

VILNIUS UNIVERSITY

Justinas
BESUSPARIS

The Assessment of the Breast Cancer Proliferation Rate and Its Intratumor Heterogeneity Using Digital Immunohistochemistry Methods

DOCTORAL DISSERTATION

Biomedical Sciences,
Medicine 06B

VILNIUS 2018

This dissertation was written during 2014–2018 at the National Center of Pathology, affiliate of Vilnius University Hospital Santaros Clinics. The research was supported by the Research Council of Lithuania (the doctoral studies were financed from the EU structural funds); in 2017, a scholarship was granted for academic accomplishments.

Academic supervisor:

Prof. dr. Arvydas Laurinavičius (Vilnius university, biomedical sciences, medicine – 06 B).

Defense board:

Chairperson – **Prof. dr. Jolanta Gulbinovič** (Vilnius university, biomedical sciences, medicine – 06 B).

Members:

Prof. dr. Tomas Poškus (Vilnius university, biomedical sciences, medicine – 06 B).

Prof. habil. dr. Dalia Pangonytė (Lithuanian university of health sciences, biomedical sciences, medicine – 06 B).

Dr. Juozas Gordevičius (Vilnius university, physical sciences, informatics – 09 P).

Dr. Darren Treanor (University of Leeds, United Kingdom, biomedical sciences, medicine – 06 B).

The dissertation will be defended at the open session research council at 9 a.m. on 7th of December 2018 in the Conference hall of Vilnius University Hospital Santaros Klinikos. Address: Santariškių str. 2, Vilnius, Lithuania.

The dissertation can be viewed in the Vilnius University Library and on the website address: <https://www.vu.lt/naujienos/ivykiu-kalendorius>

VILNIAUS UNIVERSITETAS

Justinas
BESUSPARIS

Krūties vėžio proliferacinio aktyvumo ir
jo heterogeniškumo nustatymas
skaitmeninės imunohistochemijos
metodais

DAKTARO DISERTACIJA

Biomedicinos mokslai,
Medicina 06B

VILNIUS 2018

Disertacija rengta 2014–2018 metais Valstybiniame Patologijos Centre, Vilniaus universiteto, Santaros klinikų filiale.

Mokslinius tyrimus rėmė Lietuvos mokslo taryba (doktorantūra buvo finansuojama ES struktūrinių fondų lėšomis), 2017 metais buvo gauta stipendija už akademinus pasiekimus.

Mokslinis vadovas:

Prof. dr. Arvydas Laurinavičius (Vilniaus universitetas, biomedicinos mokslai, medicina – 06B).

Gynimo taryba:

Pirmininkė – **prof. dr. Jolanta Gulbinovič** (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B).

Nariai:

Prof. dr. Tomas Poškus (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B),

Prof. habil. dr. Dalia Pangonytė (Lietuvos sveikatos mokslų universitetas, biomedicinos mokslai, medicina – 06 B),

Dr. Juozas Gordevičius (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

Dr. Darren Treanor (Lidso universitetas, Jungtinė Karalystė, biomedicinos mokslai, medicina – 06 B).

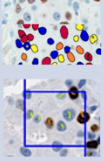
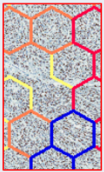
Disertacija ginama viešame disertacijos gynimo tarybos posėdyje 2018 m. gruodžio 7 d. 9 val. Vilniaus universiteto ligoninės Santaros klinikų konferencijų salėje. Adresas: Santariškių g. 2, Vilnius, Lietuva.

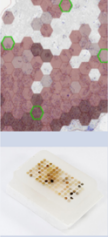
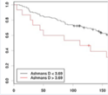
Disertaciją galima perskaityti Vilniaus universiteto bibliotekoje ir VU tinklalapyje adresu: <https://www.vu.lt/naujienos/ivykiu-kalendorius>

THESIS AT A GLANCE

THE AIM OF THE STUDY

To develop and test methodologies for a comprehensive Ki67 labeling index with an intra-tumor heterogeneity assessment based on digital immunohistochemistry methods.

Study	Objectives	Methods	Results	Conclusions
 <p>I</p>	To develop a methodology for ensuring and improving the accuracy of the approach for digital image analysis in breast cancer Ki67 IHC.	A comparison of the visual estimation of Ki67 LI, a digital image analysis result and reference values obtained with a stereology in TMA of breast cancer tissue.	We achieved a DIA misclassification rate of 5–7%, as opposed to that of 11–18% for the VE-median-based prediction.	Ki67 LI obtained by digital image analysis outperforms visual estimates, taking manual stereological counts as a reference value. An accurate Ki67 LI estimation can be achieved by DIA which is based on proper validation, calibration, and measurement error correction procedures guided by quantified bias from reference values.
 <p>II</p>	To develop a methodology for comprehensive Ki67 LI quantification with heterogeneity assessment and hot spot detection, perform analytical validation in a breast cancer patient cohort.	The WSI DIA-generated data were subsampled by hexagonal tiling, and spatial distribution parameters were calculated	The degree of proliferation measured by various automated Ki67 LI indicators was associated with higher histological grade/ more aggressive types of breast cancer. The manual hotspots of Ki67 LI were associated and comparable with the Pareto hotspot median Ki67 LI.	A systematic subsampling of DIA-generated data into hexagonal tiles would enable a comprehensive Ki67LI analysis that would reflect the aspects of intra-tumor heterogeneity and could serve as a methodology for improving digital immunohistochemistry.

Study	Objectives	Methods	Results	Conclusions
 <p data-bbox="270 753 297 776">III</p>	<p data-bbox="357 338 473 639">To optimize breast cancer tissue sampling requirements to represent Ki67 LI taking its intra-tumor heterogeneity into account.</p>	<p data-bbox="505 338 663 757">The hexagons in the HexT were chosen to simulate virtual TMA cores/ fields of view in conventional microscopy, with Ki67LI established by DIA. The sampling simulations were carried out for different heterogeneity levels.</p>	<p data-bbox="696 338 885 706">To achieve low error rates, 8 TMA cores or 4 000 nuclei are required when the heterogeneity levels are unknown. Respectively, 5–6 cores/3 000 nuclei or 11–12 cores/7 000 nuclei are required in the subgroups of homogeneous and heterogeneous tumors.</p>	<p data-bbox="917 338 1143 601">Hexagonal tiling data provide a useful model for establishing tissue sampling requirements for biomarker studies and visual estimations, which depend on intra-tissue heterogeneity and must be determined on a peruse basis.</p>
 <p data-bbox="270 1186 297 1209">IV</p>	<p data-bbox="357 820 473 1134">To evaluate the prognostic value of the comprehensive Ki67 LI estimation method in a breast cancer patient cohort.</p>	<p data-bbox="505 820 663 1106">The Ki67 LI indicators, extracted from WSI DIA-generated data, were subsampled using hexagonal tiles and compared with the data of overall patient survival.</p>	<p data-bbox="696 820 885 1188">All visual and DIA-generated indicators of the level of Ki67 expression provided significant cutoff values as single predictors of OS. Only the bimodality indicators (Ashman's D) were independent predictors of OS in the context of hormone receptor and HER2 status.</p>	<p data-bbox="917 820 1143 1239">The spatial heterogeneity indicators (the bimodality status in particular) of proliferative tumor activity, measured by the DIA of Ki67 IHC expression and analyzed using the HexT approach, can serve as an independent, prognostic indicator of OS in breast cancer patients and outperform the prognostic power of the level of proliferative activity.</p>

LIST OF ABBREVIATIONS

WSI	Whole slide image
HS	Hotspot
LI	Labeling index
Ki67 LI	Ki67 Labeling index
DIA	Digital image analysis
TMA	Tissue micro array
TMA _s	Tissue micro arrays
VE	Visual evaluation
RV	Reference value
IHC	Immunohistochemistry
TN	Triple negative
HER2	Human epidermal growth factor receptor 2
FISH	Fluorescence in situ hybridization
ER	Estrogen receptors
PgR	Progesterone receptors
GEP	Gene expression profiling
HexT	Hexagonal tiling
ROI	Region of interest
OS	Overall survival
BC	Breast cancer
ASCO/CAP	American Society of Clinical Oncology/College of American Pathologists
LIS	Laboratory information system
DCIS	Ductal carcinoma in situ
2D	Two-dimensional space
CE	Coefficient of error
HexN	Numbers of hexagons
HR	Hormone receptors

TABLE OF CONTENTS

THESIS AT A GLANCE.....	1
LIST OF ABBREVIATIONS.....	3
INTRODUCTION	6
Clinical Relevance of the Study	6
Aim of the Study	9
Study Objectives.....	9
Statements to be Defended.....	9
Scientific Novelty of the Study	10
1. REVIEW OF LITERATURE	11
1.1 Background.....	11
1.1.1 Molecular Subtypes of Breast Cancer.....	11
1.2 Issues of Ki67 Labeling Index Scoring and Interpretation.....	14
1.2.1 Guidelines	14
1.2.2 The Assessment of the Ki67 Labeling Index	15
1.2.3 Heterogeneity	17
1.2.4 Hotspots of Biomarker Expression	20
2. MATERIALS AND METHODS.....	23
2.1 Patient Cohorts	23
2.2 Tissue Preparation and Image Acquisition.....	26
2.3 Digital Image Analysis.....	27
2.3 Design of the Study and Statistical Methods.....	28
2.3.1 Study I	28
2.3.2 Study II.....	30
2.3.3 Study III.....	32
2.3.4 Study IV.....	34
3. RESULTS	36
3.1 Study I	36
3.2 Study II	41
3.3 Study III.....	44
3.4 Study IV.....	46
DISCUSSION	52
CONCLUSIONS	60
PRACTICAL RECOMMENDATIONS.....	61
FUTURE PERSPECTIVES.....	62
LIST OF PUBLICATIONS	63

4. SANTRAUKA LIETUVIŲ KALBA.....	65
4.1 Įvadas.....	65
4.1.1 Darbo tikslas.....	67
4.1.2 Darbo uždaviniai.....	67
4.1.3 Ginamieji teiginiai.....	68
4.1.4 Tyrimo naujumas.....	68
4.2 Metodai.....	69
4.3 Rezultatai.....	73
4.4 Rezultatų aptarimas.....	75
4.4 Darbo tęstinumas.....	81
4.5 Išvados.....	82
4.6 Publikacijų sąrašas.....	84
REFERENCES.....	86
ACKNOWLEDGEMENTS.....	99
PAPERS.....	100
Paper I.....	100
Paper II.....	114
Paper III.....	116
Paper IV.....	127

INTRODUCTION

Clinical Relevance of the Study

Breast cancer is the most common cancer in women both in the developed and developing countries. It is estimated that worldwide, over 464 000 women died in 2013 due to breast cancer, and 1.8 million incident cases were reported [1]. The incidence rates vary greatly worldwide from 19.3 for 100 000 women in Eastern Africa to 89.7 for 100 000 women in Western Europe. In most of the developing regions, the incidence rates are below 40 for 100 000 [2]. Breast cancer survival rates vary considerably, ranging from 80% or over in North America, Sweden and Japan, to around 60% in middle-income countries and below 40% in low-income countries [2]. The low survival rates in the less-developed countries can be explained mainly by the lack of early detection programs, resulting in a high proportion of women presenting with late-stage diseases, and by the lack of adequate diagnosis and treatment facilities.

Improvements in survival can therefore be expected from early detection and adjuvant chemotherapy treatment. Mammographic screening strategies result in the early diagnosis of breast cancer and a 25–30% decrease in breast cancer mortality in women over the age of 50 years, [3] but these strategies preferentially detect slowly growing and more well-differentiated tumors that inherently have a better prognosis and miss the fast-growing, aggressive tumors, which often present themselves as so-called interval cancers [4, 5]. Adjuvant chemotherapy and hormonal treatment have been shown to improve survival in patients with breast cancer but have potentially serious side effects and are expensive [6]. Therefore, when treating patients with hormone-sensitive primary breast cancer, the question of whether the patient requires chemotherapy additionally to endocrine treatment is crucial. Adjuvant treatment should only be given to the patients that require good prognostic factors to indicate high risk and additional factors to predict response to treatment [6].

Traditional prognostic factors, such as lymph node metastasis status, and tumor characteristics, like size or histological type, are not sufficient. Additional predictive and prognostic factors are needed to clarify the indication for adjuvant treatment, and a great number of them have been identified for breast cancer. Many of these factors are directly (cell cycle regulators [7]) or indirectly (growth factors [8-10] or angiogenesis [11]) related to cell proliferation.

Gene expression profiling (GEP) techniques open new opportunities in breast cancer prognostication and patient stratification. Molecular multigene assays are based on the identification of a set of genes (gene signature) that can be used to identify tumors with specific biological or clinical features. This allows to stratify relevant patient groups into low- and high-risk prognostic subgroups, calculate a recurrence score to estimate the risk of distant recurrence and guide further treatment. The only test that was acknowledged by the majority (80.5%) of the 14th st. Gallen (2015) panelists as providing reliable prognostic information on the benefits of additional adjuvant chemotherapy was the Oncotype DX [12]. This approach is a reverse-transcriptase-polymerase-chain-reaction assay of 21 prospectively selected genes, including the markers of proliferation, invasion, her2, estrogen and a group of reference genes [13]. More importantly, the largest proportion of genes used in this test are related to the cell proliferation cycle (Ki67, Survivin, MYBL2, CCNBL, STK15). Therefore, the high cost of multigene assays has motivated many researchers to search for alternative risk assessment models based on immunohistochemical biomarker expression [14-17].

Treatment decisions for breast cancer are significantly influenced by tumor subtype, which is defined by the expression of the estrogen receptor (ER), progesterone receptor (PGR), human epidermal growth factor receptor 2 (HER2) and the proliferation marker Ki67 [18]. The Ki67 immunohistochemistry (IHC) is a widely used method to estimate cell proliferation rate in tumor cells. A visual assessment of Ki67 IHC, commonly used in current clinical practice, has serious limitations due to a low reproducibility among the pathologists of intermediate Ki67 labeling index (Ki67LI) evaluation, where it is crucial for making clinical decisions [19-23]. A misinterpretation of the ki67 labeling index may result in a lost opportunity for patients to receive chemotherapy or may result in patients being over-treated. Therefore, standardization efforts have led to recommendations focused on a high number of cells, in a range from 500 to 2 000, needed to reach accurate visual estimations [24]. Furthermore, intra-tumoral heterogeneity is an inherent feature for this biomarker in breast cancer. Thus, counting Ki67LI in areas containing the largest proportion of positive tumor nuclei, named “hotspots” (HS), is considered to be essential in clinical practice. However, proper hotspot definitions have never been introduced and standardized by international recommendations – the optimal Ki67LI cutoff values for making clinical decisions also vary greatly (14%,

20%, 20–29%, $\leq 10\%$, 20–30%) year by year [12, 24-26]. Furthermore, the American Society of Clinical Oncology (ASCO) have published clinical practice guidelines (2016) on breast cancer [18] wherein it is recommended that the Ki67 labelling index, determined by immunohistochemistry, should not be used to guide choice on adjuvant chemotherapy with an intermediate quality of evidence base and moderate strength of recommendation. Since manual estimations are highly complicated by these factors that reduce the prognostic value of Ki67LI, novel, more comprehensive and reproducible Ki67LI estimation methodologies need to be considered, explored and applied in clinical practice.

The most unique and significant benefit for pathology practice and research can be expected from digital image analysis (DIA) applications. This opens new perspectives for pathology to serve the needs of personalized medicine by providing more accurate and reproducible measurements for tissue-based diagnosis, prognosis and prediction [27, 28]. Microscopic images, used in pathology, contain rich, multi-parametric data that can be retrieved by scanning and processing the images by numerous methods available to visualize tissues and cells as well as scan and process the images while generating rich, multi-parametric data [28].

Digitalized Ki67 immunohistochemistry has been explored broadly by applying various image analysis methods in the past several years. Automated Ki67LI estimation in breast cancer has been shown to be reproducible [22] and consistent with visual evaluations made by pathologists [29-32], suggesting the DIA's applicability for Ki67LI assessment in clinical practice. However, it has been also shown that the concordance between visual estimation and DIA is effected by tumor heterogeneity [31], which, unfortunately, is complicated to measure and has no standardization in breast cancer pathology. While some studies were focused on optimal tumor cell segmentation and automated Ki67LI estimations, others have proposed automated methods to detect the hotspots of Ki67LI [33-37]. However, the analytical and, most importantly, clinical validation of DIA were out of scope in these studies, which hinders the implementation of DIA approaches in clinical practice.

Aim of the Study

To develop and test the methodologies of a comprehensive Ki67 labeling index with intra-tumor heterogeneity assessment based on digital immunohistochemistry methods.

Study Objectives

1. To develop a methodology for ensuring and improving the accuracy of the digital image analysis approach in breast cancer Ki67 immunohistochemistry.
2. To develop a methodology for comprehensive Ki67 LI quantification with heterogeneity assessment and hotspot detection; to perform analytical validation in a breast cancer patient cohort.
3. To optimize breast cancer tissue sampling requirements to represent Ki67 LI, taking its intra-tumor heterogeneity into account.
4. To evaluate the prognostic value of the comprehensive Ki67 LI estimation method in a breast cancer patient cohort.

Statements to be Defended

1. Ki67 LI, obtained by DIA, enables a higher measurement accuracy than that of the visual estimates by pathologists and provides criterion standard with subsequent image analysis calibration and measurement error correction.
2. The intra-tumor heterogeneity of proliferative tumor activity, estimated by systematic subsampling of DIA data by hexagonal tiling, can serve as an independent prognostic indicator of OS in breast cancer patients that outperforms the prognostic power of the level of proliferative activity.

Scientific Novelty of the Study

This study covers several novel aspects:

1. A methodology for improving the accuracy of digital image analysis, based on analytical assay validation principles, was proposed for IHC. The validation and calibration procedures were based on a direct criterion standard obtained by stereology grid-count reference values.
2. A novel methodology for comprehensive tissue biomarker estimation, based on a hexagonal tiling of DIA data, was proposed, developed and tested for the use case of Ki67LI. This method enriched the biomarker measurement with spatial aspects of intra-tumor expression with analytically relevant indicators. Similar methods of image segmentation into subregions have been applied in pathology research as well as in the areas of geography and ecology. However, hexagonal grid-based spatial analytics were first developed and applied for digital microscopy image analysis tasks. More importantly, this approach was further elaborated to propose the concept of the “Pareto hotspot” representing a median biomarker expression level in the 20% of the “hottest” tumor area.
3. The heterogeneity of breast cancer proliferative activity is a well-recognized phenomenon that hampers the potential prognostic value of Ki67LI. Our approach allowed to compute the intratumor heterogeneity indicators which, for the first time, were found to outperform the “conventional” Ki67LI proliferation level indicators in multivariate prognostic modeling of the patient OS. These findings potentially strengthen the evidence base of clinical recommendations for applying Ki67LI obtained by immunohistochemistry for guiding adjuvant chemotherapy, which was eliminated by ASCO guidelines [18].
4. The hexagonal tiling simulation was utilized to establish tissue sampling requirements for the IHC biomarker measurement with regard to the individual biomarker intratumor heterogeneity characteristics. Previous studies exploited physical tissue sampling, which was the major limitation in performing repetitive sampling of the tissue and to achieve comprehensive statistical modeling.

1. REVIEW OF LITERATURE

1.1 Background

1.1.1 Molecular Subtypes of Breast Cancer

Breast cancer is a heterogeneous disease featuring distinct histological, molecular phenotypes and clinical characteristics. Historically, it has been classified based on clinic-pathological features, such as tumor spread, size and grade. However, these classifications are insufficiently accurate; better predictors of high risk and treatment response are needed to reflect the diverse biological and clinical heterogeneity of breast cancer [6, 38]. In recent times, global gene expression profiling (GEP) studies using unsupervised clustering techniques have provided molecular classification system and identified distinct clusters or intrinsic subtypes based on the quantitative expression of several genes (transcriptome profiles) [39, 40]. Perou et al. have characterized variations in gene expression patterns in a set of 65 surgical specimens of human breast tumors from 42 different individuals, using complementary DNA microarrays representing 8 102 human genes. These patterns provided a distinctive molecular portrait of each tumor and proved that breast cancer at the transcriptome level is not a single disease [39]. Despite the fact that each individual tumor features a unique GEP related to its specific biological features and genetic abnormalities, tumors clustered together to produce distinct reproducible classes based on transcriptomic profiles with common overlapping features [38]. In a particular study [39], two main clusters were identified and appeared to be related to ER expression. The ER positive cluster was enriched with ER, ER-related genes and other genes characteristic of the luminal epithelial cells, and this class was termed as “luminal” to indicate its molecular similarity to them. The other major class contained ER negative tumors and showed three distinct subclasses termed “HER2 positive,” “basal-like” and “normal breast-like.” The HER2 subgroup was characterized by an overexpression of HER2 and other genes pertaining to the HER2 amplicon. The basal-like class was largely characterized by the lack of expression of ER and HER2 and by positive expression of genes characteristic of basal-like cells of the breast and by high proliferative activity. The normal breast-like class displayed a triple-negative phenotype but did not cluster with the basal-like centroid and was characterized by expression profiles similar to those found in normal breast tissue.

Subsequent studies have also reported an association between these subtypes and patient outcome and that these classes are associated with distinct biological pathways, making them potential candidates for targeted therapy [41].

However, a gene expression array analysis is not always feasible to obtain in routine practice. Thus, a simplified classification of subtypes defined by clinic-pathological criteria has been proposed by Cheang et al. [42]. In 2011, during the 12th St. Gallen International Breast Cancer Conference, this approach was adopted by expert panel to the classification of patients for therapeutic purposes based on the recognition of intrinsic biological subtypes within the breast cancer spectrum [25]. For practical purposes, these subtypes were approximated using clinic-pathological rather than gene expression array criteria [25]. This approach uses the estrogen and progesterone receptor status obtained by immunohistochemistry, the detection of overexpression and/or amplification of the human epidermal growth factor receptor 2 (HER2) oncogene, and Ki67 labeling index, as the means of identifying tumor subtypes (Table 1). In general, systemic therapy recommendations follow the subtype classification. Thus, the “Luminal A” disease generally requires only endocrine therapy, which also forms part of the treatment of the “Luminal B” subtype. Chemotherapy is considered indicated for most patients with “Luminal B,” “Human Epidermal growth factor Receptor 2 (HER2) positive,” and “Triple negative (ductal)” disease, with the addition of trastuzumab in the “HER2 positive” disease [25]. Hence, this classification requires the availability of reliable and comprehensive measurements of ER, PgR, HER2 and Ki67 LI immunohistochemistry.

Table 1. Surrogate definitions of intrinsic subtypes of breast cancer [25, 42-45].

Intrinsic subtype	Clinicopathologic definition
Luminal A	ER and/or PgR positive, HER2 negative, Ki67 low (range varies *)
Luminal B (HER2 negative)	ER and/or PgR positive, HER2 negative, Ki67 high
Luminal B (HER2 positive)	ER and/or PgR positive, HER2 over-expressed or amplified, any Ki67
HER2 positive (non-luminal)	HER2 over-expressed or amplified, ER and PgR absent
‘Basal-like’	‘Triple negative (ductal)’, ER and PgR absent, HER2 negative

*Ki67 cut-off value varies among different recommendations.

Optimal clinical decision-making and appropriate patient identification for adjuvant breast cancer therapies are based on both prognostic and predictive tumor markers [46]. Tumor proliferation is an essential element of cancer progression, and mitosis counting has the most reproducible and independent prognostic value [6]. However, the mitotic index, which is the most established form of proliferation assessment, has limitations, as the duration of the mitotic phase can vary, especially in aneuploid tumors. Thus, the number of mitoses is not linearly correlated with the rate of proliferation [6]. Therefore, cell-cycle-associated biomarkers, such as cyclin D1, cyclin E, Ki67, p21, have been considered as prognostic factors [47]. Ki67 LI correlates with mitotic index [48, 49] and has emerged as the marker of choice with both prognostic and treatment predictive value in breast cancer [50-52].

Ki67 is a nuclear non-histone protein first identified by Gerdes et al. in the early 1980s at the University of Kiel, Germany [53]. Ki67 was found to be universally expressed among proliferating cells in many sites and absent in resting cells, making it a potential marker for evaluating the growth fraction of normal and neoplastic human cell populations [53-55]. An antibody with applicability in paraffin-embedded tissue was eventually developed and named MIB-1 for the Ki67 gene MKI67 [56], Figure 1.

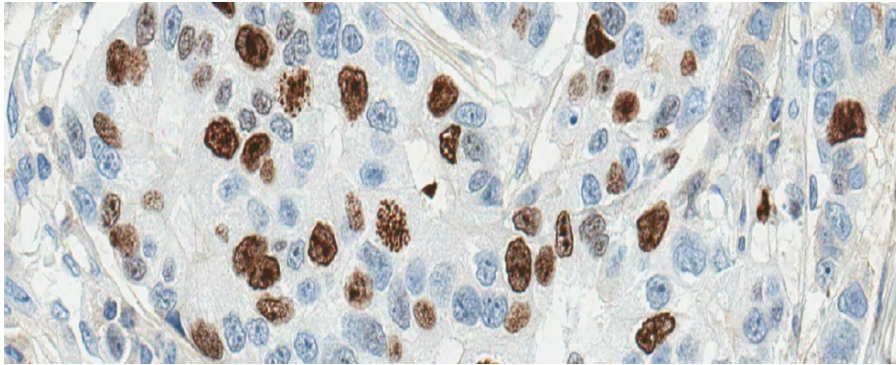


Figure 1. A sample of invasive ductal breast carcinoma tissue stained by Ki67 (MIB-1) immunohistochemistry.

The blue color indicates a counterstain (hematoxylin) that reflects Ki67 negative nuclei. The brown chromogen stains reflect positive nuclei for the Ki67 antibody.

1.2 Issues of Ki67 Labeling Index Scoring and Interpretation

1.2.1 Guidelines

Ki67 is assessed in pathology routine practice using immunohistochemistry. In 2011, the International Breast Cancer Working Group [24] published recommendations for the assessment of Ki67 LI in breast cancer. However, still no global guideline, with both reproducibility and objective standardization, has been established for Ki67 LI assessment in breast tumors. The different cut-off values have been proposed for Ki67 LI, which thereby severely limit its clinical utility. The 12th St. Gallen International Breast Cancer Conference (2011) suggested a 14% cut-off to distinguish Luminal A and Luminal B tumors [25]. In 2013, this value was questioned by the St. Gallen panel, and an increase in the cut-off value to 20% was discussed [26]. In 2015, one group of the 14th St. Gallen panelists proposed a Ki67 level of at least 20–29% as cut-off for luminal-B cancer. Another group pointed out that the Ki67 level in luminal A subtype breast cancers is likely to be $\leq 10\%$ [12]. A recent meta-analysis of 85 studies (in 32 825 patients) showed that the staining levels of 10%–20% have been the most common to dichotomize the patient populations [57]. Nevertheless, without a standardization of methodology, these cut-offs have limited value outside

of the studies from which they were derived and the centers that performed them [31].

