility, the classification of

breast cancer as Ki-67 high and low is suprisingly reliable (AWMF 2020). Thus, current guidelines recommend the evaluation of Ki-67 and the classification as high or low but do not define a mandatory cut-off. It seems plausible that DIA based classification runs into similar issues that can be overcome by adjusting the threshold.

Her2 was the most challenging biomarker in this study. The standard settings yielded substantial concordance. Optimisation using a training cohort of n = 16 cases increased the number of detected cells and changed the classification of several cases. The likelihood of a false-negative classification was reduced. However, several false-positive cases remained and the overall concordance did not increase. Depending on the method of concordance analysis, the complete cohort of n = 613 cases achieved fair to moderate concordance. Two measures improved the concordance: First, the interpretation of the DIA data was modified and the h-score was introduced as intermediate step. This increased concordance from $\kappa = 0.43$ to $\kappa = 0.48$. Second, the IHC-score categories 0 and 1+ were combined, which increased kappa to 0.55. Since categories 0 and 1+ are both considered as negative, the distinction is not clinically relevant. Overall, the level of agreement between DIA and manual scoring remained unsatisfactory despite various optimisation efforts (cf. section 3.3.3).

In the current study, manual Her2 scoring was performed by several pathologists. Each case was reviewed by one pathologist. The data were collected during routine diagnostic work-up of the samples prior to the study. To further investigate the disagreement between software and pathologist in a subsequent study, the manual scoring could be repeated and a consensus score by a group of pathologists could be determined. Comparing DIA and consensus manual score could provide clues on the origin of discordant cases. Another approach to a subsequent study would be to test different training sets, since the interpretation of DIA data relies on the cases the software has been trained with. Using different training sets that were manually scored by just one pathologist, might reveal if disagreement reflects differences between software and pathologists or between pathologists. However, interobserver analyses were not in the scope of the current study.

4.7 Comparisons to published studies

For the hormone receptors, studies have been published that show excellent concordance between manual scoring and DIA as well as between DIA and molecular quantitation of receptor expression (Jeon, Kim and Kim 2021). Such studies highlight not only the technical maturity of DIA for hormone receptors but also its biological significance.

Table 4.2: Key studies into image analysis of Her IHC

Study	Software	Cases (n)	Her2 IHC (digital vs. manual)	Her2 IHC (digital) vs. ISH	Her2, outcome
Laurinaviciene 2011	Visiopharm	195 (TMAs)	κ = 0.80	ρ = 0.67	
Brügmann 2012	Visiopharm	253 (TMAs)	κ = 0.86	Sensitivity: 99.2% Specificity: 100%	
Holten-Rossing 2015	Visiopharm	462 (TMAs)	κ = 0.74	κ = 0.74	
Hartage 2020	Visiopharm	612 (biopsy: 395 resection: 217)	κ = 0.71	ρ = 0.86	
Li 2020	Visiopharm	153 (resection)		ρ = 0.54	Prediction of pCr
Bankhead 2018	QuPath	293 (TMAs)	ρ = 0.84		Prediction of survival

Table 4.2: Key studies into image analysis of Her2 IHC compared to manual interpretation of Her2 IHC and *HER2/neu* ISH.

Her2 is more challenging to interpret. In a first step, it was demonstrated that manual Her2 interpretation using optical microscopes and whole slide scanning has very high levels of concordance (Nunes et al. 2014). Nunes et al. showed that both Her2 IHC and brightfield in situ hybridisation of the *HER2/neu* gene can be interpreted on computer screen with similar sensitivity and specificity compared to conventional microscopy. Three different IHC assays were used. The study indicates that scanned slides contain all information that are necessary for the successful interpretation of Her2 IHC and ISH. In a second step, studies on DIA of Her2 have demonstrated substantial concordance both to manual Her2 IHC interpretation and *HER2/neu* gene copy number in FISH analysis (Table 4.2). However, the level of concordance is lower compared to the analysis of hormone receptors and specific algorithms are required. The most common approach is based on the so-called membrane connectivity score. It was introduced in 2011 (Laurinaviciene et al. 2011) and has subsequently been validated in several studies using hundreds of cases (Brügmann et al. 2012; Holten-Rossing et al. 2015; Hartage et al. 2020). A study with breast cancer patients undergoing neoadjuvant treatment showed that Her2-DIA can be the best parameter to predict pathological complete response, illustrating the high relevance of DIA for actual clinical treatment (Li et al. 2020). Besides commercial software solutions, excellent results were also achieved with the free to use software platform *QuPath* (Bankhead et al. 2017).