1.2.2 The Assessment of the Ki67 Labeling Index

A visual assessment is now a method of choice for Ki67 LI evaluation in the majority of pathology laboratories and institutions. According to the International Breast Cancer Working Group recommendations for the assessment of Ki67 LI in breast cancer [24], scoring should be based on counting of at least 500 malignant invasive cells and preferably at least 1000 cells. These numbers for manual counts are relatively high to achieve in daily practice and are laborious. Therefore, many pathologists trust quick estimation “at a glance” or have their own techniques to quickly count the proportion of positive and negative nuclei in the tissue. In addition, there is no consensus about whether the Ki67 LI should be calculated as the percentage of Ki67-positive tumor cells in the whole tissue section or in the areas containing the largest proportion of positive tumor nuclei, commonly named “hot spots” [24, 58]. Additionally, various levels of spatial heterogeneity are the common feature of breast carcinomas [59-62], which burdens the evaluation of the biomarker expression in IHC slides.

A conventional interpretation of immunohistochemistry is based on human visual ability to identify tissue structures and to produce semi-quantitative estimates. This process might lack reproducibility and affect the Ki67 LI [24]. While this approach is sufficient for simple routine diagnostic tasks, it does not meet the expectations of personalized breast cancer therapies. Visual assessment has limitations, such as high inter- and intra-user variability and relatively poor reproducibility of intermediate Ki67 LI values where it is crucial for making clinical decisions [19-23]. A misinterpretation of the Ki67 labeling index may lead to an inappropriate management of the patient and lost opportunities to receive relevant treatment. Therefore, the innovative and comprehensive measurement methods of immunohistochemistry need to be established and explored in the context of clinical outcome data.

The digitalization of histology glass slides opened new perspectives in digital image analysis, which allowed to overcome the limitations of semi-quantitative estimations. The applications of digital image analysis are very broad in a field of pathology and cover automated mitotic counts [63-65], the estimations of immunohistochemically labeled cells in various tumors [66-72], tissue area quantification [73-76], lymphoid cell counts [77-82],

automated metastasis detection [83, 84] and are even utilized in scanned fluorescence in situ hybridization (FISH) images [85]. In breast cancer, digital image analysis has been successfully applied for an automated quantification of HER2 [68] and hormone receptors [69]. However, the largest amount of studies performed in the several past years have been focused on an automated Ki67 LI estimation. In a recent study, Zhong et al. [31] compared visual Ki67 LI estimates assessed by five breast pathologists and automated digital image analysis. All cases were classified into three groups by VE values ($\leq 10\%$, 11%-30% and $>30\%$ Ki67 LI) and Ki67 LI was evaluated in WSI and hotspot areas. The authors reported a perfect agreement between VE and DIA of Ki67 LI in the whole cohort of G2-G3, ER positive/HER2 negative cases. Average score and hotspot score methods both demonstrated perfect concordance between VE and DIA of Ki67 LI. The concordance was relatively lower in intermediate Ki67 LI group (11%–30%) compared with high ($>30\%$) Ki67 LI groups according to both methods. Gudlaugsson et al. [22] compared the reproducibility and prognostic value of different Ki67 LI measurement methods in 237 T1,2 N0 M0 breast cancer patient cohort without adjuvant systemic treatment. Ki67 LI assessment methods for the comparison included: a subjective “quick-scan rapid estimate” by two pathologists, ocular-square-guided counts in the hotspot with the subjectively highest Ki67 expression, computerized point-grid-sampling interactive morphometry (CIM) and automated digital image analysis (DIA). The authors concluded that Ki67 LI by DIA, but not subjective counts, was reproducible and prognostically strong.

The automated digital image analysis is a highly sensitive method that directly depends on calibration procedures. The training sets must include high numbers of varying morphology cases and the whole procedure must be performed in several steps, containing testing sets and error corrections. Consequently, the algorithms must be validated based on analytical methods and, most perfectly, clinical validation should follow the whole workflow. However, the validation is either missed in previous studies or the reference values are chosen to be semi-quantitative, such as human visual estimations, which are biased. This validation strategy is a paradox, because digital image analysis is usually used to score complex biomarkers, as it is more reproducible and objective than manual evaluation, which cannot be used as a “golden standard” to analytically validate DIA tools [86].

1.2.3 Heterogeneity

Intra-tumor heterogeneity describes the observation that different tumor cells can show distinct morphological and phenotypic profiles, such as cellular morphology, gene expression (including the expression of cell surface markers, growth factor and hormonal receptors), metabolism, motility and angiogenic, proliferative, immunogenic and metastatic potential [87]. With rare exceptions, spontaneous tumors originate from a single cell. However, at the time of clinical diagnosis, the majority of human tumors display various levels of heterogeneity. To a substantial extent, this heterogeneity might be attributed to morphological and epigenetic plasticity, but there is also strong evidence for the co-existence of genetically divergent tumor cell clones within tumors [87, 88].

The intra-tumor heterogeneity has long been recognized as a feature of some breast carcinomas in the context of IHC biomarker expression (Figure 2), microscopic and/or molecular characteristics [59-62, 69, 89-95] and represents a major challenge for the design of effective therapies. The lack of information about variability within the tumor, or between tumors with the same score, blinds clinicians to a potential readout that could represent the “biology,” eventually responsible for non-effective responses to therapy. It is intuitive that different cell populations within or between tumors could contribute to clinical refraction to therapy and thereby affect patient outcomes [96]. The concept of tumor heterogeneity leading to drug resistance was debated as early as the 1950s as the “Greenstein Hypothesis” and has become part of cancer biology doctrine [97]. In more recent times, as more targeted therapies are being developed, the issue of tumor heterogeneity has re-emerged as a factor significant to clinical strategy. Thus, there is an increasing need for clinical and pathological evaluations of tumor heterogeneity that would be aligned with an improved understanding of cancer biology [96].

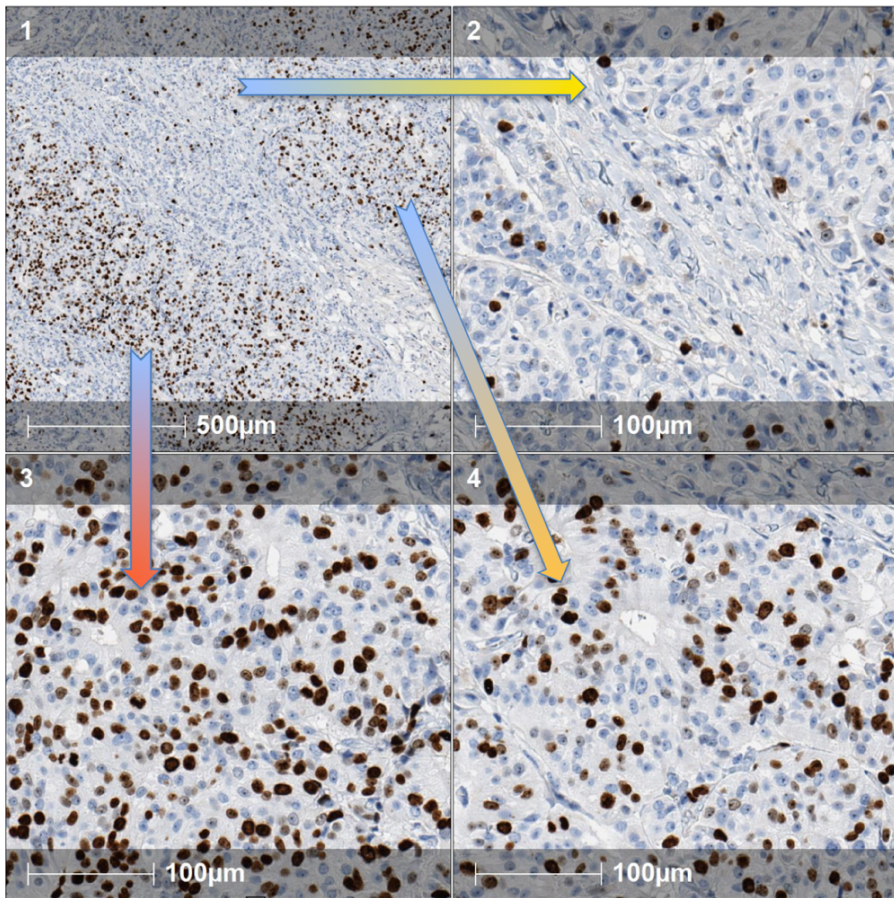


Figure 2. An example of ductal breast carcinoma with a heterogenous expression of Ki67 immunohistochemistry.

1. Tumor area with a heterogenous Ki67 LI. 2. Higher magnification of low proliferative area. 3,4 Higher magnification of the same tumor areas containing larger proportion of Ki67 IHC positive tumor cells.

There are several areas in breast cancer pathology where heterogeneity plays an important role and an accurate measure of biomarker expression is needed. It is suggested that HER2 immunohistochemistry can have a marked intra-tumor heterogeneity [98] and approximately 30% (ranging from 11% to 40%) of breast tumors exhibit heterogeneous HER2 amplification [91, 99-103]. Genomic heterogeneity refers to the coexistence of more than one population of tumor cells with distinct HER2 amplification

characteristics within the same tumor [104] and it affects the management of breast cancer patients [105]. It is reported that breast cancer patients with the HER2 heterogeneity had reduced disease-free survival [103] and influenced the effectiveness of specific therapy in metastatic HER2 positive breast cancer [106]. While great efforts were made to define the genetic HER2 heterogeneity [101, 104, 107-110], the definitions for heterogenous HER2 expression in immunohistochemistry are still lacking. Hormone receptor (ER, PR) IHC staining variations within the same tumor have been also recognized in a number of previous studies [98, 111-113]. The current guidelines of the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) recommend to use a 1% cut-off of tumor cell positivity to classify a tumor as ER expressing and select patients for endocrine therapy [44]. However, this approach ignores and does not consider intra-tumor heterogeneity, which limits the classification of ER positive/ ER negative tumors when distinct populations of tumor cells are unequally distributed within invasive tumor area [114, 115]. Moreover, about a half of patients with ER-positive disease fail to respond to endocrine therapy, and approximately 25% of women with early-stage breast cancer will develop distant metastases [69, 116, 117]. Recently, Lindström et al. investigated a large series (n = 1780) of breast cancer patients and showed that intra-tumor heterogeneity of ER is an independent long-term prognosticator with potential to change clinical management, especially for patients with luminal A tumors [69].

Spatial tumor heterogeneity also impacts traditional immunohistochemical Ki67 LI analysis and causes discordant reads between pathologists. Shui R. et al. indicated that the biological heterogeneity of Ki67 staining is an important reason of the low inter-observer reproducibility, especially of intermediate Ki67 labeling index values in which most cut-offs are located for making clinical decisions [118]. While the genomic basis of tumor heterogeneity could be easily determined by applying expensive molecular profiling techniques, such as next generation sequencing [119], defining an intra-tumor heterogeneity in the 2D space of a microscopic section is a challenging task. The determination of heterogeneity in conventional immunohistochemistry may be complicated, because each tumor has a specific histological pattern, various levels of biomarker expression, and defining heterogeneity is a very subjective task by itself. Consequently, proper definitions and standardization techniques of hotspot detection and heterogeneity indicators need to be established.

1.2.4 Hotspots of Biomarker Expression

A lot of research within the use of hotspots and different automatic methods for defining and detecting HS has been carried out in the several past years. One of the most recent studies, published by Saha et al. [34] in 2017, described an application of deep structured learning approach for the HS detection in Ki67 IHC slides and found high accuracy and reproducibility rates of the developed method. However, the reference values were based on subjective visual estimations, and this study did not suggest any HS definition criteria. In contrast, Lindberg et al. [36] tried to use objective criteria for HS definitions and developed a methodology for an automated detection of Ki67 LI hotspots by visualizing them in a heatmap that was based on the percentage of Ki67 positive nuclei calculated in small circular regions. The authors further analyzed the DIA output by each pixel and defined a hotspot as the largest continuous area with the relative maximum pixel value in the heatmap. The region area for hotspots was incrementally increased or decreased to fit the minimum requirement of 400 cells. This study proposed an alternative methodology to detect and define the HS, which guides the pathologist to quickly find the HS.

The use of clusters analysis for finding potential hotspots is also quite common in recent research papers. Some studies [33, 120] have applied clustering methods for detecting hotspots in neuroendocrine tumors of gastrointestinal tract and high-grade gliomas. In another study [33], a hybrid clustering method, referred to as a Seedlink, was developed. The study has shown that a strong improvement of inter-pathologist agreement was obtained when Seedlink-aided selections were considered. Heindl A et al. [121] investigated the clinical relevance of the spatial heterogeneity of immune infiltration in ER-positive breast cancer. The authors successfully applied a spatial clustering analysis method to identify tumor regions with statistically significant groups of immune cells by looking at each feature (immune cell) within the context of the neighboring features [121].

Spatial statistics derived from dividing images into regular regions were also applied to detect the hotspots of biomarker expression. Mostly rectangular shaped grids were used to sample tissue areas. Gudlaugsson et al. [22] described a tool based on grid structure to identify the HS of Ki67 LI in breast cancer. Squares were selectively placed on the regions with subjectively high numbers of positive nuclei, and Ki67 LI was counted by

DIA inside each square. They found that the Ki67 scores of semi-automated hotspots yielded reproducible and prognostically significant results. However, this method directly depends on the manual selections of HS area, and the Ki67 scores were based on the area but not on the number of nuclei as it should be due to the possible variation in density of cells. Other studies have applied automated methods to segment images into rectangular tiles. Thakur et al. [15] investigated the proliferative activity of the hotspots of breast cancer by applying a rectangular grid with a requirement of at least 500 cells in each tile. The results showed that the Ki67 LI, obtained in a HS area, which was defined as top 5 tiles containing the highest Ki67 positive cell ratio, predicts Oncotype DX prognostic groups with high accuracy. In one study [37], an additional adaptive $\frac{1}{4}$ step shifting and multiple analyses for each image were utilized. However, the DIA validation procedures were out of scope in the latter studies. Furthermore, the rectangular shape of the tiles is not the best choice, as the hotspots are usually very complex in shape and rectangles may not cover these areas completely.

Despite the variety of conventional and advanced machine learning technologies used for tumor cell quantification, no agreed HS definitions or parameters for heterogeneity indicators have been achieved in breast cancer Ki67 immunohistochemistry. Furthermore, the detection of hotspots by either visual or digital methods can be complicated by the nature of the tumor tissue: hotspots may vary greatly in size, shape and contrast. However, breast cancer treatment decisions are based on relatively low Ki67 LI cutoff values, which raises the need for accurate Ki67 LI measurements in pathology practice. The results of this work enabled to comprehensively measure Ki67 LI in breast cancer immunohistochemistry, detect hotspots and, most importantly, to define the degree of spatial heterogeneity in the two-dimensional space of an individual tumor, which was an impossible task before now. The proposed method could potentially serve as a decision-support system in pathology practice for more robust and more accurate Ki67 LI estimation and, most importantly, for better patient stratification to achieve the best treatment effect. Findings described in this thesis are an essential contribution toward a better understanding of how to measure spatial tumor heterogeneity in immunohistochemistry slides of various biomarkers and possibly discover new and less expensive prognostic factors. The novel hexagonal tiling methodology was also utilized for defining breast cancer sampling requirements for research studies and for clinical practice in routine pathology. The experiments were based on powerful statistical

modeling, and the level of intratumoral heterogeneity was taken into account. The practical application of these findings is important when considering the efficacy of tissue microarray construction for biomarker investigation experiments or helping to optimize visual estimations for pathologists.

2. MATERIALS AND METHODS

This section gives a brief overview of the main methods used within framework of this thesis. More specific detailed information is provided in publications, respectively (Study I – IV, [28, 122-124]). The experiments were performed on different patient cohorts. Separate parts of the study were approved by the Lithuanian Bioethics Committee and Nottingham Research Ethics Committee 2 under the title “Development of a molecular genetic classification of breast cancer.” The patients’ consents to participate in the study were obtained. All statistical analyses were performed by using SAS 9.3 software, Microsoft Excel software (Microsoft, Redmond, Washington, US) and OpenOffice Calc software (Oracle, Redwood City, California, US). The statistical significance level was set at $P < 0.05$.

2.1 Patient Cohorts

For the first part of this work (Study I), 164 tissue micro array (TMA) Ki67 IHC images of 1mm diameter tissue cores from female patients with invasive ductal carcinoma of the breast were used. The patients were treated at the Oncology Institute of Vilnius University and investigated at the National Center of Pathology, during the period of 2007–2009.

The second part of this work (Study II) included whole slide images from 302 patients with invasive breast carcinoma who had been treated by surgical excision at the National Cancer Institute (Lithuania) during 2013–2014. Patient characteristics are provided in Table 2. The same patient cohort was later used for virtual TMA simulation experiment (Study III)

Table 2. Patient and tumor characteristics of the second study population

Characteristics	Number
Age (median, min.–max.)	60 years, 24–88 years
Histological type	
Invasive ductal/no special type	271 (89.7%)
Other types	31 (10.3%)
	n=302
Subtype	
HR positive	189 (62.6%)

HER positive *	55 (18.2%)
Triple negative	55 (18.2%)
	n=299**
Axillary nodal stage	
0	171 (56.6%)
1	83 (27.5%)
2	28 (9.3%)
3	9 (2.9%)
	n=291***
Histological grade	
1	21 (7%)
2	123 (41%)
3	157 (52%)
	n=301 ****

* Includes 33 HR (hormone receptors) positive cases.

** A subtype was not established in 3 cases

*** An N stage was not established in 11 cases

**** A histological grade was not established in 1 case

In the last part of this thesis (Study IV) 152 cases from the Nottingham-Tenovus Primary Breast Carcinoma Series, aged 70 years or less, presenting with primary operable (stages I, II, and IIIa) invasive breast carcinomas between 1986 and 1998 were used. This is a well-characterized consecutive series of patients who were uniformly treated according to locally agreed clinical protocols [124-126]. Detailed patient characteristics are summarized in Table 3. The mean duration of follow-up after the surgery was 143.4 ± 71.4 months (range 5 to 248 months, median 156). Seventy-nine patients died during the follow-up period. This patient cohort was used in the framework of the international project entitled “A Comprehensive Biomarker Intratumour Heterogeneity Evaluation by Digital Immunohistochemistry Image Analysis” which was funded by European Social Fund under operational programme for human resources development for 2007-2013 priority 3 “Strengthening Capacities of Researchers” measure VP1-3.1-ŠMM-07-K „Support to Research Activities of Scientists and Other Researcher (Global Grant)”. Professor Ian O Ellis provided histological slides and pathology report data (patient age, tumor histological type, grade, estrogen and progesterone receptor scoring, HER2 (verified by a HER2 FISH test in IHC 2+ cases) and Ki67) of 152 patients. Breast pathologist and

researchers from Nottingham participated in digital image analysis experiments which are described in study IV.

Table 3. Patient and tumor characteristics of the IV study population

Characteristics	Number
Age group	
Age ≤55 years	85 (56%)
Age >55 years	67 (44%)
Histological type	
Invasive ductal/no special type	104 (68%)
Other types	48 (32%)
Subtype	
HR positive	101 (68%)
HER positive *	22 (15%)
Triple negative	26 (17%)
	n=149**
Axillary nodal stage	
1	78 (51%)
2	58 (38%)
3	16 (11%)
Axillary lymph node status	
Negative	78 (51%)
Positive	74 (49%)
Histological grade	
1	9 (6%)
2	52 (34%)
3	91 (60%)
Nottingham Prognostic Index	
Good	31 (21%)
Moderate	81 (53%)
Poor	40 (26%)
Endocrine Therapy (n = 145)	81 (56%)
Chemotherapy (n = 151)	28 (19%)

* Includes 10 HR positive cases.

** A subtype was not established in 3 cases

2.2 Tissue Preparation and Image Acquisition

For experiments described in studies I-III, paraffin sections were cut at 3 μm thickness and immunohistochemistry for Ki67 was performed with a multimer-technology based detection system, ultraView Universal DAB (Ventana, Tucson, AZ, USA), Figure 3. The Ki67 antibody (clone MIB-1; DAKO, Glostrup, DK) was applied at a 1:200 dilution for 32 minutes, followed by the Ventana BenchMark XT automated immunostainer (Ventana) standard Cell Conditioner 1 (CC1, a proprietary buffer) at 95°C for 64 minutes. Finally, the sections were developed in DAB at 37°C for eight minutes, counterstained with Mayer's hematoxylin and mounted [28]. Whole slide tissue preparation and IHC staining protocol used in study IV is previously described in [127]. Digital images were captured using the Aperio Scan-Scope XT Slide Scanner (Aperio Technologies, Vista, CA, USA) under 20x objective magnification (0.5 μm resolution).

For the first study tissue, micro arrays (TMAs) were constructed by punching one millimeter diameter cores from invasive tumor areas which were randomly selected by the pathologist. Detailed technique of tissue microarray construction is described in [128]. One TMA core per patient was used for the study. Other studies were performed on full-face tissue sections.

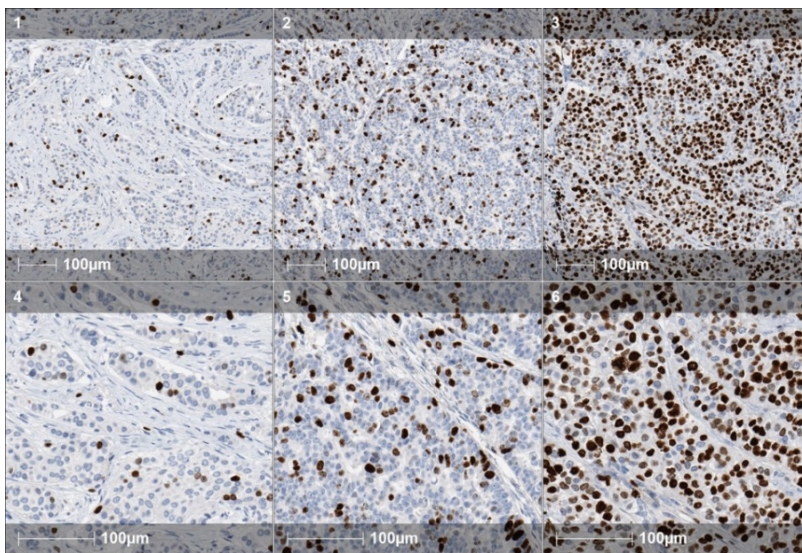


Figure 3. Examples of Ki67 immunohistochemistry in three cases of ductal breast carcinoma with a variable degree of proliferation

2.3 Digital Image Analysis

DIA was performed by using 2 image analysis software platforms: for studies I-III Aperio Genie and Nuclear v9 algorithms were applied and HALO™ Classifier Module/CytoNuclear v1.4 algorithm (Indica Labs, NM, USA) for study IV was used. The software enables automated recognition of the tumor tissue and cell segmentation in scanned images, Figure 4. The classifier was trained by the pathologist to detect tumor areas while eliminating fibrous and inflammatory stromal compartments. The nuclear analysis modes were calibrated to enumerate Ki67-positive and -negative tumor nuclear profiles in the breast cancer tissue. In the first study [28], several calibration cycles of the DIA (named DIA-0, 1 and 2, resulting in the percentage of Ki67-positive tumor cells - Ki67- DIA-0, 1 and 2, respectively) were performed to improve the accuracy of the tool by adjusting the settings of the Nuclear algorithm. Ki67-DIA-0 was obtained by the default Aperio settings for the Nuclear algorithm, Ki67-DIA-1 - by “subjective” visual assessment of the quality of the DIA results on the computer monitor; Ki67-DIA-2 was fine-tuned based on the quantitative bias established by statistical analyses comparing the Ki67-DIA-1 to reference values (RV) (Ki67-Count) [28]. In study IV, the ductal carcinoma in situ (DCIS) component was excluded by manual annotations. The quality of the automated tumor and stroma segmentation and Ki67 positivity threshold by the DIA was monitored by visual inspection [124].

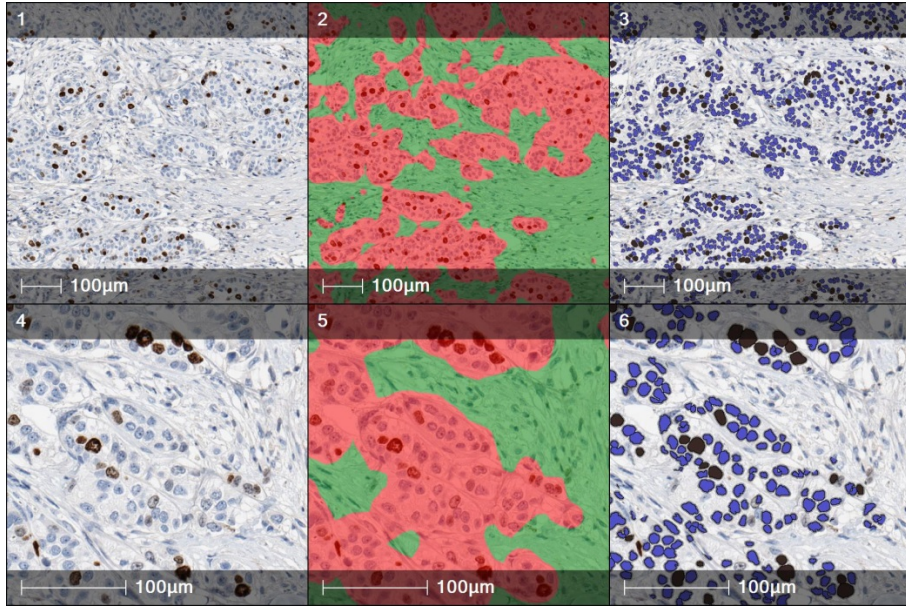


Figure 4. Examples of breast cancer Ki67 immunohistochemistry analyzed by DIA algorithms

Pictures 1, 4 show the scanned Ki67 IHC without analysis at different magnification levels; Pictures 2 and 5 illustrate the automated detection of tumor (red) and stroma (green) by previously calibrated HALO classifier algorithm; Pictures 3 and 6 depict the final DIA result by segmenting Ki67-positive (brown) and negative (blue) tumor nuclear profiles in classified tumor areas.

2.3 Design of the Study and Statistical Methods

As the main purpose of this work was complex, various techniques were used, and the whole work was divided into several parts as follows.

2.3.1 Study I

Digital Image Analysis: Calibration, Quantification and Validation

In this study, digital image analysis tool validation procedure was performed by comparing KI67 LI obtained by DIA with a reference data. The reference

value was generated by manually marking Ki67-positive and negative tumor cell profiles, using a stereological method for 2D object enumeration implemented by the Stereology module (ADCIS, Caen, France) with a test grid of systematically sampled frames overlaid on a TMA image (Figure 5). Frame size of 125 pixels and spacing of 250 pixels was chosen. The fraction of Ki67 positive tumor cell profiles was calculated as $100 \times \text{Ki67-positive nuclear profiles} / (\text{Ki67-positive nuclear profiles} + \text{Ki67-negative nuclear profiles})$. A visual assessment for the Ki67 LI on the same images as used for DIA was performed by five pathologists independently and provided semi-quantitative values (Ki67-VE-1, 2, 3, 4 and 5) expressed as the percentage of Ki67-positive tumor cell profiles. To test the degree of uncertainty of the reference value, inter-observer variation was estimated based on Ki67-Count values produced by three observers (Ki67-Count-1, 2 and 3) independently in a subset ($n = 30$) of the TMA images. Since the inter-observer variability was found to be negligible, the RV in the whole series ($n = 164$) were established by one-observer marking (Ki67-Count), splitting the job among four observers in approximately equal proportions [28].

The accuracy of the DIA and VE with regard to the RV was estimated by one-way ANOVA (Duncan multiple range test was applied for pairwise comparisons), a Pearson correlation, single and multiple linear regression analyses as well as an orthogonal linear regression based on principal component analysis. Agreement between individual measurements was also estimated based on 95% confidence intervals calculated from the RV CE and visualized by Bland and Altman plots [129, 130]. Dependence of RV ($n=30$) and VE ($n=164$) inter-observer variation on the magnitude of measurement was visualized by plots of corresponding standard deviations against the mean values of the measurements.

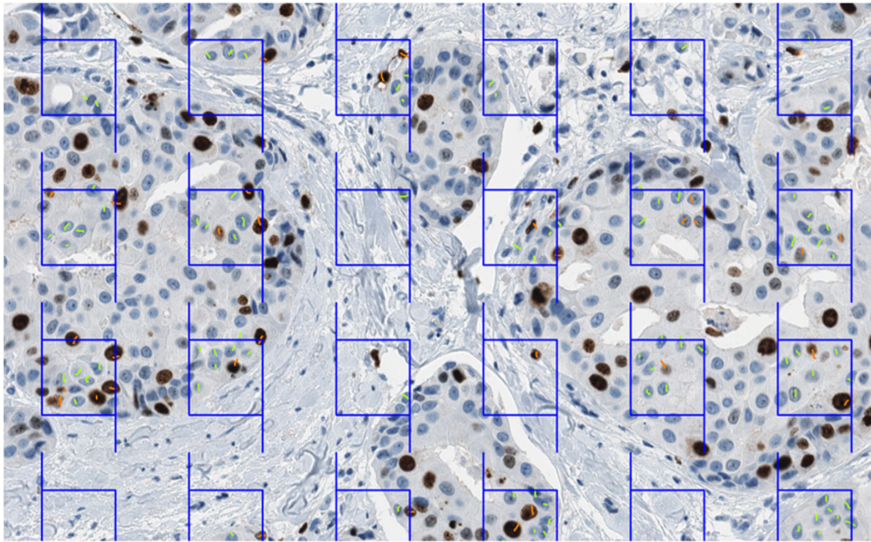


Figure 5. Test grid of frames from the stereology module overlaid on the TMA spot image

The left and bottom lines of a frame are “forbidden” – the nuclear profiles intersecting them are not marked. The short line marks (orange for Ki67 positive, green for Ki67 negative tumor cell nuclear profiles) are produced manually by the pathologist. Total numbers and Ki67 LI are computed by stereology software at the end of the procedure [28].

2.3.2 Study II

Spatial and Multiparametric Analysis: Heterogeneity Measurements, Image Segmentation by a Hexagonal Grid

In this study, a methodology for comprehensive Ki67 labeling index estimation in breast cancer tissue samples stained for Ki67 IHC was developed. It is based on the systematic subsampling of digital image analysis-generated data/images into smaller areas (hexagons), thereby enabling the computation of texture and distribution indicators for Ki67 LI intra-tissue variability. Each WSI image was subsampled by hexagons of 825pixel size, which corresponds to 0.75 mm circular diameter and 0.4421 mm² area. Hexagonal tiling (HexT) was generated to fit the area of the region of interest (ROI) in an image, and the individual nuclei extracted by DIA were assigned to an appropriate hexagon based on their coordinates.

Hexagons containing no nuclear profiles by DIA were regarded as missing data; hexagons containing fewer than 100 nuclear profiles were regarded as insufficiently sampled. A minimum requirement of 20 informative hexagons per tumor was applied in further analyses. Local Ki67 LI was calculated for each hexagon. DIA results represented by Ki67-negative and Ki67-positive tumor cell nuclei with their X and Y coordinates in the WSI were partitioned into HexT, from which intra-tumor variance indicators (Haralicks' texture parameters such as heterogeneity, entropy, dissimilarity, energy, homogeneity) were computed. The HexT data was also used for automated detection and quantitative evaluation of Ki67 LI hotspots that were based on the upper quintile of the HexT data. This indicates approximately 20% of the tumor tissue area revealing the highest biomarker expression and was conceptualized as the "Pareto hotspot." A visual representation of the of tumor analysis performed by the HexT approach and compared to pathologist visual estimation is presented in Figure 6.

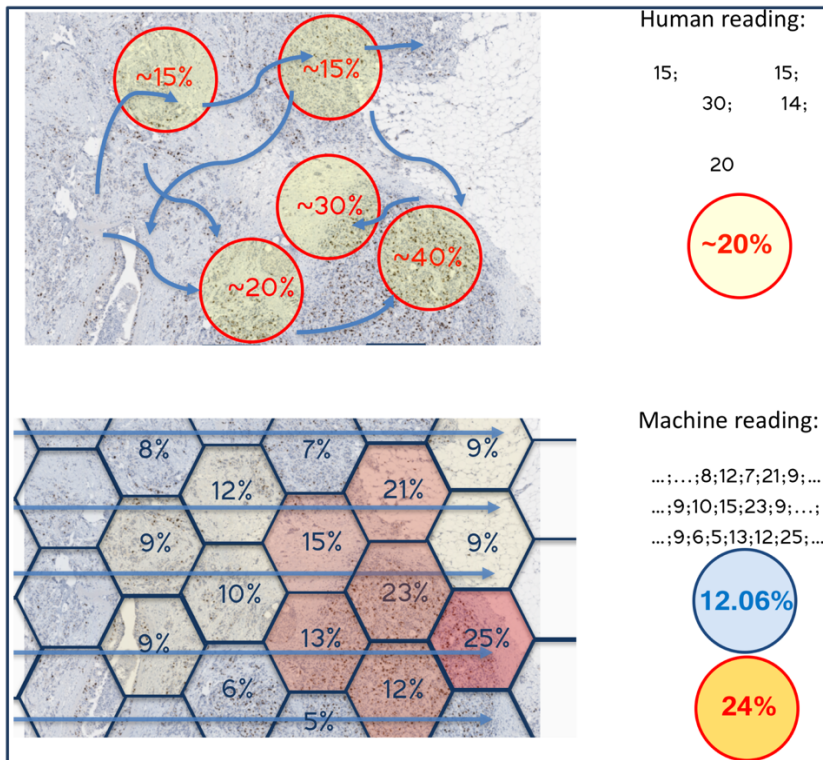


Figure 6. An improvised comparison of a visual and computerized estimation of Ki67 LI

The upper image shows a random area selection for a Ki67 LI estimation made by the pathologist. The lower image illustrates a systemic subsampling of the scanned IHC image and reveals the full area coverage by digital analysis software for an accurate Ki67LI estimation.

The automated Ki67 LI hot spot detection was validated by a visual review. Three pathologists independently reviewed 50 randomly selected WSI at low magnification and drew as many as 3 freeform annotations to delineate the Ki67 hotspots in the tumor tissue, if present. An inter-observer agreement of the visual hotspot detection was evaluated. The hotspot annotations from each observer were compared to the corresponding HexT data [122].

Summary statistics and distribution analyses were performed with significance tests based on a paired t-test, one-way ANOVA and Duncan's multiple range test for pairwise comparisons. A Fisher's exact test was used to estimate significant associations in non-parametric statistics. Inter-observer agreement was tested by kappa statistics. Pearson correlations and single and multiple linear regression analyses were performed to test pairwise linear relationships. A factor analysis was performed using the factoring method of principal component analysis. A cluster analysis was performed using the K-means algorithm [122].

2.3.3 Study III

Hexagonal Tiling Simulation for Optimizing Breast Cancer Tissue Sampling Requirements for Ki67LI Representation

A hexagonal tiling approach [122] was exploited to create a tissue sampling simulation model and test the precision of the construction of tissue microarrays and breast cancer tissue sampling, taking intra-tumoral heterogeneity into account. The hexagons in the HexT were chosen to simulate virtual TMA cores (or corresponding fields of view in conventional microscopy) with numbers of Ki67 positive and negative cells established by DIA. The patient cohort was chosen to be the same as described in paper II, [122], (n=297).

The coordinates of positive and negative nuclei extracted by the DIA were distributed into a dense HexT overlaid on each WSI. The HexT was

randomly positioned within the invasive tumor area, Figure 7. Local Ki67 LI was calculated for each hexagon to construct co-occurrence matrix used to compute Haralick texture parameters. The individual hexagons, with local Ki67 LI, were subsequently used as TMA cores for the random sampling simulations (Figure 7) and approximately resembled a TMA core of 0.75 mm circular diameter and 0.44 mm^2 area. The tumors were dichotomized into homogeneous and heterogeneous groups based on the median entropy value obtained by the HexT methodology. The sampling simulations were carried out for all three tumor classes: all/mixed, homogeneous and heterogeneous.

Two different methods were used to simulate the impact of the number of hexagons/TMA cores on the precision of the sampling to represent the Ki67 LI reported by the DIA of the entire region of interest (ROI). First, the practice of the “physical” construction of TMA, in which a set of cores are sampled only once, was simulated by randomly sampling a subset of hexagons once. Single linear regression analysis was used to compare the data in a single random selection.

Second, an error analysis was conducted by simulating many samplings of TMA subsets with hexagon numbers (HexN) of sizes $\text{HexN} = (1, 2, \dots, 15)$ per case. Each subset is sampled from the set of hexagonal tiles without replacement, but all hexagons are replaced before sampling a new subset. From the resulting sampling distribution, error measurements and other statistics can be inferred. Here, the simulations were used to infer the coefficient of error (CE) of Ki67 LI predictions using subsets differing in the number of virtual cores. The interpretations of error analysis results are made according to a putative CE value of 10% for accessible results for practical applications.

Both experiments were grouped by tumor heterogeneity and repeated for $\text{HexN} = (1, 2, \dots, 15)$ with hexagons resembling a 0.75 mm diameter TMA core and the simulations were performed with 50 000 iterations [123].

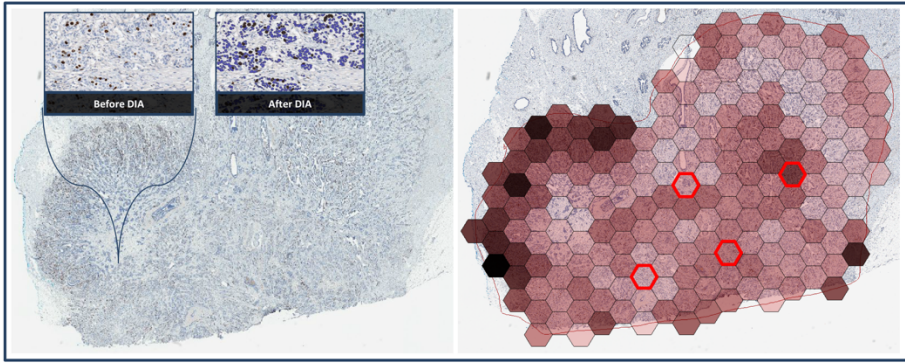


Figure 7. The hexagonal tiling of digital image analysis data for tissue subsampling simulations

Left: Ductal breast carcinoma stained by Ki67 IHC. Overlay showing high resolution tissue before and after digital image analysis. Right: Tumor with results of overlaid hexagonal grid for TMA simulation. Ki67 LI indicated by fill color. Light gray is low Ki67 LI with darker brown showing larger Ki67 LI. Red hexagons illustrate one possible subsampled set of four hexagons (HexN=4).

2.3.4 Study IV

Prognostic Value: The Comprehensive Ki67 LI for Predicting the Overall Survival of the Patients

The hexagonal tiling method [122] for the computation of intra-tissue heterogeneity parameters was tested on a different patient cohort (Nottingham, UK) with the patients' overall survival (OS) data available. HALO™ Classifier Module and CytoNuclear v1.4 algorithms were manually calibrated for the best tumor tissue recognition and cell segmentation. The HexT data (represented by the local Ki67% values and their coordinates in the WSI) were used to compute texture and distribution indicators as well as the “Pareto hotspots” for individual tumors. Four observers independently reviewed all WSI at low magnification and annotated up to three freeform areas to delineate the Ki67 hotspots in the tumor tissue within the invasive tumor component, if present. An inter-observer agreement of the visual hotspot detection was evaluated. Each observer provided a semi-quantitative score of Ki67% in the tumor tissue represented by average Ki67% and hotspot Ki67%, if detected. The final Ki67% score was calculated by substituting the average Ki67% by a hotspot

Ki67%, if established. The final Ki67% scores of the individual observers were averaged for further analyses (Ki67 Obs Mean). All results were then compared to patients' OS data.

Summary statistics and distribution analyses were performed with significance tests based on the paired sample t-test, one-way ANOVA with a Bonferoni test for pairwise comparisons. Chi-squared and Fisher's exact tests were used to estimate significant associations in non-parametric statistics. The inter-observer agreement was tested with kappa statistics. A factor analysis was performed using the factoring method of principal component analysis. Product-limit estimates were used to summarize overall survival data, and a log rank test was used for comparing OS distributions. A cox proportional hazards analysis was used to develop a multiple variable model to predict time to death. Continuous variables were dichotomized to predict OS using the web-based tool "Cutoff Finder" [131].

3. RESULTS

3.1 Study I

Digital Image Analysis: Calibration, Quantification and Validation

The Characteristics and Measurement Uncertainty of the Reference Value Dataset

Summary statistics of the reference value ($n = 30$), obtained by marking the tumor cell profiles in the test grid by three independent observers, are presented in Table 4. No significant variance between the three Ki67-Counts was revealed by one-way ANOVA ($F = 0.08$, $P = 0.9217$), while strong pairwise correlation among the values was found: $r = 0.98$, $r = 0.98$, $r = 0.97$ ($P < 0.0001$). Similarly, the total number of nuclear profiles marked did not differ significantly and strong pairwise correlation among the values was found: $r = 0.94$, $r = 0.98$, $r = 0.98$ ($P < 0.0001$).

Table 4. Summary statistics of the reference values produced by independent observers' markings ($n = 30$)

Variable	Median	Mean	Std dev	Std error	Min	Max
Ki67-Count-1	21.7	28.6	20.4	3.7	0.3	72.6
Ki67-Count-2	24	29.9	19.5	3.6	0.6	69.7
Ki67-Count-3	23	28.7	18.6	3.4	1.2	69.4
Ki67-Count- median	24	29.3	19.4	3.5	0.6	67.4
Ki67-Count- mean	23.4	29.1	19.4	3.5	0.7	66.8
Total profiles Observer 1	331	425.7	273.7	50	85	1,098
Total profiles Observer 2	509	590.7	385.4	70.4	143	1,863
Total profiles Observer 3	471.5	547.2	331.9	60.6	146	1,544
Ki67-VE-1	10	18.3	15.3	2.8	5	70
Ki67-VE-2	30	40.2	29.4	5.4	2	95
Ki67-VE-3	37.5	41.4	27.7	5.1	1	90

Ki67-VE-4	20	30.2	23	4.2	4	80
Ki67-VE-5	22.5	31	24.1	4.4	1	90
Ki67-VE-	22.5	32.5	25	4.6	2	90
median						
Ki67-VE-mean	23.4	32.2	23.2	4.2	6.2	80
Ki67-DIA-0	16.1	19.9	12.5	2.3	2.1	50
Ki67-DIA-1	18.5	24.8	15.9	2.9	1.6	65.5
Ki67-DIA-2	22.8	29.1	15.7	2.9	9.1	68.4

Note: Along with the results of cell markings, other measurements (visual assessments and DIA) are presented for a reference.

Uncertainty introduced by variance among the three observers counts to produce Ki67-Count for each individual spot was low: for the 30 spots, mean standard deviation and mean standard error were 2.6% and 1.5%, respectively. The agreement within the same confidence interval among all three measurements was 69%; whereas the pairwise agreement varied from 83% to 86%. The uncertainty of the RV generated was therefore considered satisfactory. The RV for the whole image dataset (n = 164) were based on a single observer count per spot (Ki67-Count). Yet, the subsampling uncertainty was further taken into account in the accuracy estimates.

A Comparison of Image Analysis Results and Visual Estimates to the Reference Values

Summary statistics of the RV, DIA and VE variables (n = 164) are presented in Table 5. One-way ANOVA revealed significant variance explained by the measurement method overall (P < 0.0001). Pairwise comparisons revealed no significant bias among the Ki67-Count and Ki67-VE-2 and Ki67-VE-3 estimates or Ki67-VE-5, Ki67-VE-median and Ki67-DIA-2. Meanwhile, Ki67- DIA-0, Ki67-DIA-1, Ki67-VE-1 and Ki67-VE-4 produced significantly lower values.

Pairwise correlations were highly significant (P < 0.0001). Remarkably, correlation between Ki67-Count and Ki67-DIA-0, 1 and 2 improved which each calibration cycle from r = 0.928 to r = 0.949. Notably, Ki67-Count correlated with the Ki67-VE-median strongest (r = 0.930), in comparison to the correlations with the individual VE measurements.

Single linear regression analyses for the DIA and VE results as dependent variables and the RV as explanatory variables produced highly significant ($P < 0.0001$) models in all cases. Remarkably, the determination coefficients (R-square) improved with each calibration cycle of the Ki67-DIA-0, 1, and 2 from 0.86 to 0.89 and 0.90. Notably, R-square for the VE-median (0.86) was the highest amongst the individual VE but reached only that of the Ki67-DIA-0.

Table 5. Summary statistics of the reference values produced by three observers with the corresponding data of visual estimates and digital image analysis (n = 164)

Variable	Median	Mean	Std dev	Std error	Min	Max
Ki67-Count	35.0	40.2	25.3	2.0	0.6	98.1
Ki67-DIA-2	30.1	36.5	20.2	1.6	6.4	93.0
Ki67-DIA-1	24.1	31.1	21.1	1.6	1.5	90.5
Ki67-DIA-0	20.4	25.9	18.1	1.4	2.1	85.7
Visual median	30	37.2	27.4	2.1	2	95
Visual mean	28.4	36.2	25.6	2.0	2.2	96.4
Ki67-VE-1	15	24.3	23.6	1.8	5	95
Ki67-VE-2	40	43.4	29.6	2.3	2	98
Ki67-VE-3	37.5	44.1	30.0	2.3	2	98
Ki67-VE-4	22	31.6	24.3	1.9	1	95
Ki67-VE-5	30	37.7	27.7	2.2	1	100
Total profiles observer *	2,372	2,658.7	1,390.4	108.6	464	7,452
Total profiles DIA-2	2,150.5	2,293.2	796.8	62.2	752	4,302
Total profiles DIA-1	1,920.5	2,022.7	670.1	52.3	1,012	3,788
Total profiles DIA-0	4,203.5	4,385.0	1,420.2	110.9	1,640	7,939

*Total nuclear profiles observer counts are multiplied by four in this table to be comparable to the DIA total profile numbers (the box grid used for the

observer count covers fourth of the image area). DIA, digital image analysis; VE, visual estimate.

The correspondence between the Ki67-DIA-2 and the RV was also tested, taking into account the uncertainty of the RV related to the subsampling of the tissue by the test grid. The confidence interval for the RV was calculated and the Ki67-DIA-2 values were tested for fitting the confidence interval. The R-square of the model was 0.90, the accuracy factor was 0.82. Interpretation of the plot and the slope tilt from the ellipse axis revealed a bias: an underestimation of the Ki67-Count by the Ki67-DIA-2 was observed at the higher end of the RV scale, as well as an overestimation at the low end.

The Prediction of the Reference Values by an Inverse Regression and Measurement Error Correction

Ki67-DIA-2 enabled fair accuracy and outperformed the 5 VE measurements, both individual and the median. Yet, the measurement bias for the Ki67-DIA-2 was established and enabled a measurement error correction procedure to be used to predict the ground truth in real life with maximum accuracy. Inverse regression analyses were performed to retrieve the correction criteria (Table 6). To avoid the potential impact of some non-linearity that was noted and to derive the most useful inverse regression model for an accurate prediction of the ground truth in the interval of clinical importance, a regression model $\text{Ki67-DIA-2} < 40$ was produced, based on the observations with Ki67-Count values less than 40% ($n = 92$). In addition to the single regression models, multiple regression models with inclusion of both Ki67-DIA-2 and Ki67-VE-Median gave slightly higher R-square value (0.91) than the Ki67-DIA-2 alone (0.90). Therefore, the DIA approach with the calibration of the algorithm settings-based quantified bias enabled for a most accurate measurement of the Ki67 LI, while the VE of five pathologists were consistent but gave little added value in terms of accuracy, as compared to the automated DIA measurement.

Table 6. Single and multiple linear inverse regression models for predicting reference values as a dependent variable (n = 164, P <0.0001 for all models and slope estimates)

Variable	R²	Intercept estimate	Intercept P	Slope estimate	Slope standardized estimate
Single regression models:					
Ki67-DIA-2	0.90	-3.1183	0.0165	1.1878	0.9494
Ki67-DIA-2<40*	0.75	-4.3913	0.0085	1.1472	0.8688
Ki67-DIA-1	0.89	5.0453	<0.0001	1.1309	0.9447
Ki67-DIA-0	0.86	6.8232	<0.0001	1.2916	0.9278
K67-VE-median	0.86	8.3195	<0.0001	0.8572	0.9302
Multiple regression model					
	0.91	-0.3245	0.8096		
Ki67-DIA-2				0.8068	0.6448
K67-VE-median				0.2985	0.3239

*Ki67-DIA-2 < 40 - represents a regression model for Ki67-DIA-2 with only Ki67-Count less than 40% cases included in the analysis (n = 92). DIA, digital image analysis; VE, visual estimate.

The Effect of the Prediction and Measurement Error Correction on Ki67 Dichotomization Accuracy

The effect of VE and DIA inverse regression models to predict the RV on the accuracy of patient dichotomization at RV cut-offs of clinical importance (>10, 15 and 20%) was also tested (see [28] for detailed information). A total misclassification rate at different cut-offs varied from 11 to 18% for the VE-based and 5 to 9% for the DIA-based prediction, respectively. In summary, the DIA-based prediction of the RV enabled the classification error rate half of that of the VE- based prediction, it was less than 10% at all

cut-offs tested, and could be further improved by the attempts at measurement error correction.

3.2 Study II

Spatial and Multiparametric Analysis: Heterogeneity Measurements, Image Segmentation by a Hexagonal Grid

Correlation of the Data from the Overall Image Analysis of WSI, from HexT and from Pathology Reports

The Ki67 LI calculated by the WSI DIA and the median Ki67 LI obtained from the HexT (HexSize825) data revealed perfect correlation ($r = 0.9967$, $p < 0.0001$) without any significant bias detectable by linear regression with the Hex median as the dependent variable ($r^2 = 0.997$, model $p < 0.0001$, intercept = -0.46808 , slope = 1.02341). The pathology report for Ki67 LI could be predicted from the WSI DIA Ki67 LI with some bias ($r^2 = 0.754$, model $p < 0.0001$, intercept = -4.16059 , slope = 1.21154).

The degree of proliferation measured by various Ki67 LI indicators was associated with higher histological grade and more aggressive types of breast cancer. More importantly, grade 3 tumors revealed less entropy than grade 1 or 2 tumors ($p = 0.0104$). This finding was also reflected by relevant ANOVA results where G3 tumors, presented with lower entropy values had compared to the G1 and G2 tumors ($p < 0.05$). It can be interpreted that high-grade tumors are more spatially homogenous with respect to their proliferative activity. The bimodality indicators did not reveal any significant clinic-pathological associations or relations to Ki67 LI values.

Factor Analysis of the Ki67 Indicators

Factor analysis was performed on 297 patients with a complete set of data obtained from the DIA of WSI and from the HexT analysis. Ki67 LI from pathology reports was also included in the data set. The rotated factor pattern of the 4 factors was extracted with eigenvalues of 5.8, 5.1, 2.3, and 1.9, respectively. Factor 1 was characterized by strong loading of the majority of the Haralick texture parameters; the strongest positive loading

was with the entropy indicator, and factor 1 was therefore named the entropy factor. Factor 2 was characterized by positive loading of the Ki67 LI indicators from WSI, HexT, the pathology report, and it was therefore named the proliferation factor. Similarly, factor 3 was named the bimodality factor, while factor 4 was termed the cellularity factor, based on the data obtained from both the WSI DIA and HexT data.

In summary, the factor and cluster analyses present evidence for two linearly independent features with respect to the intra-tumor heterogeneity of proliferative activity as measured by Ki67 expression. The first is based on entropy and other texture indicators, and the other is based on bimodality indicators.

Automated Hotspot Detection and Measurement. The Concept of the Pareto Hotspot

The DIA data, when subsampled into the hexagonal tiles, provided an opportunity to analyze the Ki67 LI distribution in the context of the 2D space of the tumor tissue. Furthermore, the tumor areas with high proportion of positive cells can be specifically highlighted to reveal hotspots of Ki67 expression. The Pareto principle (also known as the 80–20 rule) was applied, which states that for many events, approximately 80% of the effects come from 20% of the causes. We propose the concept of a Pareto hotspot, represented by the upper quintile of the biomarker expression in the tissue. This approach enabled us to highlight the most prominent areas of biomarker expression in the tumor tissue by overlaying the Hex with the fifth quintile of Ki67 LI on the tumor tissue image, forming the Pareto web (an example is presented Figure 8). This visualization approach does not depend on any assumptions about tumor heterogeneity, as the approximately 20% of the tumor tissue with the highest biomarker expression levels would be marked in all cases. Nevertheless, the relevance of the Pareto web must be appreciated in the context of the heterogeneity metrics for the individual tumor.