In summary, Her2-DIA is feasible, shows satisfying correlations with *HER2/neu* gene copy number and has a very high predictive value. An important reason for the different concordance rates of the current study and published data might the choice of cases and materials: In the literature, most studies used tissue microarrays (TMAs) while this study used real world biopsy specimen. TMAs are generated by combining resection specimen tissue of several patients into one paraffin block. It is a standard technique used to facilitate serial analyses. However, it may cause a positive selection of cases that have large, easy to spot tumour infiltrates which may in turn improve DIA results. In line with this view, the study by Hartage et al. (2020) used a combination of biopsy and

resection specimens and showed lower concordance (κ = 0.71) compared to the studies that used TMAs (κ = 0.74-0.86) (Hartage et al. 2020). It might be that real world cases show lower concordance rates compared to studies that reanalysed selected cases.

Table 4.3: Key studies into DIA of Ki-67

Study	Software	Cases (n)	Concordance-analysis	Outcome
Abubakar 2016	Ariol	8088	DIA vs. survival	Prognostic for survival
Stålhammar 2016	Visiopharm	436	DIA / manual vs. survival	Prognostic for survival
Healey 2017	Definiens	2653	DIA vs. manual	Prognostic for survival
Stålhammar 2018	Visiopharm	294	DIA / manual / mitotic count vs. survival	Prognostic for survival
Acs 2019	QuantCenter, QuPath	179	inter-software inter-operator	Prognostic for survival
Robertson 2020	Visiopharm, QuPath	217	Hot spot vs. global, DIA vs. Expression-analysis (PAM50)	Prognostic for survival
Hida 2020	Visiopharm	413	DIA / manual vs. survival	Prognostic for survival
Paik 2021	QuPath	240	DIA vs. Expressio-analysis (OncotyeDx)	Prognostic for survival
Aung 2021	Visiopharm, QuPath		Cell lines as reference standards technical aspects; inter-laboratory	

Table 4.3: The software used for Ki-67 quantification, number of analysed cases and type of concordance analysis are stated. All studies confirmed the prognostic value of Ki-67.

Studies on Ki-67-DIA mostly avoid comparisons between DIA and manual interpretation. Instead, DIA was used to stratify survival and demonstrate the prognostic value of Ki-67. This is in line with the observation of the current study, that manual interpretation and DIA show systematic differences and that using different cut-offs for both methods improves concordance.

Given the difficulties in Ki-67 interpretation, some authors questioned if Ki-67 was prognostic at all (Dowsett et al. 2011; Nielsen et al. 2020). A large systematic study re-investigated over 8000 breast cancer cases using the DIA system *Ariol* and showed prognostic value of Ki-67 for ER-positive breast cancer (Abubakar et al. 2016). While the *Ariol* system has not found widespread application, the data are among the best examples to demonstrate the prognostic value of Ki-67. Many other studies made similar observations and the prognostic properties of Ki-67 are now considered proven (Table 4.3). Successful studies were performed with different DIA software including the commercial products *Definiens Tissue Studio* and *Visiopharm* as well as free to use software *QuPath*.

Besides the prognostic value, several other aspects of Ki-67 scoring were investigated by DIA. By using a selection of reference cases, Acs et al. investigated inter-platform concordance of three different Ki-67 DIA software as well as inter-operator concordance. The software *Visiopharm* and *QuPath* showed excellent concordance and prognostic values. Inter-operator concordance was also high but might be a bigger source of variability than the choice of software (Acs et al. 2019).

Manual scoring is based on hot spot analysis, i.e. the number of positive cells counted in the region with the highest expression. DIA enables global analysis of all positive cells. Robertson et al. (2020) demonstrated that such global Ki-67 scoring may outperform hot spot analysis: Global scoring was shown to be an independent prognostic marker in multivariate analysis with good concordance to the expression-based prognostic test PAM50. On the other hand, Paik et al. showed that Ki-67 DIA has limited concordance with the expression-based prognostic test OncotypeDx and found higher concordance for hot spot analysis (Paik et al. 2020). Overall, Ki-67 DIA was demonstrated as strong and independent prognostic factor in ER-positive breast cancer and may outperform manual interpretation. While some questions remain to be addressed, clinical utility does seem to be proven.

4.8 Physical basis of Her2 scoring; the magnification rule

As previously stated, the human visual system perceives colours and brightness in a relative fashion that may interfere with IHC interpretation. However, humans excel at recognition of geometrical shapes, lines and other features. This capability is hardwired by several neural mechanisms, including lateral inhibition in the retina and neurons in the optical brain cortex that are activated by specific optic patterns. Accordingly, IHC interpretation protocols that rely on colour are prone to variability while IHC interpretation protocols that rely on patterns and/or shape are more robust.

The so-called magnification rule for the interpretation of Her2 IHC was instrumental in raising interobserver concordance. It relates the Her2 IHC-score to the required magnification and microscope objective. If a higher magnification is required to observe the membranous, linear Her2 staining, the IHC-score is lower. We performed additional DIA on the digitised breast cancer cases to investigate the underlying opto-physical relations (cf. section 3.4).

Interestingly, image analyses by *ImageJ* revealed a correlation between DAB staining intensity and width of the DAB colour precipitates. A stronger Her2 staining is not only darker but also wider. Accordingly, the correct IHC-score can be derived not only by assessing the colour, but also by assessing the shape of the precipitates. This interrelation explains the functioning and the success of the magnification rule. It is not limited to Her2 IHC but could possibly be used for other membranous biomarkers as well. However, one limitation remains: Even with our exploratory analysis, we were not successful to predict the result of the FISH test in Her2 2+ cases. Such cases are considered as equivocal and are re-tested using in situ hybridisation to determine the *HER2/neu* gene amplification status. The additional FISH test is time-consuming and produces extra-costs. Other studies using more recent software are showing promising results that Her2 DIA, especially using the membrane connectivity score, may lower the number of 2+ cases and

decrease the number of required FISH tests (Brügmann et al. 2012; Holten-Rossing et al. 2015). Fewer 2+ cases were demonstrated not only for TMA-studies but also for biopsy and resection specimens (Hartage et al. 2020).

4.9 Reference standards

In this study, the employed DIA software *QuantCenter* required specific settings to achieve reasonable concordance to manual IHC scoring. The pre-defined standard settings did not result in acceptable concordance. The specific settings were derived from small, representative subsets of cases. This principle is quite universal to various DIA approaches: The software has to be adjusted at each laboratory. This raises the question, which cases should be used to derive the specific settings, i.e. which cases are considered as reference standard. In this study, representative cases were selected by their manual score. By this approach, the best possible result would be perfect concordance between DIA and manual scoring. While inter-method reproducibility is an important hallmark for a new technique, the ultimate purpose of biomarkers is not concordance but prediction of clinical benefit and/or prognosis of the course of disease. DIA differs from manual scoring because software can quantify colours in absolute numbers, which the human optical system cannot. It could be possible that DIA of IHC based biomarkers could yield more accurate predictions and/or prognosis. To investigate this question, different reference standards would have to be used with known clinical benefit to anti-hormonal treatment, anti-Her2 treatment or cytotoxic chemotherapy and with known long-term course of disease. Archived biomaterials from clinical cases are limited and cannot be used for multi-institutional calibration of DIA software. A possible solution could be cell-lines with defined biomarker-expression. Cell-lines can be grown in tissue-cultures for prolonged periods of time. Cells may be genetically engineered to express biomarkers at specific levels. Thus, they may constitute excellent tools to enhance DIA standardisation at multiple sites. Indeed, a recent study by Aung et al. used cell-lines with defined Ki-67 status to systematically identify concordant and discordant Ki-67 IHC assays (Aung et al. 2021). Such standardised reference materials may improve reproducibility of Ki-67 IHC staining, while image analysis may improve IHC interpretation.