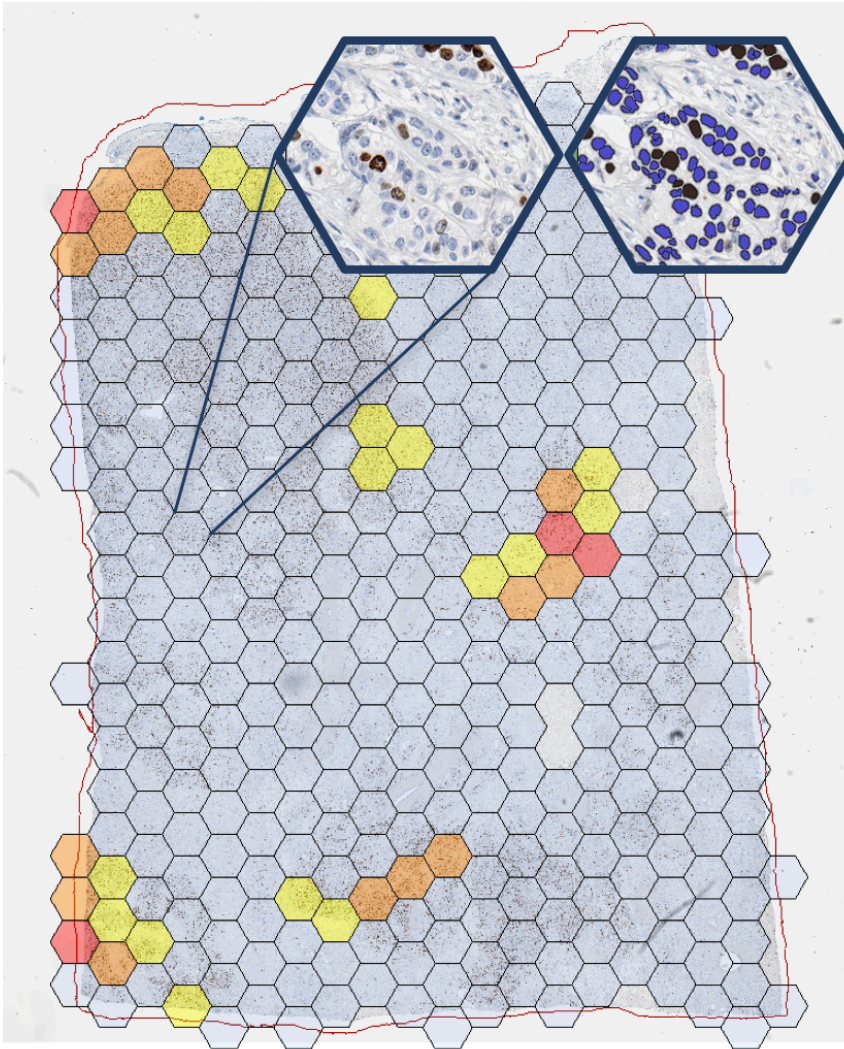


Figure 8. A visual representation of tumor analysis performed by the Hexagonal tiling approach

The hexagon grid is overlaid on the original WSI of Ki67 IHC to reflect subsampling of the DIA-generated data. The magnified hexagons illustrate side-to-side the details of the original (left) and DIA markup (right) images. The hexagon colors represent different ranks of the Pareto web, highlighting the upper fifth quintile of the HexT Ki67 LI distribution (for increased detail, yellow Hex represent the 80–90th percentile; orange Hex represent the 90–95th percentile; red Hex represent the 95–100th percentile).

The Validation of HexT Data Based on Hotspot Detection by a Visual Review of the WSI

Hotspots were identified in 20, 21 and 23 tumors by three pathologists after a visual investigation of 50 WSI. The agreement between the observers (taken pairwise) in detecting at least 1 hotspot was estimated by kappa coefficients of 0.55, 0.63, and 0.85. Consequently, the hotspots were identified in 27, 22, or 15 tumors by 1, 2 or all 3 observers, respectively. An analysis of the actual areas and hotspot overlaps, outlined by all 3 observers in the 15 tumors (as above), revealed that on average, hotspots represented 4.8% of the tumor area (range, 0.6 to 17.0 %). Meanwhile, on average, 26.0% of the hotspot areas coincided for all 3 observers (range: 1.7 to 70.8%). Pairwise comparisons revealed hotspot area overlaps of 42.0, 43.8 and 50.1%.

The hotspot annotations provided by the 3 individual observers revealed significantly higher Ki67 LI values by a paired t-test (mean differences of 8.4%, 8.7%, and 10.1%; $p < 0.0009$, $p < 0.0008$, and $p < 0.0003$, respectively) compared to the remaining area of the same tumors. The mean differences in the hotspot Ki67 LI between the observers were not significant. The mean hotspot Ki67 LI from all 3 observers was not significantly different ($p = 0.0675$) from the Ki67 LI 90th percentile (the median of the Pareto hotspot).

3.3 Study III

Hexagonal Tiling Simulation for Optimizing Breast Cancer Tissue Sampling Requirements for Representing Ki67LI

To achieve a $R^2 = 0.95$ value in the single linear regression models, a random selection of at least four, three and twelve cores were required in the mixed, homogeneous and heterogeneous tumors, respectively.

The mean coefficient of error for Ki67 LI estimates are plotted for increasing TMA core numbers in the tumor subgroups (Figure 9). To achieve the CE of 10%, 8 cores 0.75 mm in diameter were required in the mixed group of tumors. Respectively, 5–6 or 11–12 cores were required in the subgroups of homogeneous and heterogeneous tumors.

To achieve a CE of 10%, approximately 4 000 nuclei were required in the mixed group of tumors, as depicted in Figure 10. For the subgroups of

homogeneous and heterogeneous tumors to reach the same error, 3 000 and 7 000 nuclei were necessary, respectively.

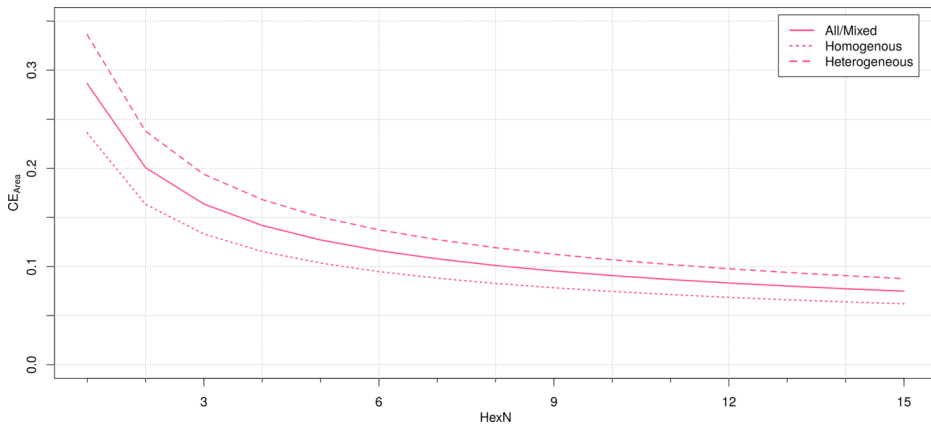


Figure 9. Error results as function of tissue area evaluated

The resampling procedure was simulated for each individual tumor case using 50 000 iterations for each count of hexagons (HexN). The analysis results are split by the tumor heterogeneity level. Error measurement (Coefficient of error) is expressed by mean of all cases.

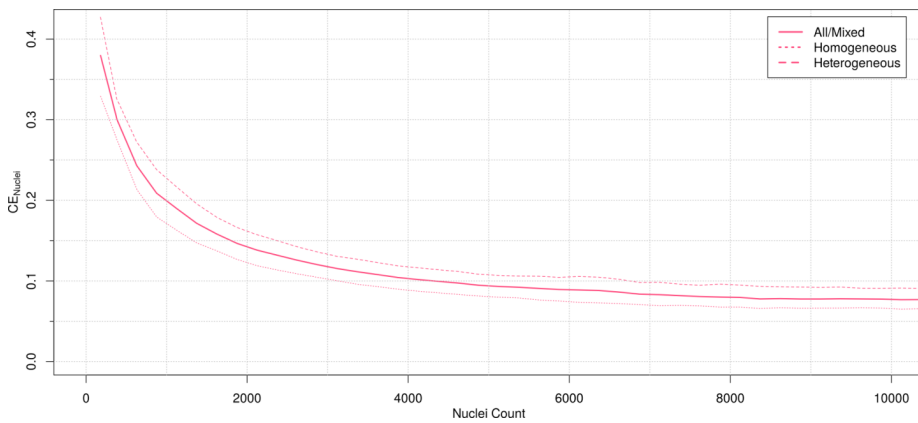


Figure 10. Error results as a function of nuclei counted

The coefficient of error plotted as a function of nuclei count. See text for transformation of TMA by the core number to nuclei count. The analysis

results are split by the tumor heterogeneity level. Error measurement (Coefficient of error) is expressed by a mean of all cases.

3.4 Study IV

Prognostic Value: The Comprehensive Ki67 LI for Predicting the Overall Survival of the Patients

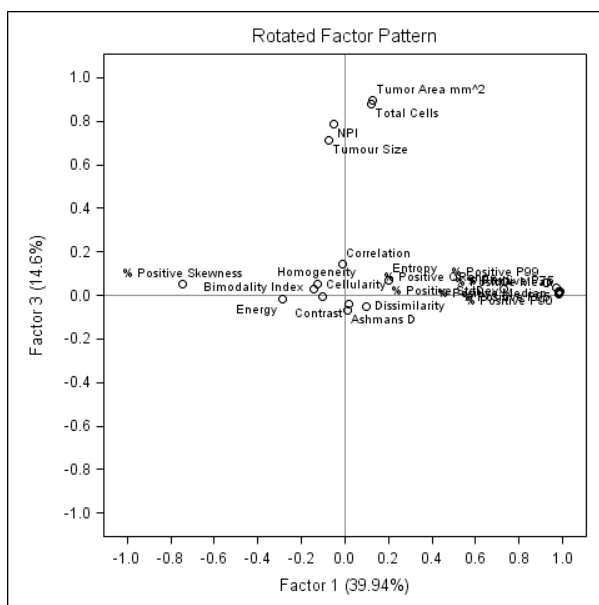
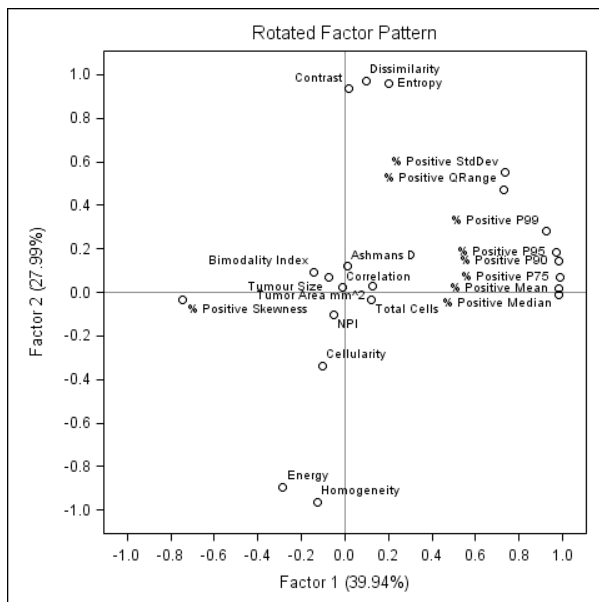
Hotspot Detection by a Visual Review of the WSI

Four observers reviewed 152 WSI, and at least one hotspot was identified in 37, 67, 32 and 27 tumors by each investigator, respectively. The agreement between the observers (taken pairwise) in detecting at least one hotspot was estimated by kappa coefficients ranging from 0.20 to 0.50. An analysis of the actual areas and hotspot overlaps, outlined by 2 or more observers in the 46 tumors, revealed that, on average, 24.4, 13.9, and 4.4% of the hotspot areas coincided between the 2, 3, and all 4 observers, respectively. The tumors with hotspots detected by at least two observers were characterized by higher entropy ($p < 0.03$), higher correlation ($p < 0.05$) and lower energy ($p < 0.02$) values but did not differ with regard to the other Haralick or bimodality indicators.

A Factor Analysis of the Comprehensive Ki67 Indicators

A factor analysis was performed on 152 patients with a complete set of DIA HexT data along with selected pathology data. The rotated factor pattern of the 5 factors, extracted with eigenvalues of 8.8, 4.2, 2.8, 1.8, and 1.3, respectively (Figure 11). Factor 1 was characterized by positive and very similar loadings of the various Ki67% indicators and was best interpreted as the “proliferation” factor. Factor 2 was characterized by strong positive loadings of the Haralick indicators of “disordered texture” (contrast, dissimilarity, entropy) and negative loadings of energy and homogeneity. Factor 3 was characterized by positive loadings of reflective of tumor sample size evaluated by DIA and pathology report along with the NPI. Factor 4 was represented by both bimodality indicators, while factor 5 was characterized by the correlation parameter and cellularity of the tumor. Associations of the tumor Ki67 indicators and the factor scores with relevant tumor characteristics were explored by ANOVA. In particular, the

histological grade (G) was associated with higher factor 1 ($p < 0.0001$) and factor 3 ($p < 0.0001$) scores as well the corresponding primary variables. Factor 2, 4, and 5 scores did not reveal any significant associations.



The Predictors of the Overall Survival of the Patients

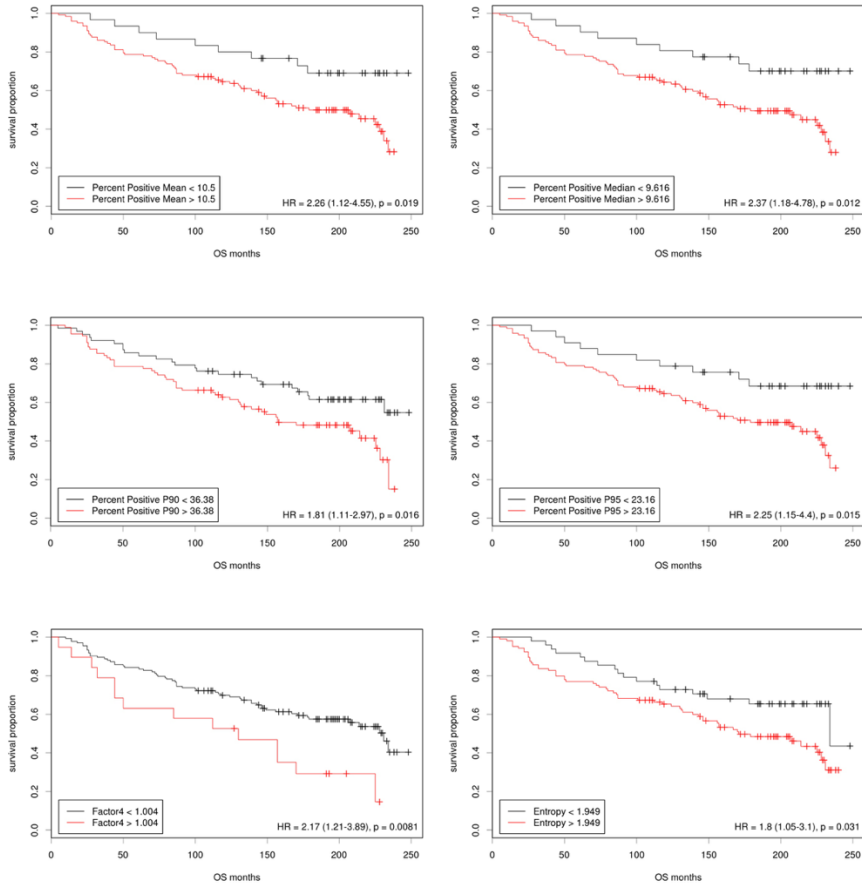
Several multivariable models were developed to account simultaneously for the comprehensive Ki67 indicators and other characteristics of the tumors to predict OS (Table 7). Model Nos. 1 and 2 revealed an independent prognostic value of worse OS for Ki67 bimodality indicators (Ashman’s D or factor 4 scores) in the context of HR and HER2 positivity. Remarkably, neither chemotherapy, nor none of the Ki67 indicators of the level of proliferative activity (Ki67 LI, Ki67 Obs Mean, Ki67 WSI or Ki67 HexT Mean, median, percentiles) could be verified as significant independent predictors of OS in this dataset.

Table 7. Cox multivariate regression models for predicting the overall survival of the patients

		Hazard ratio	95% confidence limits	P value
Model	no.			
(n=147)	1			0.0048
HR positive		0.662	(0.504, 0.869)	0.0030
Ashman’s D		1.320	(1.035, 1.685)	0.0254
Model	no.			
(n=147)	2			0.0008
HR positive		0.645	(0.489, 0.851)	0.0019
HER2 positive		2.178	(1.016, 4.669)	0.0455
Factor 4		1.592	(1.186, 2.186)	0.0020
Model	no.			
(n=141)	3			0.0030
HR positive		0.501	(0.359, 0.700)	0.0001
HER2 positive		2.800	(1.248, 6.279)	0.0125
Ashman’s D		1.322	(1.030, 1.724)	0.0288
Chemotherapy		0.384	(1.184, 0.801)	0.1107

The Ki67 indicators and factor scores were dichotomized using the web-based tool “Cutoff Finder” [131] and were analyzed using Kaplan–Meier estimates and log rank tests (Figure 12). Many indicators allowed for a

significant dichotomization of the patients into the prognostic subgroups. The bimodality of Ki67 intratumor expression, represented by factor 4 scores ($p = 0.0081$) and Ashman's D ($p = 0.0017$), provided significant cut-off values for predicting OS. The level of proliferative activity, represented by a broad range of indicators (factor 1 scores, Ki67 HexT Mean, median, percentiles, Ki67 LI, Ki67 Obs Mean, positive cell density) served as a significant single predictor as well.



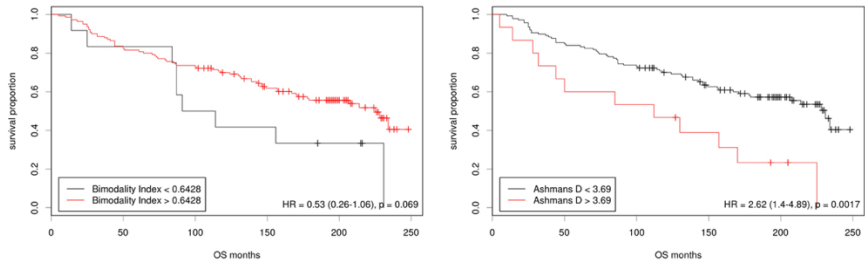


Figure 12. Cut-off values for the Ki67 indicators as single predictors of overall survival

DISCUSSION

This study reveals several aspects that are crucial for improving the prognostic value of the Ki67 proliferative tumor activity index, which plays an important role in personalized therapy decisions. First, the accuracy of the Ki67 labeling index, obtained by digital image analysis, is higher than that of pathologists' visual estimates and could be further improved by the measurement error correction procedures. More importantly, image analysis algorithms were calibrated by comparing results with reference dataset obtained by stereological counts. Second, the intra-tumor heterogeneity of proliferative tumor activity could be visualized by hexagonal tiling approach in the whole slide image along with automated detection and quantitative evaluation of Ki67LI hotspots. Additionally, the HexT approach was utilized to optimize tissue sampling requirements for breast cancer tissue to represent Ki67 LI taking its intra-tumor heterogeneity into account. Finally, the spatial heterogeneity indicators of proliferative tumor activity, measured by DIA of Ki67 IHC expression and analyzed by the HexT approach, can serve as an independent prognostic indicator of OS in breast cancer patients and outperform the prognostic power of the level of proliferative activity.

Digital image analysis and its application to digitalized immunohistochemistry slides has become a rapidly developing approach in a field of pathology research. The greatest benefit of DIA lies in the rich, multi-parametric, tissue-related data that can be retrieved by processing high-resolution scanned microscopy images. DIA enables the higher analysis capacity from IHC slides than could ever be achieved by manual counts. For instance, a huge number of cells were analyzed in various experiments within these thesis: a total of 36 million cells with an overall tumor area of 15 000 mm² and 13 million cells in the tumor area of 6 000 mm² were evaluated in experiments described in study II and study IV, respectively. In study II, 121 000 tumor cell profiles were analyzed per IHC slide on average, with a range from 11 000 to 419 000 cells. From this point of view, the retrieval of only one indicator (such as Ki67 LI) in WSI from each tumor section can be regarded as a substantial underutilization of the data.

Digital image analysis not only brings a higher analysis capacity, reproducibility and accuracy compared to manual counts, but it can also serve as an additional tool for individual risk prediction for the patients. An automated scoring of immunohistochemistry reveals stronger prognostic

stratification of patients compared to visual evaluation [132]. A recent paper revealed that micro-architectural features established by DIA can indicate a different metastatic potential of various types of breast carcinoma [133]. Stålhammar G et al. indicated that the DIA of Ki67 LI estimation outperformed manual mitotic counts, the visual estimation of Ki67 LI as well as other markers of cell proliferation, and added significant prognostic information [32]. Furthermore, an automated scoring of Ki67 contribute significantly to the multigene models that predict the risk of recurrence in breast cancer with high accuracy and sensitivity [15].

Despite the broad DIA application to improve immunohistochemistry interpretation, the validation procedures of DIA algorithms are not frequently used. The major disadvantage of DIA experiments described in some research papers is that a pathology report data or subjective visual estimations are chosen to be as a criterion standard. The DIA quality and accuracy needs to be assured before the implementation into clinical practice, much like the sources of variation must be clarified. In the first part of this work (Study I) [28], the aspects of analytical validation procedure, accuracy and quantification of the measurement bias were addressed. We aimed to develop a DIA validation and calibration methodology for automated KI67 LI estimation by comparison to the Ki67 LI obtained on the same images by stereological counts as the most appropriate “gold standard.” Results showed that visual assessment made by two pathologists produced significantly lower values and the median of visual assessment by five pathologists did not reach the accuracy obtained by the calibrated DIA tool. We also tested the potential clinical impact of the accuracy achieved by applying DIA- and VE-based predictions of Ki67 LI to dichotomize patients by frequently used cut off values at 10%, 15% and 20% and found that DIA enabled the classification error rate two times lower than that of the visual assessment. These findings are an additional evidence that single visual estimations, or “eyeballing,” cannot be used as a reliable measurement or as a reference values for DIA validation purposes when the clinical requirements for biomarker quantification accuracy are needed. Furthermore, our results support the notion that inter-observer concordance of VE is relatively low, especially in intermediate Ki67 LI group, which is the most important for making clinical decisions, as found by Ruohong et al. [118].

In this experiment (Study I), we have also found that the global bias of the DIA became not significant only after the second (quantitative)

calibration step and revealed an improvement in the prediction of the DIA outputs with each calibration step. Consequently, it induces that the initial step in DIA experiments and/or implementation into laboratory information system (LIS) should be the validation process of a selected DIA tool. The measurement accuracy can be further improved by estimating the measurement bias from the criterion standard and by adjusting the DIA tools accordingly.

In the second part of this work (Study II), a methodology for measuring spatial distribution and texture parameters in tumor sections stained for Ki67 IHC are described. As expected, the technique that is based on the systematic subsampling of scanned IHC slide into the hexagonal tiles and enabled to compute Ki67 LI intra-tumor variability parameters. The approach provides numerous benefits. First, multiple measurements of the IHC marker enable the application of distribution statistics from a single DIA run. Secondly, data of biomarker expression in 2D space enable the calculation of texture indicators in the region of interest that reflect the global measure of intra-tumor heterogeneity; these indicators, along with the distribution statistics, can be used for inter-tumor comparisons and the stratification of the tumors into homogeneous and heterogeneous categories. In similar studies, the heterogeneity assessments were made by comparing different samples taken from the same tumor. For this purpose, physical TMA sampling based techniques are used to measure the discordance of biomarker expression between several regions (TMA cores) in the tumor [98, 134, 135]. While some differences of IHC staining pattern could be identified within the tumor in TMA based studies, the confirmed heterogeneity levels of the ER, PgR, HER2 expression are uncommon (<10% of cases) when using this method [134]. The major disadvantage of a TMA-based technique is that it directly depends on tissue sampling and does not reflect the diversity of the whole tumor tissue.

Similarly to this thesis, digital image analysis applications for heterogeneity detection have been described by Potts et al. [96], who investigated the breast cancer heterogeneity of HER2 immunohistochemistry. The authors applied diversity indices used in field in the ecological sciences to evaluate cell-level and tumor-level heterogeneity in HER2 IHC tissue sections, which were analyzed by DIA. A heterogeneity heat-map for visualizing individual tumor heterogeneity and HER2 expression levels was developed. More importantly, they recognized that the number of sampled regions might be insufficient to make determinations of

tumor level heterogeneity; thus, the use of a methodology that samples the entire tumor sample on a slide may be required for this type of analysis. The hexagonal tiling approach, which was developed during this work, is a methodology that relies on systematic subsampling of automated DIA-generated data by regular polygons in arrays to measure and visualize the spatial intra-tumor texture/heterogeneity of IHC biomarker expression in whole slide image. We must note that the various percentile ranges obtained from the HexT distribution statistics may prove to be more biologically relevant and clinically useful indicators of tumor proliferative activity than a simple Ki67 LI average, especially in heterogeneous cases (the Pareto hotspot is one possible automated indicator that mimics the current clinical practice of Ki67 LI evaluation in hotspots). Furthermore, an automated highlighting of potential hotspots on the WSI (e.g., with a Pareto web) can serve as a decision-support and quality assurance tool.

In study II, multivariate analyses extracted 4 major factors of intra-tumor variance, defined as entropy, proliferation, bimodality, and cellularity. These factor scores were further used in cluster analysis, which outlined the subcategories of heterogeneous tumors with predominant entropy, bimodality, or both at different levels of proliferative activity. However, a stratification of the tumors into homogeneous and heterogeneous groups would require evidence-based definitions, preferably ones that reflect clinical outcomes. While formal definitions for bimodality do exist (for instance, if Ashman's D is more than 2), a bottom-up approach, based on the percentile distribution of the real data, could be also considered. For example, the tumors could be classified as heterogeneous if their entropy and/or bimodality indicators were in the upper quartile of the distribution.

The clinical utility of Ki67 IHC as a prognostic and predictive factor is obscured by both the lack of standardized measurement/sampling methodologies and the absence of hotspot definitions, which might be potentially achieved by DIA applications. Christgen et al. have investigated the impact of ROI size on Ki67 quantification by computer-assisted image analysis in breast cancer. After manual identification of the highly proliferative areas on WSI, they gradually increased the ROI size by expanding freeform annotations, based on the number of cells detected by the image analysis in the ROI, and showed that the median Ki67 index decreased from 55 to 15% by increasing the size of the ROI. This indicated a significant misclassification between low- and high-proliferative tumors dependent on the size the selected ROI. While manual Ki67 counts remain

the standard in clinical diagnostics, the authors proposed that the automated image analysis may be included as an optional add-on in selected cases and may help to standardize and document the hotspot size. In our experiments (studies II and IV), hotspots of Ki67LI were manually annotated by three pathologists and inter-observer agreement of the visual hotspot detection was evaluated and then compared to the corresponding HexT data. Results revealed a relatively low agreement in visual detection of at least one hotspot: kappa values ranging from 0.2 to 0.5 in study II (n = 50) and 0.55 to 0.85 in study IV (n = 152). Moreover, the size and shape of the hotspots and their spatial overlap varied greatly between the cases and observers. Similar results of high inter-observer variability in determining the HS and Ki67 LI calculations were found in some other studies [19, 22, 118, 136]. In our experiment, the HS detected by pathologists were associated with Hex containing higher Ki67 LI values and were comparable with the Pareto hotspot median Ki67 LI.