5 Summary

Breast cancer is the leading form of cancer in women in Germany with about 69.000 new cases anually and a lifetime risk of 12.9%. One in eight women will develop a malignant neoplasm of the breast. Early diagnosis measures include regular clinical examinations, mammography and biopsies of suspicious lesions. Definite diagnosis is based on histopathology. Pathologic work-up encompasses histomorphology on H&E stained slides and a set of four biomarkers that provide prognostic information about the expected course of disease and predictive information about the likeliness to benefit from different clinical treatments. In breast cancer, immunohistochemistry is currently the most common type of biomarker. Immunohistochemistry conventionally relies on manual histological interpretation, but automated techniques based on image analysis have become increasingly available.

In the present study, n = 613 breast cancer core needle biopsies from a single pathological laboratory (*Pathologie Nordhessen*, Kassel, Germany) were re-analysed by whole slide scanning of the histological specimens and image analysis of the biomarkers estrogen receptor, progesterone receptor, Her2 receptor and Ki-67 by the software package *QuantCenter* (3D Histech). The results were compared to manual biomarker interpretation by board-certified pathologists.
Digitisation of the histological slides by a state-of-the-art tile scanner (3D Histech Pannoramic P250 Flash II) required 82 seconds per slide on average (standard deviation: ± 38s) and seemed technically mature. Allocation and storage of the large files constitute major issues that require costumised solutions. Image analysis did not work with out-of-the-box settings but required optimisation on local cases. After training of the software, satisfying rates of concordance were achived for estrogen and progesterone receptors with Cohen's kappa coefficients of $\kappa = 0.86$ and $\kappa = 1.0$. In Ki-67, systematic differences between manual scoring and image analysis were noticed and the best concordance achieved was $\kappa = 0.68$. Her2 yielded a good concordance of $\kappa = 0.74$ in a training set of n = 19 representative cases but only a moderate concordance of $\kappa = 0.55$ in the complete cohort. Exploratory analysis of Her2 yielded additional information on the physical basis of manual Her2 scoring.

The findings indicate that image analysis is a mature technique that can be used to supplement the analysis of biomarkers in breast cancer. Image analysis has potential to decrease interobserver variance and to allow more precise quantitation. Yet, current software approaches require specific optimisation on local cases. The achieved concordance results from the representativeness of these training cases, which raises the question of how to define such

reference standards. A possible solution could be centrally defined testing materials, for example tissue cultures with fixed levels of biomarker expression, that could be used for standardised local optimisation.

6 Appendix

Supplementary table 1: Example of output-table created by the *QuantCenter* image analysis software

Slide:	H00138-12_ER	
Parameter	<unnamed roi>	
Annotation area	0,28	mm2
Mask area	0,05	mm2
Total count	717	pcs
Relative mask area	16,23	%
Object frequency	2518,91	pcs/mm2
Positivity Index	79,78	%
Positive mask area	83,4	%
Average positivity of Negative	158,27	
Average positivity of Weak positive	106,72	
Average positivity of Medium positive	78,67	
Average positivity of Strong positive	52,08	
Negative	145	pcs
Weak positive	145	pcs
Medium positive	188	pcs
Strong positive	239	pcs
Negative	20,22	%
Weak positive	20,22	%
Medium positive	26,22	%
Strong positive	33,33	%
Score	8	
Intensity Score	3	
Proportion Score	5	
H-Score	172,66	
Weak positive Index	25,35	%
Medium positive Index	32,87	%
Strong positive Index	41,78	%

Supplementary table 1: The table contains the results of the image analysis of estrogen receptor in case H00138. The data are saved in the csv file format.

7 Literature

Abubakar M, Orr N, Daley F, Coulson P, Ali HR, Blows F, Benitez J, Milne R, Brenner H, Stegmaier C 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(1),104

Acs B, Pelekanou V, Bai Y, Martinez-Morilla S, Toki M, Leung SCY, Nielsen TO, Rimm DL (2019): *Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study.* Lab Invest 99(1),107-117

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) (Eds): *Interdisziplinäre S3-Leitlinie für die Früherkennung, Diagnostik, Therapie und Nachsorge des Mammakarzinoms*. Langversion 4.3, register-number 032-045OL. Deutschen Krebshilfe (DKH), Bonn 2020.
https://www.leitlinienprogramm-onkologie.de/fileadmin/user_upload/Downloads/Leitlinien/Mammakarzinom_4_0/Version_4.3/LL_Mammakarzinom_Langversion_4.3.pdf
Accessed on 2020-07-05