Regarding the matter of unstandardized Ki67 IHC tissue evaluation and TMAs construction requirements, the principles of hexagonal tiling was utilized for multiple virtual simulations of tissue sampling. The extraction of Haralicks spatial parameters allowed to take tissue heterogeneity into account. We must note that the Haralick entropy threshold value is not clearly defined. Therefore, the optimal method to split the dataset it into equal parts by median was chosen. The results were very surprising and revealed different numbers of TMA cores needed depending on tumor heterogeneity: to achieve a coefficient error of 10%, 5–6 cores for homogeneous cases, 11–12 cores for heterogeneous cases, in mixed tumor population 8 TMA cores were required. These results are not in line with previous studies where, most commonly, the recommended number of TMA cores varied from 1 to 4 [137-140]. The discordance of the results may be explained by the differences in sampling iterations. Previous studies were mainly based on physical tissue sampling [139-143] or DIA experiments with artificial TMA cores [138, 144, 145]; however, the sampling was limited by relatively low numbers of repetitions (usually single random sampling). On the other hand, the experiment described in study III was modeled to virtually subsample scanned whole slide images 50 000 times with each TMA core number, which enabled to compute statistically strong error estimates. Additionally, this experiment provides evidence for minimum cell counting requirements to achieve robust KI67 LI measurements. The current clinical guidelines on the minimal number of

cells to be counted are quite arbitrary, mostly set in the range of 500 and 2000 tumor cells [24]. However, to achieve adequate precision, at least 1000 cells are recommended, while 500 cells would be acceptable as the absolute minimum [24]. Importantly, our findings reveal that to achieve 10% CE, approximately 4 000 nuclei must be counted when the intratumor heterogeneity is mixed/ unknown. These cell counts are rather large to accomplish in clinical practice for all breast carcinomas but could be feasible for cases considered as “grey zones,” e.g., in the range of Ki67 LI 10–30% [12], which would require more precise measurements.

In this study, a novel hotspot detection method based on measuring a stable proportion of the tumor tissue at the high end of the range (90th percentile) and conceptualized as the “Pareto hotspot” is proposed. This method has many advantages when compared with other studies, which have been carried out to detect the hotspots of Ki67 LI immunohistochemistry in breast cancer and other tumors. While some investigators have employed tissue texture and density based metrics [96, 146], others have proposed automated algorithms for detecting the hotspots of biomarker expression [15, 22, 33, 34, 36, 37, 120]. Nevertheless, methodologies related to manual counts [35], multiple visual estimations with an application of diversity statistics [69] or based on tissue microarray construction [134, 135] are still used to assess the heterogeneity of biomarker expression. The potential of spatial statistics derived from regular grids are broadly used in other fields that are closely related to spatial modeling (for example, geography or ecology) and has been shown to be the most efficient way for mapping spatial variation. In medical research, a rectangular grid is most commonly used due to its relative mathematical simplicity. Gudlaugsson et al. [22] described a tool based on grid structure to identify the HS of Ki67 LI in breast cancer. The squares were selectively placed on the regions with subjectively high numbers of positive nuclei, and Ki67 LI was counted by DIA inside each square. They found that Ki67 scores of semi-automated hotspots yielded reproducible and prognostically significant results. However, this method directly depends on the manual selections of the HS area, and the Ki67 scores were based on the area but not on the number of nuclei – as it should be due to the possible variation in the density of cells. Other studies have also used image segmentation techniques by automatically applying rectangular grids [15] [37]. In a recent paper [15], Ki67 LI values obtained by DIA from rectangular grids were used to predict the Oncotype DX risk categories of patients in their breast cancer patient

cohort (ER/PgR-positive, HER2-negative, lymph node-negative, stages I to II). It was found that high Ki67 indices were significantly correlated with a higher Oncotype DX risk-of-recurrence group. The Ki67 index was the major contributor to a machine learning model which, when trained solely on clinico-pathological data and Ki67 scores, identified Oncotype DX high- and low-risk patients with 97% accuracy, 98% sensitivity and 80% specificity [15]. The latter experiments are partly similar to HexT approach for HS detection, as it is based on image segmentation into equal regions. In contrast, HexT exploits hexagonal tiles, which allows for almost perfect WSI segmentation and provides a better coverage of the HS area, which is usually complex and irregular in terms of shape, whereas the rectangular or circular grid shape has a side-effect that the tissue located at the corners of the frames will never be sampled. The use of hexagons does not suffer from this: the dense HexT ensures that all parts of the tissue are considered with the same probability. Additionally, the Pareto principle measures equal tumor tissue proportion and could be easily modified to another range or rules, depending on the purpose of the study. This simple approach is less sensitive to hotspot shape and size (for instance, a small and irregular HS might be missed by measuring Ki67 LI in top 5 square tiles containing at least 500 tumor cells) and allows to avoid tissue or staining artefacts, as it could be easily identified by reviewing analysis output. In general, an employment of DIA technologies for automated HS detection may be useful in several aspects: a) it could be used to standardize the size, shape, temperature, gradient and other characteristics of the hotspots and create proper HS definitions; b) the pathologists' daily work could be optimized by only reviewing automatically highlighted HS areas and the estimates instead of spending time on counting cells and trying to detect the hidden hotspots in the whole tissue sample; c) it could help to achieve a high reproducibility and objectivity in HS measurements; d) the goal of better patient stratification for individualized therapy might be reached.

The HexT experiment (Study II) was designed to prove the principle rather than to test the clinical utility of the HexT approach, and patient follow-up was not available in the current data set. To overcome this limitation, study IV [124] was designed, and the previously described HexT principle was applied with another DIA algorithm on a breast cancer patient cohort containing long-term follow-ups and survival data. In study IV [124], a broad set of Ki67 IHC parameters, representing the level of proliferation, the pattern of distribution in the tissue, bimodality and texture indicators

were tested in prognostic models along with conventional clinico-pathologic characteristics of the breast cancer patients. All visual and machine-generated indicators of the level of Ki67 expression in this study provided significant cut-off values as single predictors of OS. However, only bimodality indicators (Ashman's D, in particular) served as the independent OS predictors in the context of HR and HER2 status and outperformed the prognostic power of the level of proliferative activity [124]. This finding was unexpected, but it may have a practical impact. Due to the biological cancer tissue variation, it is challenging to achieve a consensus for Ki67LI clinically valid cut-offs and to ensure the analytic accuracy of IHC testing where high a precision of quantification is required. The DIA-based approaches and its derivate parameters, such as bimodality or other heterogeneity indicators, may prove to be robust and less sensitive to these variations. Our data suggest that the variability of intra-tumor proliferative activity may be a fundamental feature of tumor aggressiveness affecting the final outcome of the disease, even more important than the average level of proliferation in tumor tissue. At least it is an independent factor of the disease behavior.

The whole work contains several limitations. In study I, the guidelines for analytical test validation were not strictly followed, since the subject (IHC image) is different from the analytical test samples used in medicine. Firstly, the criterion standard was established by one observer markings, splitting the job (whole series $n = 164$) among four observers in approximately equal proportions. Since the inter-observer variability was found to be negligible in the testing set ($n = 30$), it was considered to further rely on one observer's counts. Second, the repeatability of the tests was not tested, as it would require time-consuming efforts to repeat the stereological count manually, and it was not the main focus to investigate the intra-observer agreement of stereological counts. Third, the DIA prediction accuracy was not validated on an independent dataset, since it requires another set of criterion standard data. In a hexagonal tiling experiment (Study II), relatively large surgical excision samples of breast cancer tissue were used, and the approach was not tested on core needle biopsy material. It remains to be investigated if small core biopsy samples are sufficient for texture statistics due to the potential lack of tumor tissue in relation to the applied Hex size. Additionally, the DCIS component was not excluded in our analyses. Although we did not find evidence that DCIS could significantly impact hotspot detection in our study, a clinical study design

would require manual or automated exclusion of DCIS. Despite these limitations, the validation step opens better perspectives to use automated DIA tools to investigate the tissue heterogeneity and clinical utility aspects of Ki67 and other IHC biomarker expression.

In general, the results of this thesis suggest that adequate accuracy levels of Ki67 LI measurements can hardly be achieved by manual counts and highlight the importance of high-capacity, computer-based IHC measurement techniques to improve the efficiency of testing. In addition, automated hotspot or tumor heterogeneity detection with standard definitions by DIA would provide another advantage compared to the visual evaluation by conventional microscopy or inspection of whole slide images. Most importantly, this work is an additional evidence that automated Ki67 LI and its heterogeneity indicators could potentially guide the clinical choices for breast cancer treatment.

CONCLUSIONS

1. A methodology for ensuring and improving the accuracy of the digital image analysis approach in breast cancer Ki67 immunohistochemistry was developed. Ki67 LI obtained by digital image analysis outperforms visual estimates, taking manual stereological counts as a reference value, and could be further improved by the measurement error correction attempts. Automated Ki67 LI allows to dichotomize patients at reference value cut-offs of clinical importance (>10, 15 and 20%) at a two-times lower misclassification rate compared to that of the visual assessment consensus of five pathologists.
2. The hexagonal tiling approach, based on the systematic subsampling of DIA-generated data into a HexT array, enables the computation of texture and spatial distribution indicators for Ki67 LI intra-tumor variability. Breast cancer cases could be dichotomized into homogenous and heterogenous with Ki67 LI based on these indicators. The HexT approach allowed to visualize Ki67 LI intra-tissue heterogeneity in the whole slide image along with an automated detection and quantitative evaluation of Ki67 hotspots, which were based on the upper quintile of the HexT data, conceptualized as the “Pareto hotspot.” Furthermore, this approach can potentially be applied to numerous different IHC markers and tissues as an effective way of reflecting intra-tissue heterogeneity for decision support and quality assurance.

3. Hexagonal tiling data provide a useful model for establishing tissue sampling requirements for biomarker studies and visual estimations, which depend on intra-tissue heterogeneity and must be determined on a peruse basis.
4. The spatial heterogeneity indicators (the bimodality status in particular) of proliferative tumor activity, measured by the DIA of Ki67 IHC expression and analyzed by the HexT approach, can serve as an independent prognostic indicator of OS in breast cancer patients and outperform the prognostic power of the level of proliferative activity.

PRACTICAL RECOMMENDATIONS

1. Analytical validation procedures, based on appropriate reference data, should be used for digital image analysis algorithms in research and diagnostic testing.
2. The heterogeneity indicators of a biomarker expression should be taken into account for Ki67 LI estimation. To achieve low error estimates, when evaluating Ki67 LI by cell counting, approximately 4,000 nuclei must be evaluated if the intratumor heterogeneity is unknown. In breast cancer cases of the lower proliferative activity (Ki67 LI<20%), a larger sampling is required to achieve the same error rates as for the highly proliferative tumors.
3. Tissue heterogeneity impact should be taken into account for selection of an optimal number of TMA cores for biomarker research studies: for Ki67 LI in breast cancer, the number of 5–6 TMA cores is sufficient for a homogeneous expression in the tissue, 8 cores for tumors with mixed heterogeneity and at least 11 cores for heterogeneous tumors.

FUTURE PERSPECTIVES

Implementing Digital Image Analysis into Clinical Practice

The biggest advantage and clinical utility of the DIA tools is that they could be implemented in laboratory information systems and could serve as decision-support tools for pathologists. Since the DIA algorithms are analytically/clinically validated and tested, they can be further introduced into LIS, which would be the final goal of this study. Besides the fact that this process was out of scope in this particular thesis, it is in progress in the National Center of Pathology (Vilnius, Lithuania), and a pilot study has been already run. The initial results indicated that the implementation process should be coordinated with pathologists by clarifying their expectations and informing them that a digital analysis will never replace human interaction but could assist and significantly reduce the amount of work.

Heterogeneity Definitions and Clinical Validation

The proposed hexagonal tiling approach enables multiple definitions of hotspots besides the Pareto principle; however, they were not investigated in the scope of the present study, since the HS definitions, as well as the various spatial heterogeneity parameters, would be best elaborated in the context of clinical outcome data. The DIA-generated spatial heterogeneity parameters can be computed with DIA-based data; however, the stratification of the tumors into homogeneous and heterogeneous groups would require evidence-based definitions – preferably ones that reflect clinical outcomes. The next steps of this study will include the mentioned tasks by expanding patient cohorts, collecting survival data and investigating the prognostic value of DIA-generated parameters.

The Applicability of the Hexagonal Tiling Approach

The hexagonal tiling approach was tested on the Ki67 immunohistochemistry of breast cancer cases. However, it could be easily adapted for other purposes or tissues. Neuroendocrine tumors of the gastrointestinal tract might be among the candidates, where automated and comprehensive Ki67 LI estimations would serve to facilitate laborious

manual attempts to measure the relatively low Ki67 LI. The cut-off values for clinical decisions are 1%, 3% and 20%, which is sometimes almost impossible to measure with only visual estimations. DIA approaches and the detection of hidden hotspots could facilitate this task and provide a better stratification of the patients. Regarding breast cancer, only a small amount of studies were made in investigating hormone receptor (ER/PR) or HER2 heterogeneity parameters in the context of patient survival data. This task could also be possibly covered by applying heterogeneity measurement methods, which are developed in this study. Nevertheless, the initial analytical validation, based on accurate reference value construction, of the DIA algorithms must be applied prior to this as well as any other experiments.

LIST OF PUBLICATIONS

This thesis is based on the results presented in the following articles:

1. Laurinavicius A, Plancoulaine B, Laurinaviciene A, Herlin P, Meskauskas R, Baltrusaityte I, **Besusparis J**, Dasevicius D, Elie N, Iqbal Y, Bor C: A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. *Breast Cancer Research*, 2014. 16(2): p. R35.
2. Plancoulaine B, Laurinaviciene A, Herlin P, **Besusparis J**, Meskauskas R, Baltrusaityte I, Iqbal Y, Laurinavicius A: A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data. *Virchows Archiv*, 2015;467 (6):711–722.
3. Laurinavicius, A, Plancoulaine B, Rasmusson A, **Besusparis J**, Augulis R, Meskauskas R, Herlin P, Laurinaviciene A, Abdelhadi Muftah A. A, Miligy I, Aleskandarany M, Rakha E. A, Green A. R, Ellis I. O: Bimodality of intratumor Ki67 expression is an independent prognostic factor of overall survival in patients with invasive breast carcinoma. *Virchows Archiv*, 2016. 468(4): p. 493–502.

4. **Besusparis, J**, Plancoulaine B, Rasmusson A, Augulis R, Green A. R, Ellis I. O, Laurinaviciene A, Herlin P, Laurinavicius A: Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer. *Diagnostic Pathology*, 2016. 11(1): p. 82.

Publications related to the content but not directly included in this thesis:

1. Laurinavicius A, **Besusparis J**, Didziapetryte J, Radziuviene G, Meskauskas R, Laurinaviciene A: Digital immunohistochemistry: new horizons and practical solutions in breast cancer pathology. *Diagnostic Pathology* 2013, 8(Suppl 1):S15
2. Plancoulaine B, Meskauskas R, Baltrusaityte I, **Besusparis J**, Herlin P, Laurinavicius A: Digital immunohistochemistry wizard: image analysis-assisted stereology tool to produce reference data set for calibration and quality control. *Diagnostic Pathology* 2014, 9(Suppl 1):S8
3. Laurinaviciene A, Baltrusaityte I, Meskauskas R, **Besusparis J**, Lesciute-Krilaviciene D, Raudeliunas D, Iqbal Y, Herlin P, Laurinavicius A: Digital immunohistochemistry platform for the staining variation monitoring based on integration of image and statistical analyses with laboratory information system. *Diagnostic Pathology* 2014, 9(Suppl 1):S10

4. SANTRAUKA LIETUVIŲ KALBA

4.1 Įvadas

Darbo aktualumas

Krūties vėžys yra labiausiai paplitęs piktybinis susirgimas tarp moterų visame pasaulyje: 2013 metais diagnozuota 1,8 mln. naujų krūties vėžio atvejų, dėl šios priežasties per minėtą laikotarpį mirė 464 000 moterų [1, 2]. Išgyvenamumo rodikliai varijuoja nuo 80 % Šiaurės Amerikoje, Japonijoje, Švedijoje iki 40 % besivystančiose šalyse [2]. Šių rodiklių gerinimas gali būti pasiektas taikant ankstyvos diagnostikos programas arba adjuvantinę chemoterapiją, kuri derinant su hormoniniu gydymu yra labai efektyvi, tačiau pasižymi potencialiai pavojingais šalutiniais reiškiniais ir yra brangi [6]. Dėl šios priežasties gydant hormonų receptoriams pozityvius krūties navikus yra itin svarbu nuspręsti, ar papildomai yra reikalinga adjuvantinė chemoterapija, į kurios skyrimo indikacijas yra įtraukti patikimi prognostiniai faktoriai aukštai rizikai identifikuoti [6]. Tradiciniai prognostiniai faktoriai, tokie kaip morfologiniai naviko bruožai ar metastazių limfmazgiuose statusas, yra nepakankami, todėl pastaraisiais metais buvo rasta daug naujų krūties vėžio prognostinių ir predikcinių žymenų, kurių dauguma yra susiję su naviko ląstelių ciklo reguliacija ir proliferacija [7-10].

Krūties navikų gydymo taktikos pasirinkimas priklauso nuo naviko potipio, kuris yra apibrėžiamas remiantis imunohistochemine estrogenų receptorių (ER), progesterono receptorių (PgR) raiška, žmogaus epitelio augimo faktoriaus 2 (HER2) statusu ir Ki67 proliferacinio indekso (Ki67 PI) vertėmis navikiniame audinyje. Luminaliniai A navikai yra gydomi vien tik endokrinine terapija, kuri yra ir luminalinio B potipio gydymo strategijos dalis. Papildomai chemoterapija yra taikoma daugumai pacienčių, sergančių luminaliniu B, HER2 teigiamu arba trigubai neigiamu krūties vėžiu [25]. Pagrindinė skirtis tarp luminalinių A ir luminalinių B navikų yra Ki67 proliferacinis aktyvumas, kurio ribinės vertės yra žemos ir įvairiose rekomendacijose varijuoja nuo 14 % iki 30 % [12, 24-26]. Norint taikyti minėtą klasifikaciją yra būtinas patikimas ir išsamus ER, PgR, HER2 ir Ki67 PI imunohistocheminių reakcijų vertinimas, kuris yra ypač svarbus priimant tinkamą klinikinį sprendimą ir parenkant gydymo strategiją.

Ki67 yra plačiai naudojamas imunohistocheminis žymuo, skirtas proliferuojančių ląstelių daliai tiriamajame audinyje nustatyti. Vizualus Ki67

proliferacinio indekso vertinimas yra pirmo pasirinkimo metodas kasdienėje praktikoje. Tačiau, literatūros duomenimis, vizualaus vertinimo atkartojamumas tarp skirtingų patologų yra žemas, ypač turint omeny ribines, kliniškai svarbiausias vertes nuo 10 % iki 20 % [118]. Klaidingas Ki67 proliferacinio indekso įvertinimas gali lemti netinkamai pasirinktą krūties navikų gydymo taktiką. Dėl šios priežasties vizualaus vertinimo rekomendacijose yra nurodoma vertinti skaičiuojant didelius kiekius (mažiausiai 500–1 000) naviko ląstelių branduolių [24]. Tačiau tokie skaičiai yra per aukšti rankiniam vertinimui ir sunkiai įgyvendinami rutiniame darbe. Daugelis patologų renkasi subjektyvų vizualų vertinimą arba taiko individualius metodus, pagreitinančius proliferacinio indekso įvertinimą. Įprastinė imunohistocheminių preparatų interpretacija yra pagrįsta žmogaus vizualių sugebėjimų identifikuoti audinių struktūras ir atlikti pusiau kiekybinius vertinimus. Šis procesas pasižymi žemu atkartojamumu ir gali reikšmingai paveikti galutinį Ki67 proliferacinio indekso vertinimą tiriant naviką [24]. Nors šis metodas yra pakankamas įprastiems diagnostiniams tikslams pasiekti, jis gali būti netinkamas priimant personalizuoto gydymo sprendimus.

Intranavikinis Ki67 žymens raiškos heterogeniškumas – labai būdingas krūties navikų bruožas [59-62], kuris nėra plačiai tyrinėtas ir potencialiai yra viena iš pagrindinių neefektyvios terapijos priežasčių [96]. Proliferacinio indekso vertinimas yra rekomenduojamas tuose naviko plotuose, kuriuose žymens raiška yra ryškiausia. Vis dėlto nėra vienos nuomonės, ar Ki67 PI turėtų būti apskaičiuojamas kaip Ki67 teigiamų naviko ląstelių procentas visame invazyvaus naviko plote, ar tik vadinamuosiuose „karštuosiuose taškuose“ [24, 58]. Pabrėžtina, kad dabartinėse gairėse „karštųjų taškų“ apibrėžimai ir imunohistocheminių žymenų intranavikinio heterogeniškumo nustatymas nėra standartizuoti. Ankstesni literatūros duomenys apie intranavikinio heterogeniškumo nustatymą ir apskaičiavimą histologiniuose preparatuose yra riboti. Ki67 proliferacinio indekso ribinės vertės kasmet žymiai varijuoja (14 %, 20 %, 20–29 %, ≤ 10 %, 20–30 %) [12, 24-26]. 2016 metais Amerikos klinikinės onkologijos draugija (angl. ASCO) išleido klinikinės praktikos gaires [18], kuriose yra nurodoma, kad dėl vidutinio įrodymų lygio Ki67 proliferacinis indeksas neturėtų būti taikomas kaip atrankos kriterijus skiriant adjuvantinę chemoterapiją. Kadangi vizualus vertinimas yra sudėtingesnis dėl minėtų faktorių, kurie sumažina Ki67 PI prognostinę vertę, naujų, išsamių ir patikimų vertinimo metodologijų paieška yra labai aktuali.

Klinikiniuose tyrimuose skaitmeninės vaizdo analizės metodų taikymas tampa vis dažnesne alternatyva vizualiam vertinimui. Literatūros duomenimis, automatizuotas Ki67 proliferacinio aktyvumo vertinimas pasižymi aukštu atkartojamumu ir potencialiai galėtų būti pritaikytas kasdienėje praktikoje [27]. Tačiau intranavikinis Ki67 žymens raiškos heterogeniškumo nustatymas nėra standartizuotas ir tai yra pagrindinė prasto atitikimo tarp vizualaus ir automatizuoto vertinimo priežastis [31]. Ankstesniuose skaitmeninės vaizdo analizės taikymo darbuose buvo gilinamasi į optimalų naviko ląstelių segmentavimą ir automatizuotą žymenų karštųjų taškų aptikimą [33-37] išvengiant vaizdo analizės algoritmų analitinės ir klinikinės validacijos. Dėl šios priežasties automatizuoto imunohistocheminių žymenų vertinimo ir skaitmeninės vaizdo analizės pritaikomumas kasdienėje praktikoje vis dar yra ribotas.

4.1.1 Darbo tikslas

Sukurti metodologiją, užtikrinančią krūties navikų Ki67 proliferacinio indekso vertinimo tikslumą, ir įvertinti šio indekso intranavikinio heterogeniškumo lygį audinyje, naudojant skaitmeninę vaizdo analizę.

4.1.2 Darbo uždaviniai

1. Sukurti metodologiją, užtikrinančią skaitmeninės vaizdo analizės algoritmų tikslumą Ki67 proliferacinio indekso vertinimui krūties navikų imunohistocheminiuose preparatuose.
2. Sukurti metodologiją išsamiam Ki67 proliferacinio indekso skaičiavimui, jo heterogeniškumo nustatymui ir karštųjų taškų aptikimui, atlikti analitinę naujų metodų validaciją pacienčių, sergančių krūties vėžiu, imtyje.
3. Nustatyti imunohistocheminių Ki67 krūties navikų preparatų vertinimo ir audinių mikrogardelių mėginių ėmimo parametrus, atsižvelgiant į intranavikinį Ki67 proliferacinio indekso heterogeniškumo lygį audinyje.
4. Įvertinti naujo Ki67 proliferacinio indekso skaičiavimo metodo prognostinę vertę pacienčių, sergančių krūties piktybiniais navikais, imtyje.

4.1.3 Ginamieji teiginiai

1. Remiantis palyginimu su pamatine Ki67 PI verte, teigtina, kad kalibruotais skaitmeninės vaizdo analizės metodais nustatyto Ki67 proliferacinio indekso tikslumas yra didesnis nei patologo vizualaus vertinimo atveju.
2. Krūties navikų Ki67 proliferacinio aktyvumo intranavikinis heterogeniškumas, nustatytas segmentuojant skaitmeninės vaizdo analizės duomenis į šešiakampes gardeles, yra nepriklausomas prognostinis krūties vėžiu sergančių pacienčių bendro išgyvenamumo rodiklis ir pasižymi stipresne prognostine verte nei vien proliferacinio aktyvumo lygis.

4.1.4 Tyrimo naujumas

1. *Analitinė skaitmeninės vaizdo analizės validacija.* Aprašoma imunohistocheminių preparatų skaitmeninės vaizdo analizės tikslumo gerinimo metodologija pagrįsta analitinės validacijos procedūromis, kurios atliktos remiantis pamatinėmis vertėmis, apskaičiuotomis taikant stereologijos principus. Ankstesnėse panašiose studijose analitinės validacijos procedūros arba nebuvo atliekamos, arba pamatinėmis vertėmis buvo pasirenkami subjektyvaus vertinimo rezultatai.
2. *Šešiakampių gardelių principas.* Sukurta nauja imunohistocheminių žymenų vertinimo ir erdvinų parametrų nustatymo metodologija, pagrįsta šešiakampių gardelių principu ir išbandyta Ki67 proliferacinio indekso matavimuose. Skenuotų preparatų segmentavimo į šešiakampes gardeles principas anksčiau nebuvo taikytas patologijoje. Metodika taip pat pritaikyta imunohistocheminių žymenų karštiesiems taškams aptikti.
3. *Heterogeniškumo parametrų nustatymas.* Darbe pristatomi nauji ir standartizuoti metodai, skirti intranavikiniam Ki67 PI heterogeniškumui kiekybiškai įvertinti ir erdvinės tekstūros parametrams apskaičiuoti. Šiame darbe pirmą kartą nustatyta, kad skaitmeninės vaizdo analizės metodais apskaičiuotų išvestinių heterogeniškumo parametrų prognostinė vertė yra didesnė nei įprastinių Ki67 PI matavimo metodologijų. Ši išvada gali potencialiai sustiprinti Ki67 PI prognostinę vertę.
4. *Ki67 vizualaus vertinimo ir audinių mikrogardelių eksperimentų kriterijai.* Šešiakampių gardelių principas buvo pritaikytas Ki67 PI vizualaus vertinimo bei audinių mikrogardelių eksperimentų ėminių

kriterijams nustatyti, atsižvelgiant į intranavikinį heterogeniškumą. Eksperimentai atlikti remiantis išsamiu statistiniu modeliavimu ir daugybiniais pakartojimais. Pabrėžtina, kad ankstesni panašūs tyrimai buvo paremti fiziniu audinių mėginių ėmimu, kuris buvo pagrindinis apribojimas patikimam statistiniam modeliavimui atlikti.