Aung TN, Acs B, Warrell J, Bai Y, Gaule P, Martinez-Morilla S, Vathiotis I, Shafi S, Moutafi M, Gerstein M et al. (2021): *A new tool for technical standardization of the Ki67 immunohistochemical assay*. Mod Pathol 34(7),1261-1270

Badve S, Kumar GL (Eds.): *Predictive Biomarkers in Oncology: Applications in Precision Medicine.* 1st Edition; Springer-Verlag, Berlin 2018

Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, McQuaid S, Gray RT, Murray LJ, Coleman HG et al. (2017): *QuPath: Open source software for digital pathology image analysis*. Sci Rep 7(1),16878

Bankhead P, Fernández JA, McArt DG, Boyle DP, Li G, Loughrey MB, Irwin GW, Harkin DP, James JA, McQuaid S, Salto-Tellez M, Hamilton PW (2018): *Integrated tumor identification and automated scoring minimizes pathologist involvement and provides new insights to key biomarkers in breast cancer.* Lab Invest 98(1),15-26

Berufsverband deutscher Pathologen (Eds): *Leitfaden Digitale Pathologie in der Diagnostik – Befunderstellung an digitalen Bildern.* 1st Edition; Berufsverband deutscher Pathologen e.V., Berlin 2018.
https://www.pathologie.de/pathologie/digitale-pathologie/?eID=downloadtool&uid=1734
Accessed on 2021-02-15

Brügmann A, Eld M, Lelkaitis G, Nielsen S, Grunkin M, Hansen JD, Foged NT, Vyberg M (2012): *Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains.* Breast Cancer Res Treat 132(1),41-9

Cameron D, Piccart-Gebhart MJ, Gelber RD, Procter M, Goldhirsch A, de Azambuja E, Castro G Jr, Untch M, Smith I, Gianni L et al. (2017): *11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial.* Lancet 389(10075),1195-1205

Campbell WS, Lele SM, West WW, Lazenby AJ, Smith LM, Hinrichs SH (2012): *Concordance between whole-slide imaging and light microscopy for routine surgical pathology.* Hum Pathol 43(10),1739-44

Campbell WS, Hinrichs SH, Lele SM, Baker JJ, Lazenby AJ, Talmon GA, Smith LM, West WW (2014): *Whole slide imaging diagnostic concordance with light microscopy for breast needle biopsies.* Hum Pathol 45(8),1713-21

Cantu GC, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, Elhennawy K, Schwendicke F (2020): *Detecting caries lesions of different radiographic extension on bitewings using deep learning*. J Dent 100(103425)

Cianfrocca M, Goldstein LJ (2004): *Prognostic and predictive factors in early-stage breast cancer*. Oncologist 9(6),606-16

Collins LC, Botero ML, Schnitt SJ (2005): *Bimodal frequency distribution of estrogen receptor immunohistochemical staining results in breast cancer: an analysis of 825 cases*. Am J Clin Pathol 123(1),16-20

Dabbs DJ (Ed.): *Diagnostic Immunohistochemistry: Theranostic and Genomic Applications.* 5th Edition; Elsevier Publishers, Amsterdam 2018

Davies C, Godwin J, Gray R, Clarke M, Cutter D, Darby S, McGale P, Pan HC, Taylor C, Wang YC et al. (2011): *Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials.* Lancet 378(9793),771-84

Dieci MV, Radosevic-Robin N, Fineberg S, van den Eynden G, Ternes N, Penault-Llorca F, Pruneri G, D'Alfonso TM, Demaria S, Castaneda C et al. (2018): *Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer.* Semin Cancer Biol 52(2),16-25

Dietel M, Klöppel G (Eds.): *Pathologie. Mamma, Weibliches Genitale, Schwangerschaft und Kindererkrankungen*. 3rd Edition; Springer-Verlag, Berlin 2013

Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, Ellis M, Henry NL, Hugh JC, Lively T et al. (2011): *Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group.* J Natl Cancer Inst 103(22),1656-64

Fletcher DM (Ed.): *Diagnostic Histopathology of Tumors*. 5th Edition; Elsevier Publishers, Amsterdam 2020

Forsyth DA, Ponce J: *Computer Vision: A Modern Approach*. 2nd Edition; Pearson Education Publishers, London 2015

Gerdes J, Schwab U, Lemke H, Stein H (1983): *Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation.* Int J Cancer 31(1),13-20

Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, Senn HJ; Panel members (2013): *Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013.* Ann Oncol 24(9),2206-23

Gottardis MM, Robinson SP, Satyaswaroop PG, Jordan VC (1988): *Contrasting actions of tamoxifen on endometrial and breast tumor growth in the athymic mouse*. Cancer Res 48(4),812-5

Hartage R, Li AC, Hammond S, Parwani AV (2020): *A Validation Study of Human Epidermal Growth Factor Receptor 2 Immunohistochemistry Digital Imaging Analysis and its Correlation with Human Epidermal Growth Factor Receptor 2 Fluorescence In situ Hybridization Results in Breast Carcinoma.* J Pathol Inform 4(11),2

Harvey JM, Clark GM, Osborne CK, Allred DC (1999): *Estrogen Receptor Status by Immunohistochemistry Is Superior to the Ligand-Binding Assay for Predicting Response to Adjuvant Endocrine Therapy in Breast Cancer*. J Clin Oncol 17(5),1474-81

Healey MA, Hirko KA, Beck AH, Collins LC, Schnitt SJ, Eliassen AH, Holmes MD, Tamimi RM, Hazra A (2017): *Assessment of Ki67 expression for breast cancer subtype classification and prognosis in the Nurses' Health Study*. Breast Cancer Res Treat 166(2),613-622

Hida AI, Omanovic D, Pedersen L, Oshiro Y, Ogura T, Nomura T, Kurebayashi J, Kanomata N, Moriya T (2020): *Automated assessment of Ki-67 in breast cancer: the utility of digital image analysis using virtual triple staining and whole slide imaging. Histopathology* 77(3),471-480

Holten-Rossing H, Møller Talman ML, Kristensson M, Vainer B (2015): *Optimizing HER2 assessment in breast cancer: application of automated image analysis.* Breast Cancer Res Treat 152(2),367-75

Humphries MP, Maxwell P, Salto-Tellez M (2021): *QuPath: The global impact of an open source digital pathology system.* Comput Struct Biotechnol J 19,852-859

Ingold Heppner B, Untch M, Denkert C, Pfitzner BM, Lederer B, Schmitt W, Eidtmann H, Fasching PA, Tesch H, Solbach C et al. (2016): *Tumor-Infiltrating Lymphocytes: A Predictive and Prognostic Biomarker in Neoadjuvant-Treated HER2-Positive Breast Cancer.* Clin Cancer Res 22(23),5747-5754

Jeon T, Kim A, Kim C (2021): *Automated immunohistochemical assessment ability to evaluate estrogen and progesterone receptor status compared with quantitative reverse transcription-polymerase chain reaction in breast carcinoma patients.* J Pathol Transl Med 55(1),33-42

Keum N, Greenwood DC, Lee DH, Kim R, Aune D, Ju W, Hu FB, Giovannucci EL (2015): *Adult weight gain and adiposity-related cancers: a dose-response meta-analysis of prospective observational studies*. J Natl Cancer Inst 107(2),djv088

Klauschen F, Müller KR, Binder A, Bockmayr M, Hägele M, Seegerer P, Wienert S, Pruneri G, de Maria S, Badve S et al. (2018): *Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning.* Semin Cancer Biol 52(2),151-157

Kumar V, Abbas AK, Aster JC (Eds.): *Robbins and Cotran Pathologic Basis of Disease.* 9th Edition; Elsevier Publishers, Amsterdam 2015

Landis JR, Koch GG (1977): *The Measurement of Observer Agreement for Categorical Data.* Biometrics 33(1),159–74

Laurinaviciene A, Dasevicius D, Ostapenko V, Jarmalaite S, Lazutka J, Laurinavicius A (2011): *Membrane connectivity estimated by digital image analysis of HER2 immunohistochemistry is concordant with visual scoring and fluorescence in situ hybridization results: algorithm evaluation on breast cancer tissue microarrays.* Diagn Pathol 23(6),87