4.2 Metodai

Šiame skyrelyje trumpai aptariami tik pagrindiniai metodai, taikyti disertaciniame darbe. Detalesnė informacija yra pateikta ankstesnėse publikacijose (Etapai I–IV [28, 122–124]). Darbe naudotos trys skirtingos krūties piktybiniais navikais sergančių pacienčių imtys. Atskiros darbo dalys buvo patvirtintos Lietuvos bioetikos komiteto ir Notingamo mokslo tiriamųjų darbų etikos komiteto. Pacienčių sutikimas dalyvauti tyrime buvo gautas. Statistinė analizė atlikta naudojant SAS 9.3, *Microsoft Excel* (*Microsoft*, Redmondas, Vašingtono valstija, JAV) ir *OpenOffice Calc* (*Oracle*, Redvud Sitis, Kalifornijos valstija, JAV) programinės įrangos paketus. Statistinio reikšmingumo lygmuo buvo naudotas ties $p < 0,05$ reikšmėmis.

Pacienčių imtys

1. 164 pacienčių audinių mikrogardelių histologiniai preparatai, dažyti Ki67 imunohistocheminiu žymeniu. Pacientės gydytos 2007–2009 metais Nacionaliniame vėžio institute, Vilniuje. Tyrime naudota pacienčių klinikinė ir histologinio atsakymo informacija. (Imtis naudota 1-am uždaviniui įgyvendinti.)
2. 302 pacienčių audinių viso pjūvio histologiniai preparatai, dažyti Ki67 imunohistocheminiu žymeniu. Pacientės gydytos 2013–2014 metais Nacionaliniame vėžio institute, Vilniuje. Tyrime naudota pacienčių klinikinė ir histologinio atsakymo informacija. (Imtis naudota 2-am ir 3-iam uždaviniams įgyvendinti.)
3. 152 pacienčių audinių viso pjūvio histologiniai preparatai, dažyti Ki67 imunohistocheminiu žymeniu. Pacientės gydytos 1986–1998 metais Notingame (Jungtinė Karalystė). Tyrime naudota pacienčių klinikinė ir histologinio atsakymo informacija bei išgyvenamumo duomenys. (Imtis naudota 4-am uždaviniui įgyvendinti.)

Audinių paruošimas

Formaline fiksuoti ir parafinuoti krūties navikų audinio preparatai buvo panaudoti ruošiant 3 µm storio histologinius pjūvius tolesniam imunohistocheminiam dažymui. Imunohistocheminės reakcijos atliktos naudojant monokloninius Ki67 antikūnus (klonas MIB-1; DAKO, Glostrupas, DK). Detalūs imunohistocheminio dažymo aprašymai yra pateikti ankstesnėse publikacijose [28, 124]. Histologiniai preparatai buvo skenuoti *Aperio Scan-Scope XT Slide* skeneriu (*Aperio Technologies*, Vista, Kalifornijos valstija, JAV) taikant 20x padidinimą (0,5 µm rezoliucija).

Pirmajame etape audinių mikrogardelių konstrukcijai buvo panaudoti 1 mm skersmens krūties navikinio audinio stulpeliai, kurie buvo atsitiktinai pasirinkti patologo. Iš kiekvienos pacientės buvo panaudota po vieną TMA stulpelį. Audinių mikrogardelių konstrukcijos eksperimentas yra plačiau aprašytas [128]. Kituose šio darbo etapuose (II–IV) buvo naudojami viso pjūvio histologiniai preparatai.

Skaitmeninė vaizdo analizė

Skaitmeninė vaizdo analizė buvo atlikta naudojant dvi skirtingas programines įrangos platformas: I–III etapuose taikyti *Aperio Genie and Nuclear v9* algoritmai, IV etape skaitmenizuoti histologiniai preparatai analizuoti HALO™ audinių klasifikavimo modulių ir *CytoNuclear v1.4* algoritmu (*Indica Labs*, Naujosios Meksikos valstija, JAV). Naudojant programines įrangos algoritmus buvo atliktas automatizuotas naviko ploto atpažinimas ir kiekybinė naviko ląstelių branduolių profilių analizė. Audinio klasifikavimo moduliai buvo kalibruoti invazyvių navikinių kompleksų aptikimui, eliminuojant stromos, uždegimo ir kitus audinio plotus, turinčius artefaktų. Naviko ląstelių analizė atlikta automatizuotu būdu kiekybiškai klasifikuojant ląstelių branduolių profilius į turinčius neigiamą ir teigiamą Ki67 imunohistocheminę raišką. I etape atlikti keli vaizdo analizės algoritmų kalibracijos žingsniai: 1. Baziniai nustatymai, 2. Nustatymai pagal subjektyvų analizės rezultatų kokybės vertinimą, 3. Modifikuojant branduolių atpažinimo algoritmą pagal statistinį paklaidos vertinimą, gautą lyginant antro algoritmo modifikacijos žingsnio duomenis su pamatinėmis Ki67 vertėmis.

I etapas. *Skaitmeninė vaizdo analizė: kalibracija, kiekybinė analizė ir validacija*

Pamatinių Ki67 proliferacinio indekso verčių generavimą rankiniu būdu atliko 3 patologai žymėdami ir skaičiuodami teigiamus ir neigiamus naviko ląstelių profilius taikant stereologinį tinklelį. Skaitmeninė skenuotų preparatų analizė atlikta taikant *Aperio Genie, Nuclear v9* algoritmus. Skaitmeninės vaizdo analizės algoritmų kalibracija atlikta trimis etapais (1. Baziniai nustatymai, 2. Intuityvi kalibracija pagal vizualų vertinimą, 3. Programinės įrangos nustatymų koregavimas pagal statistinės analizės rezultatus, gautus antrojo kalibracijos etapo analizės rezultatus lyginant su pamatine verte). Automatizuoto Ki67 proliferacinio indekso (trijų kalibracijos etapų) vertės ir vizualaus (5 patologų) vertinimo vidurkis bei atskiros individualių vertintojų Ki67 PI vertės palygintos su pamatine Ki67 PI verte. Palyginimui naudoti koreliacijos, tiesinės regresijos ir vienfaktorinės dispersinės analizės ANOVA modeliai. Duomenų palyginimui poromis taikytas Duncano daugialypio lyginimo testas.

II etapas. *Erdvinė ir multiparametrinė analizė: heterogeniškumo matavimai, šešiakampių gardelių taikymas*

Skenuoti Ki67 imunohistocheminiai krūties navikų preparatai segmentuoti į lygius regionus taikant šešiakampių gardelių principą. Atlikta skaitmeninė skenuotų preparatų analizė *Aperio Genie, Nuclear v9* algoritmais ir Ki67 proliferacinis indeksas apskaičiuotas kiekvienoje šešiakampėje gardelėje bei visame invazyvaus naviko plote. Naudojant šiuos duomenis ir „bendrosios masės matricą“ (angl. *Co-occurrence matrix*) apskaičiuoti proliferacinio indekso heterogeniškumo parametrai. Atliktas automatizuoto proliferacinio indekso karštųjų taškų aptikimas, remiantis viršutiniu segmentavimo rezultatų kvintiliu, kuris buvo pavadintas Pareto karštuoju tašku. Automatizuoto Ki67 vertinimo rezultatai palyginti su klinikiniais duomenimis ir vizualiu patologo vertinimu. Atlikta faktorinė duomenų analizė.

III etapas. Šešiakampių gardelių principo pritaikymas Ki67 PI vizualaus vertinimo ir audinių mikrogardelių eksperimentų ėminių kriterijų nustatymui

Šiame etape buvo atliktas skenuotų Ki67 imunohistocheminio dažymo krūties navikų preparatų segmentavimas į šešiakampes gardeles ir pritaikytas virtualiam audinių mikrogardelių konstrukcijos eksperimentui. Remiantis skaitmeninės vaizdo analizės rezultatais Ki67 PI reikšmės įvertintos skirtinguose šešiakampių gardelių „virtualių audinių mikrogardelių“ rinkiniuose ($N = 1, \dots, 15$) ir palygintos su Ki67 PI vertėmis visame konkreto atvejo plote. Procedūra pakartota po 50 000 kartų kiekviename „virtualių audinių mikrogardelių“ rinkinyje ir kiekvienu tiriamuoju atveju ($N = 297$). Norint palyginti rezultatus, vizualaus vertinimo kriterijų bei audinių mikrogardelių konstrukcijos parametrų nustatymui apskaičiuotos klaidos koeficiento reikšmės. Tokiu būdu įvertinama, kokia yra audinio kiekio įtaka galutiniam Ki67 PI vertinimui. Tiriamieji atvejai ir eksperimentai suskirstyti į atskiras grupes pagal žymens raiškos heterogeniškumo lygį audinyje, naudojant II etape aprašytą metodologiją.

IV etapas. Klinikinė validacija: automatizuotų Ki67 PI parametrų palyginimas su pacienčių bendro išgyvenamumo duomenimis

Automatizuoto krūties navikų Ki67 proliferacinio indekso ir jo heterogeniškumo nustatymo metodologija ištestuota krūties navikų imtyje (Notingamas, Jungtinė Karalystė), turinčioje ilgamečio stebėjimo rezultatus. Skaitmeninė skenuotų preparatų analizė atlikta taikant HALO™ Classifier Module/CytoNuclear v1.4 algoritmus, kurie buvo kalibruoti patikimam naviko struktūrų atpažinimui ir kiekybiniam naviko ląstelių branduolių vertinimui. Šešiakampių gardelių principas buvo pritaikytas erdvinės naviko tekstūros matavimui, heterogeniškumo parametrų nustatymui ir automatizuotam Ki67 PI karštųjų taškų aptikimui. Ki67 PI vertinimo rezultatai palyginti su pacienčių išgyvenamumo duomenimis taikant Coxo multiparametrinės regresijos analizę ir Kaplano–Meierio išgyvenamumo kreives (nuolatiniai kintamieji buvo dichotomizuoti naudojant viešos prieigos įrankį „Cutoff Finder“ [131]). Atlikta faktorinė duomenų analizė. Kiekybiniams duomenims palyginti naudota multifaktorinės variacijos

analizė ANOVA (palyginimui poromis – Bonferoni testas).

4.3 Rezultatai

I etapas. *Skaitmeninė vaizdo analizė: kalibracija, kiekybinė analizė ir validacija*

Vidutinė apskaičiuota pamatinė Ki67 proliferacinio indekso vertė – $40,2 \pm 25,3$ %, lyginant su vizualaus 5 patologų vertinimo mediana ir kalibruoto skaitmeninės vaizdo analizės algoritmo rezultatu, atitinkamai $37,2 \pm 27,4$ % ir $36,5 \pm 20,2$ %. Vizualaus 5 patologų vertinimo mediana ir kalibruoto skaitmeninės vaizdo analizės algoritmo rezultato koreliacijos koeficientų reikšmės, lyginant su pamatinėmis Ki67 PI vertėmis, atitinkamai 0,930 ir 0,949 ($p < 0,0001$). Tiesinės regresijos modelių rezultatai (pamatinę Ki67 PI vertę laikant nepriklausomu kintamuoju): vizualaus 5 patologų vertinimo mediana – $R^2 = 0,86$ ($p < 0,0001$), kalibruoto skaitmeninės vaizdo analizės algoritmo rezultatas – $R^2 = 0,9$ ($p < 0,0001$). Pacienčių dichotomizavimo į kliniškai svarbias grupes ties ribinėmis vertėmis > 10 %, > 15 %, > 20 % tikslumas (klaidingai suklasifikuotų atvejų dalis), lyginant su pamatinėmis Ki67 PI vertėmis: a) naudojant vizualaus 5 patologų vertinimo medianą – 11 %, 14 %, 18 %; b) naudojant kalibruoto skaitmeninės vaizdo analizės algoritmo Ki67 PI rezultatą – 7 %, 9 %, 7 %.

II etapas. *Erdvinė ir multiparametrinė analizė: heterogeniškumo matavimai, šešiakampių gardelių taikymas*

Gauta ideali koreliacija ($r = 0,9967$, $p < 0,0001$) tarp Ki67 PI, automatizuotai apskaičiuoto visame histologiniame pjūvyje, ir Ki67 PI medianos šešiakampėse gardelėse. Aukštesnis Ki67 PI asocijuotas su aukštesniu naviko diferenciacijos laipsniu ($p < 0,0001$); aukštesnės Ki67 PI entropijos reikšmės yra būdingos aukštesnio diferenciacijos laipsnio navikams ($p < 0,0001$). Faktorinės analizės metu išskirti 4 svarbūs faktoriai: 1. Heterogeniškumo faktorius, 2. Proliferacijos faktorius, 3. Bimodališkumo faktorius, 4. Ląstelingumo faktorius. Ki67 PI Pareto karštųjų taškų plote, kuris persidengia su manualiniu žymėjimu, buvo statistiškai reikšmingai didesnis nei likusiame naviko plote (T testas: vidutiniai skirtumai 5,8 %, 5,8 % ir 6,7 %; $p < 0,0003$, $p < 0,0002$ ir $p < 0,0001$). Trijų vertintojų Ki67 PI vidurkis karštuosiuose taškuose reikšmingai nesiskyrė ($p = 0,0675$) nuo Pareto karštojo taško įvertinimo (90-ojo procentilio).

III etapas. *Šešiakampių gardelių principo pritaikymas Ki67 PI vizualaus vertinimo ir audinių mikrogardelių eksperimentų ėminių kriterijų nustatymui*

Optimalūs audinių ėminių ir Ki67 PI vertinimo parametrai varijuoja priklausomai nuo audinio heterogeniškumo lygio. Norint pasiekti 10 % klaidos koeficiento vertę yra reikalinga naudoti 8 TMA stulpelius nežinomo heterogeniškumo navikams; homogeniškiems navikams – 5 TMA stulpelius; heterogeniškiems navikams – 11 TMA stulpelių. Norint pasiekti 10 % klaidos koeficiento vertę reikalinga įvertinti 4 000 naviko ląstelių branduolių vizualiam Ki67 PI nustatymui, kai intranavikinio heterogeniškumo lygis yra nežinomas. Vertinant atvejus, kai Ki67 PI yra žemas (< 20 %), yra reikalingas didesnis ėminių / ląstelių skaičius, norint pasiekti tą pačią paklaidos tikimybę, kaip ir aukšto Ki67 PI navikų atveju.

IV etapas. *Klinikinė validacija: automatizuotų Ki67 PI parametrų palyginimas su pacienčių bendro išgyvenamumo duomenimis*

Vidutinis pacienčių stebėjimo laikas – $143,4 \pm 71,4$ mėnesio, 79 iš 152 pacienčių mirė. Trijuose Coxo regresijos modeliuose vienas iš automatizuotu būdu apskaičiuotų heterogeniškumo indikatorių (bimodališkumo parametras) buvo identifikuotas kaip nepriklausomas blogos pacienčių prognozės veiksnys. Nė vienas iš Ki67 proliferacinio aktyvumo parametrų (apskaičiuotų tiek rankiniu, tiek automatizuotu būdu) nebuvo identifikuotas kaip nepriklausomas veiksnys, prognozuojantis bendro pacienčių išgyvenamumo rodiklius. Faktorinės analizės metu išskirti 5 faktoriai, iš kurių svarbiausi yra 1. Proliferacijos faktorius, 2. Heterogeniškumo faktorius ir 4. Bimodališkumo faktorius. Multifaktorinės variacijos analizės (ANOVA) duomenimis, aukštesnis 1 faktorius asocijuotas su aukštesniu diferenciacijos laipsniu ($p < 0,0001$). Trigubai neigiami navikai pasižymėjo aukštesnėmis 1 faktoriaus reikšmėmis nei hormonų receptoriams pozityvi grupė.

4.4 Rezultatų aptarimas

Doktorantūros studijų metu atliktų tyrimų rezultatai atskleidžia keletą aspektų, kurie yra svarbūs siekiant pagerinti krūties navikų gydymo taktikai pasirinkti naudojamo Ki67 proliferacinio indekso prognostinę vertę. Nustatyta, kad automatizuoto Ki67 proliferacinio indekso vertinimo tikslumas yra didesnis nei patologų vizualaus vertinimo atveju. Pabrėžtina, kad vaizdo analizės algoritmai buvo kalibruojami kaip pamatinės vertės naudojant Ki67 PI įverčius, gautus atlikus nešališkus skaičiavimus pagal stereologijos taisykles. Antra, proliferacinio naviko aktyvumo erdvinis heterogeniškumas gali būti pamatuotas ir vizualizuotas taikant šešiakampių gardelių principą, kurį taip pat įmanoma pritaikyti biožymenų raiškos karštųjų taškų aptikimui ir jų kiekybiniam įvertinimui. Be to, šešiakampių gardelių metodas buvo pritaikytas ir audinių mikrogardelių eksperimentų reikalavimų optimizavimui, atsižvelgiant į krūties vėžio Ki67 PI heterogeniškumą. Krūties navikų proliferacinio aktyvumo erdvinio heterogeniškumo rodikliai, išmatuoti taikant skaitmeninę vaizdo analizę ir šešiakampes gardeles, gali būti nepriklausomas kintamasis, prognozuojantis pacienčių bendro išgyvenamumo rodiklius ir pranokstantis vien tik proliferacinio aktyvumo įverčius.

Vaizdo analizės metodų taikymas skaitmenizuotų imunohistocheminių preparatų vertinimui yra sparčiai besivystantis ir tobulėjantis procesas patologijos tyrimų srityje. Didžiausias skaitmeninės vaizdo analizės privalumas – dideli audinio informacijos kiekiai, kurie gali būti lengvai išgaunami analizuojant aukštos raiškos skenuotus histologinius preparatus. Automatizuotos analizės pajėgumas yra gerokai didesnis nei rankinio preparatų vertinimo. Pavyzdžiui, šiame tyrime buvo išanalizuota 36 mln. ląstelių bendrame 15 000 mm² naviko plote bei 13 mln. ląstelių 6 000 mm² naviko plote, atitinkamai eksperimentuose, aprašytuose II ir IV tyrimuose. Antrojo tyrimo metu viename histologiniame preparate vidutiniškai įvertinta 121 000 navikinių ląstelių profilių. Remiantis šiuo požiūriu vienintelio indikatorius (t. y. Ki67 PI) apskaičiavimas galėtų būti laikomas nepakankamu duomenų panaudojimu. Skaitmeninei vaizdų analizei yra būdingas ne vien tik didesnis tikslumas, atkartojamumas ir pajėgumas, tačiau automatizuotas imunohistocheminių preparatų vertinimas pasižymi ir aukštesne prognostine verte lyginant su vizualiu vertinimu [132]. Remiantis navikų mikroskopinės architektūros požymiais, apskaičiuotais skaitmeninės vaizdų analizės metodais, įmanoma stratifikuoti krūties navikus į turinčius

skirtingą metastatinį potencialą [133]. G. Stålhammaras su bendraautoriais aprašo eksperimentą, kurio metu skaitmenizuotas Ki67 PI vertinimas pranoko rankinį mitozijų skaičiavimą ir Ki67 PI nustatymą įprastiniais metodais bei turėjo papildomos prognostinės informacijos [32]. Be to, automatizuoto Ki67 PI vertinimo rezultatai reikšmingai koreliuoja su molekulinėmis modeliais, pasižyminčių itin aukštu ligos recidyvo prognozavimo jautrumu ir specifiskumu, rezultatais [15].

Nepaisant sparčiai tobulėjančios programinės įrangos, skirtos skaitmeninei histologinių preparatų analizei, minėti algoritmai yra beveik netaikomi kasdienėje patologijos praktikoje. Validacijos procedūrų netikslumai yra didžiausias ankstesnių eksperimentų, tyrinėjusių vaizdo analizės algoritmų pritaikomumą, trūkumas. Ankstesniuose darbuose algoritmų validacija arba nebūdavo atliekama, arba pamatinėmis vertėmis buvo pasirenkami subjektyvaus vertinimo rezultatai, kurie negali būti taikomi kaip „tiesos kriterijus“. Šio darbo pirmajame etape buvo atlikta išsami skaitmeninės vaizdo analizės algoritmų kalibracija ir sukurtas vadinamasis „aukso standartas“, paremtas stereologijos taisyklėmis ir rankiniu naviko ląstelių profilių skaičiavimu. Šiame etape nustatyta, kad 5 patologų atlikto vizualaus Ki67 PI vertinimo vidurkis nepasiekė tikslumo lygio, gauto taikant kalibruotą skaitmeninės vaizdo analizės algoritmą. Taip pat buvo apskaičiuotas vizualaus ir automatizuoto vertinimo tikslumas skirstant pacientus į kliniškai svarbias grupes pagal dažniausiai naudojamas Ki67 PI ribines vertes (10 %, 15 % ir 20 %). Eksperimento metu nustatyta, kad paklaidos tikimybė yra du kartus didesnė, jei yra naudojamas patologo vertinimas. Šis rezultatas buvo iš dalies netikėtas, tačiau tai yra papildomi įrodymai, kad vizualus pusiau kiekybinis imunohistocheminių Ki67 preparatų vertinimas negali būti naudojamas kaip patikimas matavimas, kuriuo remiantis priimami klinikiniai sprendimai arba nustatomos pamatinės vertės, atliekant skaitmeninės vaizdo analizės algoritmų validavimą. Šiame eksperimente taip pat nustatyta, kad skaitmeninės vaizdo analizės kokybė gerėjo po kiekvieno kalibracijos žingsnio, o geriausias rezultatas buvo pasiektas statistiškai išanalizavus algoritmų paklaidas. Remiantis šiais rezultatais galima teigti, kad pirmas žingsnis skaitmeninės vaizdo analizės eksperimentuose turėtų būti kruopšti algoritmų validacija, naudojant pamatines vertes, apskaičiuotas nešališkais vertinimo metodais.

Doktorantūros studijų metu buvo sukurtas naujas ir pažangus automatizuotas metodas, skirtas erdvinei biologinių žymenų tekstūrai ir heterogeniškumui histologiniuose preparatuose nustatyti ir įvertinti.

Metodika yra pagrįsta sisteminiu skenuotų preparatų skaitmeninės vaizdo analizės rezultatų padalijimu į lygias šešiakampes gardeles. Kaip ir tikėtasi, taikant šį metodą buvo nustatyti naviko erdvinės tekstūros parametrai ir jiems suteiktos skaitinės reikšmės, kurios vėliau kartu su pasiskirstymo duomenimis gali būti panaudotos atvejų palyginimui ir navikų stratifikacijai į skirtingo heterogeniškumo grupes. Ankstesniuose panašiuose tyrimuose navikų erdvinio heterogeniškumo nustatymas buvo atliktas lyginant mėginius, paimtus iš to paties naviko skirtingų regionų. Dažniausiai buvo naudojami fiziniai audinių mikrogardelių (TMA) mėginiai, kuriuose lyginti tiriamojo biologinio žymens raiškos skirtumai tarp TMA stulpelių, paimtų iš to paties naviko [98, 134, 135]. Nors, šių eksperimentų duomenimis, tam tikri navikinio audinio variacijos skirtumai gali būti identifikuoti, tačiau heterogeniška ER, PgR, HER2 raiška buvo aptinkama nedažnai (< 10 % atvejų) [134]. Pagrindinis TMA metodo trūkumas nustatant biologinių žymenų heterogeniškumą yra tiesioginė rezultatų priklausomybė nuo audinių ėminių skaičiaus ir viso eksperimento struktūros. Dažnai per mažas TMA ėminių kiekis neatspindi viso navikinio audinio įvairovės ir gerokai apriboja heterogeniškumo tyrimus tokio tipo eksperimentuose.

Skaitmeninės vaizdo analizės metodų taikymas intranavikiniam heterogeniškumo lygiui įvertinti nėra naujas. S. J. Pottasas kartu su bendraautorais pritaikė ekologijos mokslų srityje naudojamus variacijos rodiklius krūties navikų imunohistocheminės HER2 raiškos heterogeniškumo nustatymui histologiniuose preparatuose, iš anksto pasirinktuose regionuose [96]. Tyrimo metu, naudojant skaitmeninės vaizdo analizės rezultatus, buvo sukurtas virtualus žemėlapis, parodantis HER2 ekspresijos lygį skirtingose naviko vietose. Tyrimo išvadose autoriai pabrėžia, kad, norint tiksliai nustatyti erdvinę variaciją, analizėje naudotų regionų skaičius gali būti nepakankamas. Tokie tyrimai turėtų būti atlikti naudojant viso pjūvio histologinius preparatus, kuriuose būtų analizuojamas visas naviko plotas, o ne smulkūs jo fragmentai. Pabrėžtina, kad rengiant disertaciją sukurta šešiakampių gardelių metodika yra paremta sisteminiu skaitmeninės vaizdo analizės rezultatų segmentavimu ir naviko erdvinės tekstūros parametrų nustatymu visame atlikto pjūvio histologiniame preparate. Suklasifikavus šios analizės metu gautus rezultatus (Ki67 PI atskiruose šešiakampiuose) į skirtingus intervalus, gautus duomenis galima būtų panaudoti naujų, kompleksinių ir potencialiai biologiškai reikšmingesnių indikatorių kūrimui nei vien Ki67 proliferacinio indekso vertės.

Šiame darbe yra aprašomas naujas ir pranašus metodas, skirtas biožymenų imunohistocheminės raiškos karštiesiems taškams aptikti. Jis paremtas stabilios naviko dalies (90-ojo procentilio) vertinimu ir pavadintas Pareto karštuoju tašku. Literatūros duomenimis, skaitmeninės vaizdo analizės algoritmai yra plačiai taikomi biožymenų raiškos karštiesiems taškams aptikti audiniuose [15, 22, 33, 34, 36, 37, 120]. Nepaisant to, dalyje ankstesnių eksperimentų vis dar yra taikomi metodai, paremti rankiniu skaičiavimu [35], daugybiniu vizualiu vertinimu [69] arba audinių mikrogardelių konstrukcijos principų taikymu [134, 135]. Kitose mokslo srityse (ekologijoje, geografijoje) erdvinės paviršiaus tekstūros matavimams yra plačiai taikomi statistiniai metodai, paremti tiriamojo objekto segmentavimu į smulkesnius lygius fragmentus. Medicinos srities mokslo tiriamuosiuose darbuose šiam tikslui dažniausiai yra naudojamos stačiakampio formos gardelės [15, 22, 37]. Pavyzdžiui, E. Gudlaugssonas su bendraautorais [22] aprašo metodą, skirtą Ki67 PI karštiesiems taškams aptikti krūties navikuose, selektyviai pasirenkant keturkampio formos laukus su vizualiai didžiausia Ki67 teigiamų ląstelių dalimi ir juose atliekant skaitmeninę vaizdo analizę. Tyrimo metu nustatyta, kad tokiu būdu apskaičiuotos Ki67 PI vertės pasižymi aukštu atkartojamumu ir yra prognostiškai reikšmingos. Vis dėlto toks metodas tiesiogiai priklauso nuo rankiniu būdu pasirinktų analizės plotų, o rezultatai gauti neatsižvelgus į galimą skirtingą ląstelių tankį navike. Šie eksperimentai iš dalies yra panašūs į doktorantūros studijų metu atliktus darbus, nes histologiniai preparatai buvo virtualiai segmentuojami į lygius smulkesnius regionus. Tačiau šešiakampių gardelių principas, kuris anksčiau dar niekada nebuvo pritaikytas histologinių preparatų segmentavimui, pasižymi beveik idealiu viso naviko ploto padengimu ir yra nepriklausomas nuo neregulios bei kompleksiškos karštojo taško formos ir dydžio. Naudojant stačiakampes gardeles, audinyje lieka nepadengtų regionų (ypač audinio kraštuose), o šešiakampių gardelių principas užtikrina, kad yra vienoda tikimybė pasirinkti bet kurį audinio regioną. Pabrėžtina, kad taikant Pareto principą yra vertinama stabili naviko ląstelių dalis (20 % aukščiausią Ki67 PI vertę turinčių šešiakampių gardelių), tačiau šios taisyklės gali būti lengvai modifikuojamos, priklausomai nuo eksperimento struktūros ir tikslų. Apibendrinant galima teigti, kad skaitmeninės vaizdo analizės taikymas karštajam taškui aptikti gali būti naudingas dėl kelių priežasčių: a) tai gali būti panaudota standartizuoto karštojo taško apibrėžimui kurti, aprašant jų formą, dydį, „temperatūrą“ ir kitas savybes, b) objektyviam ir aukštu

atkartojamumu pasižyminčiam karštojo taško aptikimui, c) taikant šią metodiką yra įmanoma išryškinti naviko vietas (šešiakampių gardelių grupes), kuriose Ki67 imunohistocheminė raiška yra ryškiausia, tokiu būdu palengvinant kasdienį patologo darbą identifikuojant sunkiai randamus karštuosius taškus tiriamajame audinyje, d) pacienčių atrankos individualizuotai terapijai gerinti.