Leung SCY, Nielsen TO, Zabaglo L, Arun I, Badve SS, Bane AL, Bartlett JMS, Borgquist S, Chang MC, Dodson A et al. (2016): *Analytical validation of a standardized scoring protocol for Ki67: phase 3 of an international multicenter collaboration.* NPJ Breast Cancer 2,16014

Lewis Phillips GD, Li G, Dugger DL, Crocker LM, Parsons KL, Mai E, Blättler WA, Lambert JM, Chari RV, Lutz RJ et al. (2008): *Targeting HER2-positive breast cancer with trastuzumab-DM1, an antibody-cytotoxic drug conjugate*. Cancer Res 68(22),9280-90

Li AC, Zhao J, Zhao C, Ma Z, Hartage R, Zhang Y, Li X, Parwani AV (2020): *Quantitative digital imaging analysis of HER2 immunohistochemistry predicts the response to anti-HER2 neoadjuvant chemotherapy in HER2-positive breast carcinoma*. Breast Cancer Res Treat 180(2),321-329

Lord CJ, Ashworth A (2017): *PARP inhibitors: Synthetic lethality in the clinic*. Science 355(6330),1152-1158

Millar EK, Browne LH, Beretov J, Lee K, Lynch J, Swarbrick A, Graham PH (2020): *Tumour Stroma Ratio Assessment Using Digital Image Analysis Predicts Survival in Triple Negative and Luminal Breast Cancer.* Cancers (Basel) 12(12),3749

Nelson MH, Dolder CR (2006): *Lapatinib: a novel dual tyrosine kinase inhibitor with activity in solid tumors.* Ann Pharmacother 40(2),261-9

Niazi MKK, Parwani AV, Gurcan MN (2019): *Digital pathology and artificial intelligence. Lancet Oncol* 20(5),e253-e261

Nielsen TO, Leung SCY, Rimm DL, Dodson A, Acs B, Badve S, Denkert C, Ellis MJ, Fineberg S, Flowers M et al. (2020): *Assessment of Ki67 in Breast Cancer: Updated Recommendations from the International Ki67 in Breast Cancer Working Group.* J Natl Cancer Inst 113(7),808-819

Nixon MS, Aguado AS: *Feature Extraction and Image Processing for Computer Vision*. 4th Edition; Academic Press publishers, Cambride 2019

Nunes C, Rocha R, Buzelin M, Balabram D, Foureaux F, Porto S, Gobbi H (2014): *High agreement between whole slide imaging and optical microscopy for assessment of HER2 expression in breast cancer: whole slide imaging for the assessment of HER2 expression*. Pathol Res Pract 210(11),713-8

Paik S, Kwon Y, Lee MH, Kim JY, Lee DK, Cho WJ, Lee EY, Lee ES (2021): *Systematic evaluation of scoring methods for Ki67 as a surrogate for 21-gene recurrence score*. NPJ Breast Cancer 7(1),13

Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, Beckwith BA, Evans AJ, Lal A, Parwani AV; College of American Pathologists Pathology and Laboratory Quality Center (2013): *Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center.* Arch Pathol Lab Med 137(12),1710-22

Paravina RD, Pérez MM, Ghinea R (2019): *Acceptability and perceptibility thresholds in dentistry: A comprehensive review of clinical and research applications*. J Esthet Restor Dent 31(2),103-112

Polley MY, Leung SC, Gao D, Mastropasqua MG, Zabaglo LA, Bartlett JM, McShane LM, Enos RA, Badve SS, Bane AL, Borgquist S et al. (2013): *An international Ki67 reproducibility study*. J Natl Cancer Inst 105(24),1897-906

Rivenson Y, De Haan K, Wallace WD, Ozcan A (2020): *Emerging Advances to Transform Histopathology Using Virtual Staining*. BME Frontiers 2020(9647163),11

Robert Koch-Institut (Eds.): *Krebs in Deutschland für 2015/2016*. 12th Edition; Zentrum für Krebsregisterdaten, Berlin 2019.
https://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/kid_2019/krebs_in_deutschland_2019.pdf?__blob=publicationFile
Accessed on 2020-05-07