Darbo metu nustatyta, kad skirtingų vertintojų sutarimas aptinkant karštuosius Ki67 PI taškus audinyje yra gana prastas. II ir IV etapuose karštieji taškai buvo apibrėžti ir įvertinti trijų patologų. Lyginant tarp skirtingų vertintojų, sutarimas aptinkant bent vieną karštąjį tašką audinyje varijavo priklausomai nuo vertintojų porų (kapa koeficientai nuo 0,2 iki 0,5 II etapo metu ($n = 50$) ir nuo 0,55 iki 0,85 IV etapo metu ($n = 152$)). Vertinant karštojo taško dydžio ir formos persidengimą tarp skirtingų patologų, erdvinė variacija buvo gauta labai ryški. Kaip ir tikėtasi, šie duomenys atitinka nurodytus literatūroje [19, 22, 118, 136] ir akcentuoja biologinių žymenų raiškos karštųjų taškų apibrėžimo standartizavimo svarbą. Šiame darbe patologų aptiktų karštųjų taškų įverčiai buvo panašūs į esančius šešiakampėse gardelėse, turinčiose aukštesnę Ki67 PI vertę, ir buvo lyginti su Pareto karštojo taško įverčio mediana.

Atlikus faktoriinę šešiakampių gardelių segmentavimo duomenų (išvestinių heterogeniškumo parametrų) analizę buvo išryškinti 4 pagrindiniai faktoriai: entropijos, proliferacijos, bimodališkumo ir ląstelingumo. Šie faktoriai vėliau buvo panaudoti klasterinėje analizėje, kurioje išryškėjo heterogeniškų navikų potipiai su dominuojančia entropija, bimodališkumu ir skirtingomis proliferacinio indekso vertėmis. Tačiau siekiant tiksliau stratifikuoti navikus į homogeniškus ir heterogeniškus yra būtini įrodymais pagrįsti apibrėžimai, kurie idealiu atveju būtų paremti pacienčių išgyvenamumo duomenų analize.

Šešiakampių gardelių segmentavimo principas buvo pritaikytas Ki67 imunohistocheminio vertinimo standartizacijos pagerinimui ir reikalavimų TMA eksperimentams nustatymui, atliekant daugybinius virtualių TMA gardelių modeliavimo eksperimentus. Rezultatai suskirstyti atsižvelgiant į audinio heterogeniškumo lygį. Šio eksperimento metu buvo gauti nauji ir dalinai prieštaringi literatūros duomenims rezultatai. Nustatyta, kad optimalūs preparatų vertinimo parametrai priklauso nuo tiriamojo žymens raiškos heterogeniškumo lygio audinyje. Šio tyrimo metu nustatyta, kad patikimam TMA eksperimentų modeliavimui yra reikalingi didesni audinių ėminių kiekiai, nei yra aprašoma literatūroje. 5–6 audinių mikrogardelių

stulpeliai yra būtini, kai žymens raiška – homogeniška, 11 stulpelių – heterogeniškiems navikams ir 8 stulpeliai, kai heterogeniškumo lygis yra nežinomas. Ankstesniuose panašiuose darbuose rekomenduojamas TMA stulpelių skaičius varijuoja nuo 1 iki 4. Tokie ryškūs nesutapimai tarp literatūros ir šio eksperimento duomenų gali būti paaiškinami labai mažu pakartojimų dažniu ankstesniuose eksperimentuose, kurie dažniausiai yra paremti vienu atsitiktinio modelio sukūrimu. Tai ir yra pagrindinė priežastis, apsunkinanti išsamų statistinį modeliavimą. Šio eksperimento metu virtualus audinių mikrogardelių modeliavimo eksperimentas buvo atliktas po 50 000 kartų kiekvienu atveju, kiekviename skirtingame virtualių TMA stulpelių rinkinyje. Pirmą kartą nustatyti vizualaus krūties navikų Ki67 imunohistocheminio žymens vertinimo parametrai, remiantis išsamaus statistinio modeliavimo rezultatais. Nustatyta, kad norint pasiekti mažą paklaidos tikimybę vizualiai vertinant Ki67 PI krūties piktybiniuose navikuose, kai heterogeniškumo lygis yra nežinomas, tikslinga įvertinti apie 4 000 naviko ląstelių branduolių. Taip pat yra būtina skaičiuoti daugiau naviko ląstelių branduolių, kai Ki67 PI yra žemas (< 20 %). Pagal dabartines rekomendacijas minimalus ląstelių skaičius, būtinas vizualiam vertinimui, yra 500–1 000 branduolių [24]. Pabrėžtina, kad šio tyrimo metu nustatytas minimalus ląstelių kiekis viršija esamas rekomendacijas 2 kartus ir turėtų būti taikomas, kai intranavikinio heterogeniškumo lygis yra nežinomas. Tokie aukšti ląstelių vertinimo reikalavimai, žinoma, yra sunkiai įmanomi kasdienėje praktikoje visais atvejais, tačiau potencialiai galėtų būti taikomi, kai yra reikalingas labai tikslus proliferacinio indekso vertinimas norint priimti kliniskus sprendimus (10–30 % Ki67 PI intervale).

Pirmą kartą nustatyta, kad krūties navikų Ki67 PI erdvinio heterogeniškumo rodikliai, apskaičiuoti taikant šešiakampių gardelių principą, pasižymi didesne prognostine verte nei konservatyvus navikų proliferacinio aktyvumo vertinimas. Ketvirtajame doktorantūros etape automatizuota heterogeniškumo matavimo metodologija buvo pritaikyta krūties navikų imčiai, turinčiai ilgamečio pacienčių stebėjimo ir išgyvenamumo duomenis. Šiame eksperimente buvo ištirta įvairių Ki67 proliferacinio indekso, audinio erdvinės tekstūros, bimodališkumo parametru ir vizualaus vertinimo duomenų prognostinė vertė. Pabrėžtina, kad daugumai kiekybinių Ki67 automatizuoto ir vizualaus vertinimo parametru buvo nustatytos statistiškai patikimos ribinės vertės stratifikuojant pacientes į kliniškai svarbias grupes. Tačiau tik bimodališkumo parametrai (Ashmano D koeficientas) buvo identifikuoti kaip nepriklausomi bendro pacienčių

išgyvenamumo rodikliai hormonų receptoriams ir HER2 pozityviuose navikų subtipuose bei pranoko prognostinę įprastinių Ki67 PI parametru vertę. Šis rezultatas buvo netikėtas, tačiau gali turėti praktinę reikšmę. Dėl ryškios krūties navikų biologinės įvairovės ir audinio heterogeniškumo variacijos yra problemiška apibrėžti Ki67 PI ribines vertes ir užtikrinti tikslų kiekybinį Ki67 imunohistocheminių preparatų vertinimą. Skaitmeninės vaizdo analizės metodais apskaičiuoti išvestiniai Ki67 PI bimodališkumo parametrai, tikėtina, yra mažiau jautrūs šiems veiksniams. Šio tyrimo rezultatai įrodo, kad intranavikinio proliferacinio aktyvumo variacijos rodikliai gali būti svarbesnis ligos agresyvumo rodiklis, lemiantis individualią prognozę, nei vien Ki67 PI vertinimas.

Apibendrinant pasakytina, kad šios disertacijos rezultatai patvirtina, jog aukštas Ki67 imunohistocheminio vertinimo tikslumas yra sunkiai pasiekiamas įprastiniais vizualaus vertinimo metodais. Pabrėžtina informacinių technologijų svarba Ki67 PI vertinimo efektyvumui gerinti, apskaičiuojant standartizuotus heterogeniškumo parametrus ir automatizuotu būdu aptinkant karštuosius žymens raiškos taškus viso pjūvio histologiniuose preparatuose. Išvestiniai audinio erdvinės tekstūros parametrai, apskaičiuoti skaitmeninės vaizdo analizės metodais, gali būti potencialiai naudojami pasirenkant krūties navikų gydymo taktiką ateityje.

4.4 Darbo tęstinumas

Skaitmeninės vaizdo analizės įdiegimas į klinikinę praktiką

Reali skaitmeninės vaizdo analizės įrankių klinikinė nauda gali būti pasiekta įdiegus juos į patologijos laboratorijų informacines sistemas ir naudojant kaip sprendimų palaikymo įrankius patologų kasdienėje praktikoje. Tai ir būtų pagrindinis šioje disertacijoje aprašomų eksperimentų tikslas. Nepaisant to, kad vaizdo analizės algoritmų integracija nebuvo šio disertacinio darbo uždavinys, Valstybiniame patologijos centre (Vilnius, Lietuva) jau yra atlikti bandomieji eksperimentai, kurių metu validuotas vaizdo analizės algoritmas buvo įdiegtas į informacinę laboratorijos sistemą ir pritaikytas rutininiam Ki67 PI vertinimui krūties navikuose.

Heterogeniškumo apibrėžimai ir klinikinė validacija

Taikant šiame darbe aprašytą šešiakampių gardelių principą yra įmanoma

modeliuoti keletą skirtingų biožymenų karštųjų taškų apibrėžimų, kurie šiame darbe nebuvo nagrinėti. Geriausias būdas validuoti šiuos skaitmeninės vaizdo analizės metodus pagrįstus apibrėžimus yra palyginti juos su pacienčių išgyvenamumo duomenimis ir nustatyti jų prognostinę reikšmę. Biožymenų raiškos erdvinio heterogeniškumo parametrus galima apskaičiuoti remiantis skaitmeninės analizės rezultatais, tačiau navikų stratifikavimas į homogeniškus ir heterogeniškus taip pat turi būti pagrįstas klinikiniais eksperimentais. Kituose šio darbo etapuose yra numatytas išvestinių vaizdo analizės parametrų prognostinės vertės įvertinimas plečiant tiriamųjų pacienčių grupes ir renkant išgyvenamumo duomenis.

Šešiakampių gardelių principo pritaikomumas

Šešiakampių gardelių metodas buvo išbandytas krūties navikų Ki67 imunohistocheminiuose preparatuose, tačiau jį galima lengvai pritaikyti ir kitiems tikslams ar audiniams. Automatizuotas Ki67 PI vertinimas galėtų palengvinti kruopštų proliferacinio indekso vertinimą virškinamojo trakto neuroendokrininiuose navikuose, kai ribinės vertės labai žemos (1 %, 3 % ir 20 %), o vizualus vertinimas yra labai apsunkintas. Skaitmeninė vaizdo analizė ir karštųjų taškų aptikimas galėtų palengvinti šią užduotį ir užtikrinti geresnį pacienčių skirstymą į kliniškai reikšmingas grupes.

Kita sritis, kurioje skaitmeninės vaizdo analizės algoritmai galėtų būti sėkmingai pritaikyti, yra estrogenų, progesterono ir HER2 receptorių vertinimas krūties navikuose. Literatūros duomenimis, automatizuotas heterogeniškumo nustatymas minėtuose biologiniuose žymenyse buvo labai retai taikomas. Šiame darbe aprašyti metodai galėtų būti lengvai pritaikyti minėtai užduočiai atlikti, prieš tai įgyvendinus išsamią naujai sukurtų skaitmeninės vaizdo analizės algoritmų validaciją.

4.5 Išvados

1. Sukurta metodologija, užtikrinanti ir pagerinanti skaitmeninės vaizdo analizės algoritmų Ki67 proliferacinio indekso vertinimo tikslumą. Krūties navikų Ki67 proliferacinio aktyvumo indeksas, nustatytas skaitmeninės vaizdo analizės metodais, pranoksta vizualų patologo vertinimą, lyginant su pamatine verte, apskaičiuota taikant stereologinius metodus. Skirstant krūties karcinomomis sergančias pacientes į terapiniu požiūriu svarbias grupes (Ki67 PI > 10, 15 ir 20 %) paklaidos tikimybė

yra 2 kartus mažesnė remiantis automatizuoto Ki67 proliferacinio indekso matavimo vertėmis nei taikant vizualų 5 patologų Ki67 proliferacinio indekso vertinimo vidurkį. Vaizdo analizės algoritmai gali būti toliau sėkmingai tobulinami, matuojant jų paklaidą ir analizuojant galimas sisteminės neatitikimų priežastis.

2. Sukurta tikslaus Ki67 PI vertinimo metodologija, paremta šešiakampių gardelių segmentavimo principu. Šiuo metodu yra įmanoma pamatuoti ir vizualizuoti Ki67 proliferacinio aktyvumo indekso erdvinę tekstūrą navike bei apskaičiuoti žymens raiškos intranavikinio heterogeniškumo parametrus, skaitmeninės vaizdo analizės metodais analizuotuose Ki67 imunohistocheminiuose preparatuose. Išvestiniai heterogeniškumo parametrai gali būti pritaikyti krūties navikų dichotomizavimui į turinčius homogenišką ir heterogenišką Ki67 imunohistocheminę raišką. Šis principas buvo panaudotas automatizuotam Ki67 PI karštųjų taškų aptikimui, kuris yra kiekybiškai apskaičiuojamas remiantis viršutiniu segmentavimo rezultatų kvintiliu ir pavadintas Pareto karštuoju tašku. Šešiakampių gardelių metodas taip pat pasižymi lengvu pritaikomumu kitose patologijos srityse tiriant įvairius imunohistocheminius žymenis ar audinius, siekiant nustatyti jų raiškos heterogeniškumą arba atlikti kokybės kontrolės procedūras.
3. Optimalūs audinių mikrogardelių mėginių ėmimo ir / arba preparatų vertinimo parametrai priklauso nuo tiriamojo žymens raiškos heterogeniškumo lygio audinyje. Mokslo tiriamosiose studijose nurodoma, kad Ki67 imunohistocheminio žymens tyrimuose, atliekamuose su krūties piktybinių navikų audiniais, 5–6 audinių mikrogardelių stulpeliai yra būtini, kai žymens raiška – homogeniška, 11 stulpelių – heterogeniškiems navikams ir 8 stulpeliai, kai heterogeniškumo lygis yra nežinomas. Norint pasiekti mažą paklaidos tikimybę vizualiai vertinant Ki67 proliferacinį aktyvumą krūties piktybiniuose navikuose, kai heterogeniškumo lygis yra nežinomas, tikslinga įvertinti apie 4 000 naviko ląstelių branduolių. Vertinant atvejus, kai Ki67 PI yra žemas (< 20 %), yra reikalingas didesnis ėminių / ląstelių skaičius, norint pasiekti tą pačią paklaidos tikimybę, kaip ir aukšto Ki67 PI navikuose.
4. Erdvinio navikų proliferacinio aktyvumo heterogeniškumo parametrai (ypač bimodališkumo statusas), apskaičiuoti taikant skaitmeninę vaizdo analizę ir šešiakampių gardelių segmentavimo principą, gali būti naudojami kaip nepriklausomi prognostiniai faktoriai, apibūdinantys

krūties piktybiniais navikais sergančių pacienčių bendro išgyvenamumo rodiklius ir pranokstantys prognostinę Ki67 proliferacinio aktyvumo reikšmę.

4.6 Publikacijų sąrašas

Publikacijos, tiesiogiai susijusios su disertacijos tikslu ir uždaviniais:

1. Laurinavicius, A, Plancoulaine, B, Laurinaviciene, A, Herlin, P, Meskauskas, R, Baltrusaityte, I, **Besusparis, J**, Dasevicius, D, Elie, N, Iqbal, Y, Bor, C., *A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue*. Breast Cancer Research, 2014. **16**(2): p. R35.
2. Plancoulaine, B, Laurinaviciene, A, Herlin, P, **Besusparis, J**, Meskauskas, R, Baltrusaityte, I, Iqbal, Y, Laurinavicius, A., *A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data*. Virchows Archiv, 2015. **467**(6): p. 711–722.
3. Laurinavicius, A, Plancoulaine, B, Rasmusson, A, **Besusparis, J**, Augulis, R, Meskauskas, R, Herlin, P, Laurinaviciene, A, Abdelhadi Muftah, A. A, Miligy, I, Aleskandarany, M, Rakha, E. A, Green, A. R, Ellis, I. O., *Bimodality of intratumor Ki67 expression is an independent prognostic factor of overall survival in patients with invasive breast carcinoma*. Virchows Archiv, 2016. **468**(4): p. 493–502.
4. **Besusparis, J**, Plancoulaine, B, Rasmusson, A, Augulis, R, Green, A. R, Ellis, I. O, Laurinaviciene, A, Herlin, P, Laurinavicius, A., *Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer*. Diagnostic Pathology, 2016. **11**(1): p. 82.

Publikacijos, netiesiogiai susijusios su disertacijos uždaviniais:

1. Laurinavicius, A, **Besusparis, J**, Didziapetryte, J, Radziuviene, G, Meskauskas, R, Laurinaviciene, A., *Digital immunohistochemistry: new horizons and practical solutions in breast cancer pathology*. Diagnostic Pathology, 2013. **8**(Suppl 1): S15.
2. Plancoulaine, B, Meskauskas, R, Baltrusaityte, I, **Besusparis, J**, Herlin, P, Laurinavicius, A., *Digital immunohistochemistry wizard: image analysis-assisted stereology tool to produce reference data set for calibration and quality control*. Diagnostic Pathology, 2014. **9**(Suppl 1): S8.
3. Laurinaviciene, A, Baltrusaityte, I, Meskauskas, R, **Besusparis, J**, Lesciute-Krilaviciene, D, Raudeliunas, D, Iqbal, Y, Herlin, P, Laurinavicius, A., *Digital immunohistochemistry platform for the staining variation monitoring based on integration of image and statistical analyses with laboratory information system*. Diagnostic Pathology, 2014. **9**(Suppl 1): S10.

REFERENCES

1. Global Burden of Disease Cancer, C., et al., *The Global Burden of Cancer 2013*. JAMA Oncol, 2015. **1**(4): p. 505-27.
2. Coleman, M.P., et al., *Cancer survival in five continents: a worldwide population-based study (CONCORD)*. Lancet Oncol, 2008. **9**(8): p. 730-56.
3. Kerlikowske, K., et al., *Efficacy of screening mammography. A meta-analysis*. JAMA, 1995. **273**(2): p. 149-54.
4. Cowan, W.K., et al., *A study of interval breast cancer within the NHS breast screening programme*. J Clin Pathol, 2000. **53**(2): p. 140-6.
5. Groenendijk, R.P., et al., *Screen-detected breast cancers have a lower mitotic activity index*. Br J Cancer, 2000. **82**(2): p. 381-4.
6. van Diest, P.J., E. van der Wall, and J.P. Baak, *Prognostic value of proliferation in invasive breast cancer: a review*. J Clin Pathol, 2004. **57**(7): p. 675-81.
7. van Diest, P.J., et al., *Cyclin D1 expression in invasive breast cancer. Correlations and prognostic value*. Am J Pathol, 1997. **150**(2): p. 705-11.
8. de Jong, J.S., et al., *Expression of growth factors, growth inhibiting factors, and their receptors in invasive breast cancer. I: An inventory in search of autocrine and paracrine loops*. J Pathol, 1998. **184**(1): p. 44-52.
9. de Jong, J.S., et al., *Concerted overexpression of the genes encoding p21 and cyclin D1 is associated with growth inhibition and differentiation in various carcinomas*. Mol Pathol, 1999. **52**(2): p. 78-83.
10. de Jong, J.S., et al., *Expression of growth factors, growth-inhibiting factors, and their receptors in invasive breast cancer. II: Correlations with proliferation and angiogenesis*. J Pathol, 1998. **184**(1): p. 53-7.

11. de Jong, J.S., P.J. van Diest, and J.P. Baak, *Hot spot microvessel density and the mitotic activity index are strong additional prognostic indicators in invasive breast cancer*. *Histopathology*, 2000. **36**(4): p. 306-12.
12. Untch, M., et al., *Primary Therapy of Patients with Early Breast Cancer: Evidence, Controversies, Consensus: Opinions of German Specialists to the 14th St. Gallen International Breast Cancer Conference 2015 (Vienna 2015)*. *Geburtshilfe Frauenheilkd*, 2015. **75**(6): p. 556-565.
13. Paik, S., et al., *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer*. *N Engl J Med*, 2004. **351**(27): p. 2817-26.
14. Klein, M.E., et al., *Prediction of the Oncotype DX recurrence score: use of pathology-generated equations derived by linear regression analysis*. *Mod Pathol*, 2013. **26**(5): p. 658-64.
15. Thakur, S.S., et al., *The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer*. *PLoS One*, 2018. **13**(1): p. e0188983.
16. Kim, H.S., et al., *Optimizing the Use of Gene Expression Profiling in Early-Stage Breast Cancer*. *J Clin Oncol*, 2016. **34**(36): p. 4390-4397.
17. Harowicz, M.R., et al., *Algorithms for prediction of the Oncotype DX recurrence score using clinicopathologic data: a review and comparison using an independent dataset*. *Breast Cancer Res Treat*, 2017. **162**(1): p. 1-10.
18. Harris, L.N., et al., *Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline*. *J Clin Oncol*, 2016. **34**(10): p. 1134-50.
19. Polley, M.Y., et al., *An international Ki67 reproducibility study*. *J Natl Cancer Inst*, 2013. **105**(24): p. 1897-906.
20. Voros, A., et al., *An intra- and interobserver reproducibility analysis of the Ki-67 proliferation marker assessment on core biopsies of breast cancer patients and its potential clinical implications*. *Pathobiology*, 2013. **80**(3): p. 111-8.
21. Voros, A., et al., *Different methods of pretreatment Ki-67 labeling index evaluation in core biopsies of breast cancer patients treated with*

- neoadjuvant chemotherapy and their relation to response to therapy*. *Pathol Oncol Res*, 2015. **21**(1): p. 147-55.
22. Gudlaugsson, E., et al., *Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer*. *Histopathology*, 2012. **61**(6): p. 1134-44.
23. Varga, Z., et al., *How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists*. *PLoS One*, 2012. **7**(5): p. e37379.
24. Dowsett, M., et al., *Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group*. *J Natl Cancer Inst*, 2011. **103**(22): p. 1656-64.
25. Goldhirsch, A., et al., *Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011*. *Ann Oncol*, 2011. **22**(8): p. 1736-47.
26. Untch, M., et al., *13th st. Gallen international breast cancer conference 2013: primary therapy of early breast cancer evidence, controversies, consensus - opinion of a german team of experts (zurich 2013)*. *Breast Care (Basel)*, 2013. **8**(3): p. 221-9.
27. Madabhushi, A., et al., *Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data*. *Comput Med Imaging Graph*, 2011. **35**(7-8): p. 506-14.
28. Laurinavicius, A., et al., *A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue*. *Breast Cancer Res*, 2014. **16**(2): p. R35.
29. Kushnarev, V.A., E.S. Artemyeva, and A.G. Kudaybergenova, *[Comparison of digital and visual methods for Ki-67 assessment in invasive breast carcinomas]*. *Ark Patol*, 2018. **80**(2): p. 38-42.
30. Maeda, I., et al., *Comparison between Ki67 labeling index determined using image analysis software with virtual slide system and that determined visually in breast cancer*. *Breast Cancer*, 2016. **23**(5): p. 745-51.
31. Zhong, F., et al., *A Comparison of Visual Assessment and Automated Digital Image Analysis of Ki67 Labeling Index in Breast Cancer*. *PLoS One*, 2016. **11**(2): p. e0150505.

32. Stalhammar, G., et al., *Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer*. *Histopathology*, 2017.
33. Lopez, X.M., et al., *Clustering methods applied in the detection of Ki67 hot-spots in whole tumor slide images: an efficient way to characterize heterogeneous tissue-based biomarkers*. *Cytometry A*, 2012. **81**(9): p. 765-75.
34. Saha, M., et al., *An Advanced Deep Learning Approach for Ki-67 Stained Hotspot Detection and Proliferation Rate Scoring for Prognostic Evaluation of Breast Cancer*. *Sci Rep*, 2017. **7**(1): p. 3213.
35. Romero, Q., et al., *A novel model for Ki67 assessment in breast cancer*. *Diagn Pathol*, 2014. **9**: p. 118.
36. Wessel Lindberg, A.S., et al., *Quantitative tumor heterogeneity assessment on a nuclear population basis*. *Cytometry A*, 2017. **91**(6): p. 574-584.
37. Lu, H., et al., *Automated Selection of Hotspots (ASH): enhanced automated segmentation and adaptive step finding for Ki67 hotspot detection in adrenal cortical cancer*. *Diagn Pathol*, 2014. **9**: p. 216.
38. Rakha, E.A. and A.R. Green, *Molecular classification of breast cancer: what the pathologist needs to know*. *Pathology*, 2017. **49**(2): p. 111-119.
39. Perou, C.M., et al., *Molecular portraits of human breast tumours*. *Nature*, 2000. **406**(6797): p. 747-52.
40. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*. *Proc Natl Acad Sci U S A*, 2001. **98**(19): p. 10869-74.
41. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets*. *Proc Natl Acad Sci U S A*, 2003. **100**(14): p. 8418-23.
42. Cheang, M.C., et al., *Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer*. *J Natl Cancer Inst*, 2009. **101**(10): p. 736-50.
43. Wolff, A.C., et al., *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, 2007. **131**(1): p. 18-43.

44. Hammond, M.E., et al., *American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version).* Arch Pathol Lab Med, 2010. **134**(7): p. e48-72.

45. Nielsen, T.O., et al., *Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma.* Clin Cancer Res, 2004. **10**(16): p. 5367-74.