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(1),161-175

Robertson JFR, Di Leo A, Johnston S, Chia S, Bliss JM, Paridaens RJ, Lichfield J, Bradbury I, Campbell C (2021): *Meta-analyses of visceral versus non-visceral metastatic hormone receptor-positive breast cancer treated by endocrine monotherapies*. NPJ Breast Cancer 7(1),11

Royal College of Pathologists Working Group (1991): *Pathology reporting in breast cancer screening.* J Clin Pathol 44(9),710-25

Rüschoff J, Lebeau A, Kreipe H, Sinn P, Gerharz CD, Koch W, Morris S, Ammann J, Untch M, Nicht-interventionelle Untersuchung (NIU) HER2 Study Group (2017): *Assessing HER2 testing quality in breast cancer: variables that influence HER2 positivity rate from a large, multicenter, observational study in Germany.* Mod Pathol 30(2),217-226

Scheel AH, Penault-Llorca F, Hanna W, Baretton G, Middel P, Burchhardt J, Hofmann M, Jasani B, Rüschoff J (2018): *Physical basis of the 'magnification rule' for standardized Immunohistochemical scoring of HER2 in breast and gastric cancer.* Diagn Pathol 13(1),19

Schneider CA, Rasband WS, Eliceiri KW (2012): NIH *Image to ImageJ: 25 years of image analysis.* Nat Methods 9(7),671-5

Schwendicke F, Golla T, Dreher M, Krois J (2019): *Convolutional neural networks for dental image diagnostics: A scoping review*. J Dent 91(103226)

Sobecki M, Mrouj K, Camasses A, Parisis N, Nicolas E, Llères D, Gerbe F, Prieto S, Krasinska L, David A et al. (2016): The *cell proliferation antigen Ki-67 organises heterochromatin.* Elife 5,e13722

Stålhammar G, Fuentes Martinez N, Lippert M, Tobin NP, Mølholm I, Kis L, Rosin G, Rantalainen M, Pedersen L, Bergh J et al. (2016): *Digital image analysis outperforms manual biomarker assessment in breast cancer.* Mod Pathol 29(4),318-29

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

Tumorzentrum München (Eds.): *Mammakarzinome: Empfehlungen zur Diagnostik, Therapie und Nachsorge (Manuale Tumorzentrum München)*. 16th Edition; Zuckschwerdt, München 2017

Weinberg RA: *The Biology of Cancer*. 2nd Edition; Norton & Company publishers; New York 2014

WHO Classification of Tumours (Eds.): *Breast Tumours.* 5th Edition; World Health Organisation, Geneva 2019

Wittekind C (Ed.): *TNM Klassifikation maligner Tumoren: Korrigierter Nachdruck 2020.* 8th Edition; Wiley-VCH Publishers, Weinheim 2020

Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A et al. (2007): *American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer.* Arch Pathol Lab Med 131(1),18-43

Wolff AC, Hammond ME, Hicks DG, Dowsett M, McShane LM, Allison KH, Allred DC, Bartlett JM, Bilous M, Fitzgibbons P et al. (2013): *Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update.* J Clin Oncol 1(31),3997-4013

Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, Bilous M, Ellis IO, Fitzgibbons P, Hanna W et al. (2018): *Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update.* Arch Pathol Lab Med 142(11),1364-1382

Danksagung

An dieser Stelle möchte ich allen Dank aussprechen, ohne die meine Dissertation nicht möglich gewesen wäre:

Herrn PD Dr. Peter Middel danke ich für die Möglichkeit diese Arbeit unter seiner Leitung in der Pathologie Nordhessen durchzuführen.

Herrn PD Dr. Andreas Scheel danke ich für die freundliche Betreuung und seine ständige Diskussionsbereitschaft.

Für die Bereitstellung der histopathologischen Objektträger und den kollegialen Umgang vor Ort danke ich dem Team der Pathologie Nordhessen in Kassel.

Firma Sysmex (Norderstedt) danke ich für die Leihgabe des verwendeten Objektträgerscanners und Firma 3DHistech (Budapest, Ungarn) für die Bereitstellung der *QuantCenter* Software.