46. Weigel, M.T. and M. Dowsett, *Current and emerging biomarkers in breast cancer: prognosis and prediction.* Endocr Relat Cancer, 2010. **17**(4): p. R245-62.

47. Ross, J.S., et al., *Breast cancer biomarkers and molecular medicine.* Expert Rev Mol Diagn, 2003. **3**(5): p. 573-85.

48. Weidner, N., D.H. Moore, 2nd, and R. Vartanian, *Correlation of Ki-67 antigen expression with mitotic figure index and tumor grade in breast carcinomas using the novel "paraffin"-reactive MIB1 antibody.* Hum Pathol, 1994. **25**(4): p. 337-42.

49. Spyrtatos, F., et al., *Correlation between MIB-1 and other proliferation markers: clinical implications of the MIB-1 cutoff value.* Cancer, 2002. **94**(8): p. 2151-9.

50. Yerushalmi, R., et al., *Ki67 in breast cancer: prognostic and predictive potential.* Lancet Oncol, 2010. **11**(2): p. 174-83.

51. Viale, G., *Pathological work up of the primary tumor: getting the proper information out of it.* Breast, 2011. **20 Suppl 3**: p. S82-6.

52. Thor, A.D., et al., *Comparison of mitotic index, in vitro bromodeoxyuridine labeling, and MIB-1 assays to quantitate proliferation in breast cancer.* J Clin Oncol, 1999. **17**(2): p. 470-7.

53. Gerdes, J., et al., *Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation.* Int J Cancer, 1983. **31**(1): p. 13-20.

54. Lopez, F., et al., *Modalities of synthesis of Ki67 antigen during the stimulation of lymphocytes.* Cytometry, 1991. **12**(1): p. 42-9.

55. Gerdes, J., et al., *Cell cycle analysis of a cell proliferation-associated human nuclear antigen defined by the monoclonal antibody Ki-67*. J Immunol, 1984. **133**(4): p. 1710-5.
56. Cattoretti, G., et al., *Monoclonal antibodies against recombinant parts of the Ki-67 antigen (MIB 1 and MIB 3) detect proliferating cells in microwave-processed formalin-fixed paraffin sections*. J Pathol, 1992. **168**(4): p. 357-63.
57. Stuart-Harris, R., et al., *Proliferation markers and survival in early breast cancer: a systematic review and meta-analysis of 85 studies in 32,825 patients*. Breast, 2008. **17**(4): p. 323-34.
58. Jang, M.H., et al., *A comparison of Ki-67 counting methods in luminal Breast Cancer: The Average Method vs. the Hot Spot Method*. PLoS One, 2017. **12**(2): p. e0172031.
59. Almendro, V. and G. Fuster, *Heterogeneity of breast cancer: etiology and clinical relevance*. Clin Transl Oncol, 2011. **13**(11): p. 767-73.
60. Martelotto, L.G., et al., *Breast cancer intra-tumor heterogeneity*. Breast Cancer Res, 2014. **16**(3): p. 210.
61. Marusyk, A., V. Almendro, and K. Polyak, *Intra-tumour heterogeneity: a looking glass for cancer?* Nat Rev Cancer, 2012. **12**(5): p. 323-34.
62. Polyak, K., *Heterogeneity in breast cancer*. J Clin Invest, 2011. **121**(10): p. 3786-8.
63. Veta, M., et al., *Mitosis Counting in Breast Cancer: Object-Level Interobserver Agreement and Comparison to an Automatic Method*. PLoS One, 2016. **11**(8): p. e0161286.
64. Malon, C., et al., *Mitotic figure recognition: agreement among pathologists and computerized detector*. Anal Cell Pathol (Amst), 2012. **35**(2): p. 97-100.
65. Roux, L., et al., *Mitosis detection in breast cancer histological images An ICPR 2012 contest*. J Pathol Inform, 2013. **4**: p. 8.
66. Koelzer, V.H., et al., *Digital image analysis improves precision of PD-L1 scoring in cutaneous melanoma*. Histopathology, 2018.

67. Dubinski, W., et al., *Assessment of the prognostic significance of endoglin (CD105) in clear cell renal cell carcinoma using automated image analysis*. Hum Pathol, 2012. **43**(7): p. 1037-43.
68. Holten-Rossing, H., et al., *Optimizing HER2 assessment in breast cancer: application of automated image analysis*. Breast Cancer Res Treat, 2015. **152**(2): p. 367-75.
69. Lindstrom, L.S., et al., *Intratumor Heterogeneity of the Estrogen Receptor and the Long-term Risk of Fatal Breast Cancer*. J Natl Cancer Inst, 2018.
70. Gandara-Cortes, M., et al., *Breast cancer subtype discrimination using standardized 4-IHC and digital image analysis*. Virchows Arch, 2018. **472**(2): p. 195-203.
71. Tang, L.H., et al., *Objective quantification of the Ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: a comparison of digital image analysis with manual methods*. Am J Surg Pathol, 2012. **36**(12): p. 1761-70.
72. Charalampoudis, P., et al., *Thyroid hormone receptor alpha (TRa) tissue expression in ductal invasive breast cancer: A study combining quantitative immunohistochemistry with digital slide image analysis*. Eur J Surg Oncol, 2017. **43**(8): p. 1428-1432.
73. Daunoravicius, D., et al., *Quantification of myocardial fibrosis by digital image analysis and interactive stereology*. Diagn Pathol, 2014. **9**: p. 114.
74. Masugi, Y., et al., *Quantitative assessment of liver fibrosis reveals a nonlinear association with fibrosis stage in nonalcoholic fatty liver disease*. Hepatol Commun, 2018. **2**(1): p. 58-68.
75. Bihari, C., et al., *Quantitative fibrosis estimation by image analysis predicts development of decompensation, composite events and defines event-free survival in chronic hepatitis B patients*. Hum Pathol, 2016. **55**: p. 63-71.
76. Zhou, Y., et al., *An Inexpensive Digital Image Analysis Technique for Liver Fibrosis Quantification in Chronic Hepatitis B Patients*. Ann Hepatol, 2017. **16**(6): p. 881-887.

77. McIntire, P.J., et al., *Hot Spot and Whole-Tumor Enumeration of CD8(+) Tumor-Infiltrating Lymphocytes Utilizing Digital Image Analysis Is Prognostic in Triple-Negative Breast Cancer*. Clin Breast Cancer, 2018.
78. Abas, F.S., et al., *Computer-assisted quantification of CD3+ T cells in follicular lymphoma*. Cytometry A, 2017. **91**(6): p. 609-621.
79. Schaadt, N.S., et al., *Image analysis of immune cell patterns in the human mammary gland during the menstrual cycle refines lymphocytic lobulitis*. Breast Cancer Res Treat, 2017. **164**(2): p. 305-315.
80. Gaudio, F., et al., *Computer-driven quantitative image analysis in the assessment of tumor cell and T cell features in diffuse large B cell lymphomas*. Ann Hematol, 2018. **97**(4): p. 663-668.
81. Steele, K.E., et al., *Measuring multiple parameters of CD8+ tumor-infiltrating lymphocytes in human cancers by image analysis*. J Immunother Cancer, 2018. **6**(1): p. 20.
82. Eriksen, A.C., et al., *Computer-assisted stereology and automated image analysis for quantification of tumor infiltrating lymphocytes in colon cancer*. Diagn Pathol, 2017. **12**(1): p. 65.
83. Valkonen, M., et al., *Metastasis detection from whole slide images using local features and random forests*. Cytometry A, 2017. **91**(6): p. 555-565.
84. Ehteshami Bejnordi, B., et al., *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer*. JAMA, 2017. **318**(22): p. 2199-2210.
85. Radziuviene, G., et al., *Automated Image Analysis of HER2 Fluorescence In Situ Hybridization to Refine Definitions of Genetic Heterogeneity in Breast Cancer Tissue*. Biomed Res Int, 2017. **2017**: p. 2321916.
86. Aeffner, F., et al., *The Gold Standard Paradox in Digital Image Analysis: Manual Versus Automated Scoring as Ground Truth*. Arch Pathol Lab Med, 2017. **141**(9): p. 1267-1275.
87. Marusyk, A. and K. Polyak, *Tumor heterogeneity: causes and consequences*. Biochim Biophys Acta, 2010. **1805**(1): p. 105-17.
88. Dick, J.E., *Stem cell concepts renew cancer research*. Blood, 2008. **112**(13): p. 4793-807.

89. Turashvili, G. and E. Brogi, *Tumor Heterogeneity in Breast Cancer*. Front Med (Lausanne), 2017. **4**: p. 227.
90. Esparza-Lopez, J., et al., *Breast Cancer Intra-Tumor Heterogeneity: One Tumor, Different Entities*. Rev Invest Clin, 2017. **69**(2): p. 66-76.
91. Seol, H., et al., *Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its clinicopathological significance*. Mod Pathol, 2012. **25**(7): p. 938-48.
92. Almendro, V., et al., *Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity*. Cell Rep, 2014. **6**(3): p. 514-27.
93. Park, S.Y., et al., *Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype*. J Clin Invest, 2010. **120**(2): p. 636-44.
94. Wang, Y., et al., *Clonal evolution in breast cancer revealed by single nucleus genome sequencing*. Nature, 2014. **512**(7513): p. 155-60.
95. Navin, N., et al., *Inferring tumor progression from genomic heterogeneity*. Genome Res, 2010. **20**(1): p. 68-80.
96. Potts, S.J., et al., *Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue*. Lab Invest, 2012. **92**(9): p. 1342-57.
97. Weinhouse, S., *Discussion of Doctor Greenstein's paper*. Cancer Res, 1956. **16**(7): p. 654-7.
98. Nassar, A., et al., *Intratumoral heterogeneity of immunohistochemical marker expression in breast carcinoma: a tissue microarray-based study*. Appl Immunohistochem Mol Morphol, 2010. **18**(5): p. 433-41.
99. Murthy, S.S., et al., *Assessment of HER2/Neu status by fluorescence in situ hybridization in immunohistochemistry-equivocal cases of invasive ductal carcinoma and aberrant signal patterns: a study at a tertiary cancer center*. Indian J Pathol Microbiol, 2011. **54**(3): p. 532-8.
100. Hanna, W.M., et al., *HER2 in situ hybridization in breast cancer: clinical implications of polysomy 17 and genetic heterogeneity*. Mod Pathol, 2014. **27**(1): p. 4-18.

101. Chang, M.C., et al., '*Genetic heterogeneity*' in *HER2/neu testing by fluorescence in situ hybridization: a study of 2,522 cases*. *Mod Pathol*, 2012. **25**(5): p. 683-8.
102. Bartlett, A.I., et al., *Heterogeneous HER2 gene amplification: impact on patient outcome and a clinically relevant definition*. *Am J Clin Pathol*, 2011. **136**(2): p. 266-74.
103. Ohlschlegel, C., et al., *HER2 genetic heterogeneity in breast carcinoma*. *J Clin Pathol*, 2011. **64**(12): p. 1112-6.
104. Rakha, E.A., et al., *Updated UK Recommendations for HER2 assessment in breast cancer*. *J Clin Pathol*, 2015. **68**(2): p. 93-9.
105. Cottu, P.H., et al., *Intratumoral heterogeneity of HER2/neu expression and its consequences for the management of advanced breast cancer*. *Ann Oncol*, 2008. **19**(3): p. 595-7.
106. Slamon, D.J., et al., *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene*. *Science*, 1987. **235**(4785): p. 177-82.
107. Vance, G.H., et al., *Genetic heterogeneity in HER2 testing in breast cancer: panel summary and guidelines*. *Arch Pathol Lab Med*, 2009. **133**(4): p. 611-2.
108. Bartlett, J.M., et al., *HER2 testing in the UK: recommendations for breast and gastric in-situ hybridisation methods*. *J Clin Pathol*, 2011. **64**(8): p. 649-53.
109. Bernasconi, B., et al., *Genetic heterogeneity in HER2 testing may influence therapy eligibility*. *Breast Cancer Res Treat*, 2012. **133**(1): p. 161-8.
110. Hsu, C.Y., et al., *Proposal of modification for the definition of genetic heterogeneity in HER2 testing in breast cancer*. *Arch Pathol Lab Med*, 2010. **134**(2): p. 162; author reply 163.
111. Romain, S., et al., *Biological heterogeneity of ER-positive breast cancers in the post-menopausal population*. *Int J Cancer*, 1994. **59**(1): p. 17-9.
112. Chhieng, D.C., et al., *Intratumor heterogeneity of biomarker expression in breast carcinomas*. *Biotech Histochem*, 2004. **79**(1): p. 25-36.

113. Davis, B.W., et al., *Receptor heterogeneity of human breast cancer as measured by multiple intratumoral assays of estrogen and progesterone receptor*. Eur J Cancer Clin Oncol, 1984. **20**(3): p. 375-82.
114. Beca, F. and K. Polyak, *Intratumor Heterogeneity in Breast Cancer*. Adv Exp Med Biol, 2016. **882**: p. 169-89.
115. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing*. N Engl J Med, 2012. **366**(10): p. 883-892.
116. Colleoni, M., et al., *Annual Hazard Rates of Recurrence for Breast Cancer During 24 Years of Follow-Up: Results From the International Breast Cancer Study Group Trials I to V*. J Clin Oncol, 2016. **34**(9): p. 927-35.
117. Sheppard, V.B., et al., *Frailty and adherence to adjuvant hormonal therapy in older women with breast cancer: CALGB protocol 36990I*. J Clin Oncol, 2014. **32**(22): p. 2318-27.
118. Shui, R., et al., *An interobserver reproducibility analysis of Ki67 visual assessment in breast cancer*. PLoS One, 2015. **10**(5): p. e0125131.
119. Yates, L.R., *Intratumoral heterogeneity and subclonal diversification of early breast cancer*. Breast, 2017. **34 Suppl 1**: p. S36-S42.
120. Khan Niazi, M.K., et al., *Perceptual clustering for automatic hotspot detection from Ki-67-stained neuroendocrine tumour images*. J Microsc, 2014. **256**(3): p. 213-25.
121. Heindl, A., et al., *Relevance of Spatial Heterogeneity of Immune Infiltration for Predicting Risk of Recurrence After Endocrine Therapy of ER+ Breast Cancer*. J Natl Cancer Inst, 2018. **110**(2).
122. Plancoulaine, B., et al., *A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data*. Virchows Arch, 2015.
123. Besusparis, J., et al., *Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer*. Diagn Pathol, 2016. **11**(1): p. 82.
124. Laurinavicius, A., et al., *Bimodality of intratumor Ki67 expression is an independent prognostic factor of overall survival in patients with invasive breast carcinoma*. Virchows Arch, 2016. **468**(4): p. 493-502.

125. Rakha, E.A., et al., *Prognostic significance of Nottingham histologic grade in invasive breast carcinoma*. J Clin Oncol, 2008. **26**(19): p. 3153-8.
126. Abd El-Rehim, D.M., et al., *High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses*. Int J Cancer, 2005. **116**(3): p. 340-50.
127. Aleskandarany, M.A., et al., *Growth fraction as a predictor of response to chemotherapy in node-negative breast cancer*. Int J Cancer, 2010. **126**(7): p. 1761-9.
128. Laurinavicius, A., et al., *Immunohistochemistry profiles of breast ductal carcinoma: factor analysis of digital image analysis data*. Diagn Pathol, 2012. **7**: p. 27.
129. Bland, J.M. and D.G. Altman, *Measuring agreement in method comparison studies*. Stat Methods Med Res, 1999. **8**(2): p. 135-60.
130. Giavarina, D., *Understanding Bland Altman analysis*. Biochem Med (Zagreb), 2015. **25**(2): p. 141-51.
131. Budczies, J., et al., *Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization*. PLoS One, 2012. **7**(12): p. e51862.
132. Loughrey, M.B., et al., *Validation of the systematic scoring of immunohistochemically-stained tumour tissue microarrays using QuPath digital image analysis*. Histopathology, 2018.
133. Roxanis, I., et al., *The significance of tumour microarchitectural features in breast cancer prognosis: a digital image analysis*. Breast Cancer Res, 2018. **20**(1): p. 11.
134. Allott, E.H., et al., *Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification*. Breast Cancer Res, 2016. **18**(1): p. 68.
135. Supernat, A., et al., *Tumor heterogeneity at protein level as an independent prognostic factor in endometrial cancer*. Transl Oncol, 2014. **7**(5): p. 613-9.
136. Chung, Y.R., et al., *Interobserver Variability of Ki-67 Measurement in Breast Cancer*. J Pathol Transl Med, 2016. **50**(2): p. 129-37.

137. Ilyas, M., et al., *Guidelines and considerations for conducting experiments using tissue microarrays*. *Histopathology*, 2013. **62**(6): p. 827-39.
138. Goethals, L., et al., *A new approach to the validation of tissue microarrays*. *J Pathol*, 2006. **208**(5): p. 607-14.
139. Torhorst, J., et al., *Tissue microarrays for rapid linking of molecular changes to clinical endpoints*. *Am J Pathol*, 2001. **159**(6): p. 2249-56.
140. Zhang, D., et al., *Reliability of tissue microarrays in detecting protein expression and gene amplification in breast cancer*. *Mod Pathol*, 2003. **16**(1): p. 79-84.
141. Alkushi, A., *Validation of tissue microarray biomarker expression of breast carcinomas in Saudi women*. *Hematol Oncol Stem Cell Ther*, 2009. **2**(3): p. 394-8.
142. Mucci, N.R., et al., *Neuroendocrine expression in metastatic prostate cancer: evaluation of high throughput tissue microarrays to detect heterogeneous protein expression*. *Hum Pathol*, 2000. **31**(4): p. 406-14.
143. Schmidt, L.H., et al., *Tissue microarrays are reliable tools for the clinicopathological characterization of lung cancer tissue*. *Anticancer Res*, 2009. **29**(1): p. 201-9.
144. Quintayo, M.A., et al., *Virtual tissue microarrays: a novel and viable approach to optimizing tissue microarrays for biomarker research applied to ductal carcinoma in situ*. *Histopathology*, 2014. **65**(1): p. 2-8.
145. Pedersen, M.B., et al., *Digital pathology for the validation of tissue microarrays in peripheral T-cell lymphomas*. *Appl Immunohistochem Mol Morphol*, 2014. **22**(8): p. 577-84.
146. Haroske, G., et al., *Cellular sociology of proliferating tumor cells in invasive ductal breast cancer*. *Anal Quant Cytol Histol*, 1996. **18**(3): p. 191-8.

ACKNOWLEDGEMENTS

This thesis would not have been possible without all the support and contribution received from many kind people. I would like to gratefully thank everyone for being together, in particular:

My beloved wife and family, for your enormous support throughout writing this thesis, faith, love and all my life in general.

Prof. Arvydas Laurinavičius, my supervisor, who initiated this project, for your motivation, inspiration and patience. The door of your office was always open whenever I ran into a trouble or had a question about my research or writing.

Allan Rasmuson, for all your help and huge assistance in statistical modeling, preparation of the manuscripts and for all your patience keeping me on track.

Benoit Plancoulaine and **Paulette Herlin**, the originators of hexagonal tiling approach, for all your intelligence, experience in mathematics, fresh ideas, positivity and assistance in statistical analysis of the data.

Renaldas Augulis, for all your assistance, sleepless nights and weekends we were working together before deadlines.

My sincere thanks also goes to **prof. Ian O Ellis** and **prof. Emad Rakha**, who kindly provided me an opportunity to join their team as a guest, and who gave access to the research facilities.

Finally, I must express my very profound gratitude to my **Parents** for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

PAPERS

Papers I and III were reprinted with a confirmed permission from *Springer Nature*. Papers II and IV are available online.

Paper I

A METHODOLOGY TO ENSURE AND IMPROVE ACCURACY OF KI67 LABELLING INDEX ESTIMATION BY AUTOMATED DIGITAL IMAGE ANALYSIS IN BREAST CANCER TISSUE

Laurinavicius, A, Plancoulaine, B, Laurinaviciene, A, Herlin, P,
Meskauskas, R, Baltrusaityte, I, Besusparis, J, Dasevicius, D, Elie, N,
Iqbal, Y, Bor, C.

Breast Cancer Research, 2014. 16(2): p. R35.

RESEARCH ARTICLE

Open Access

A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue

Arvydas Laurinavicius^{1,2*}, Benoit Plancoulaine³, Aida Laurinaviciene^{1,2}, Paulette Herlin^{1,3}, Raimundas Meskauskas^{1,2}, Indra Baltrušaityte^{1,2}, Justinas Besusparis^{1,2}, Darius Dasevicius^{1,2}, Nicolas Elie³, Yasir Iqbal¹, Catherine Bor^{3,4} and Ian O Ellis^{1,5}

Abstract

Introduction: Immunohistochemical Ki67 labelling index (Ki67 LI) reflects proliferative activity and is a potential prognostic/predictive marker of breast cancer. However, its clinical utility is hindered by the lack of standardized measurement methodologies. Besides tissue heterogeneity aspects, the key element of methodology remains accurate estimation of Ki67-stained/counterstained tumour cell profiles. We aimed to develop a methodology to ensure and improve accuracy of the digital image analysis (DIA) approach.

Methods: Tissue microarrays (one 1-mm spot per patient, n = 164) from invasive ductal breast carcinoma were stained for Ki67 and scanned. Criterion standard (Ki67-Count) was obtained by counting positive and negative tumour cell profiles using a stereology grid overlaid on a spot image. DIA was performed with Aperio Genie/Nuclear algorithms. A bias was estimated by ANOVA, correlation and regression analyses. Calibration steps of the DIA by adjusting the algorithm settings were performed: first, by subjective DIA quality assessment (DIA-1), and second, to compensate the bias established (DIA-2). Visual estimate (Ki67-VE) on the same images was performed by five pathologists independently.

Results: ANOVA revealed significant underestimation bias ($P < 0.05$) for DIA-0, DIA-1 and two pathologists' VE, while DIA-2, VE-median and three other VEs were within the same range. Regression analyses revealed best accuracy for the DIA-2 (R-square = 0.90) exceeding that of VE-median, individual VEs and other DIA settings. Bidirectional bias for the DIA-2 with overestimation at low, and underestimation at high ends of the scale was detected. Measurement error correction by inverse regression was applied to improve DIA-2-based prediction of the Ki67-Count, in particular for the clinically relevant interval of Ki67-Count < 40%. Potential clinical impact of the prediction was tested by dichotomising the cases at the cut-off values of 10, 15, and 20%. Misclassification rate of 5-7% was achieved, compared to that of 11-18% for the VE-median-based prediction.

Conclusions: Our experiments provide methodology to achieve accurate Ki67-LI estimation by DIA, based on proper validation, calibration, and measurement error correction procedures, guided by quantified bias from reference values obtained by stereology grid count. This basic validation step is an important prerequisite for high-throughput automated DIA applications to investigate tissue heterogeneity and clinical utility aspects of Ki67 and other immunohistochemistry (IHC) biomarkers.

* Correspondence: arvydas.laurinavicius@vpcl.lt

¹Department of Pathology, Forensic Medicine and Pharmacology, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

²National Center of Pathology, affiliate of Vilnius University Hospital Santariskiu Clinics, Vilnius, Lithuania

Full list of author information is available at the end of the article



© 2014 Laurinavicius et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Introduction

Rapid development of digital pathology technologies, enabling high-resolution scanning of microscopy slides, brings great efficiencies in data storage, transfer and usage in research, clinical practice and education [1-3]. The most unique and significant benefit for pathology practice and research can be expected from digital image analysis (DIA) applications, opening new perspectives for pathology to serve the needs of personalized medicine, by providing more accurate and reproducible measurements for tissue-based diagnosis, prognosis and prediction [4,5]. Microscopic images, used in pathology, contain an enormous amount of data that can be retrieved by numerous methods available to visualise tissue, cell and molecular components, scan and process the images, generating rich multi-parametric data of broad dynamic range. In a broader context of biology, the quest for quantitative microscopy, with support of bio-image informatics, raises the perspective that the days of manually chosen "representative" images are numbered and such images will be replaced by quantitative measures based on the underlying image data [6]. Similarly, pathology is becoming a quantitative or analytical discipline and has to adopt both benefits and obligations that come together [7].

The most immediate benefits of DIA come with increased capacity, precision and accuracy, compared to visual evaluation or counting, used in pathology diagnosis and research. While the capacity and precision (reproducibility and repeatability) aspects are rather obvious, the concept of accuracy (objectivity, correspondence to ground truth, criterion standard or reference values) is less familiar to anatomic pathologists and is frequently confused with the reproducibility aspect. This is probably due to the fact that anatomic pathology has been a qualitative and semi-quantitative discipline for many years, while pathology diagnosis itself was seen as the ground truth in medicine. Therefore, reproducibility rather than accuracy of pathology diagnosis or evaluation was mostly the focus. On the other hand, targeted therapies should be validated against and along with specific biomarker tests, leading to the development of standard testing procedures and clinically validated cut-off values. The validated tests and therapies are considered clinically useful; however, usefulness should not become a substitute for accuracy or objectivity [8].

Standardization of DIA for optimal use in pathology involves many aspects - from tissue processing, sampling, staining, scanning, to DIA settings and proper test validation requirements, as extensively reviewed [8,9]. Although no studies have performed a full scale investigation of every aspect of the DIA process, the combined evidence shows that DIA is able to reproduce data at an acceptable level, with no more variability than manual assessment using conventional microscopy. Meanwhile, validation of DIA has been performed by comparing

digital results with manual estimates, either quantitative or semi-quantitative, or by comparing DIA with another form of criterion standard, for example, fluorescence *in situ* hybridization, or by comparing DIA with clinical (often prognostic) information [9].

Although these validation approaches are common and useful, a criterion standard in these studies is still indirect and may be subject to its own bias. Ideally, to validate and calibrate the DIA tools one should seek the most direct reference values (RV) that answer the same question as the algorithm is intended to do [7]. This means that the same feature in the same image has to be measured by an independent and most possibly objective way; therefore, stereologically sound methods have to be re-introduced to serve the validation and quality assurance of DIA tools; in other words, the DIA tools have to produce stereologically valid results [7,9].

Most useful DIA applications in pathology can be expected today in the area of immunohistochemistry (IHC), a widely-used and relatively inexpensive technology, enabling a broad spectrum of tissue-based biomarkers for personalized therapies; therefore, raising requirements for IHC quantification and accuracy. Not surprisingly, many DIA studies have been targeting IHC markers in breast cancer and other pioneering areas of personalized therapies. As an example, a paradox of an outstanding issue of the cell proliferation marker Ki67 in breast (and other) cancers can be recognized: it is regarded as an important prognostic and predictive factor; however, its clinical utility is hindered by the absence of harmonized methodology of the test [10,11]. Besides the need for accurate enumeration of the proportion of Ki67-positive tumour cell profiles (Ki67 labelling index - Ki67 LI), the issue is further complicated by marked intra-tumour heterogeneity of Ki67 expression in many cases, therefore, demanding standardized sampling of the tissue for the analysis. Although DIA is welcomed, current clinical recommendation asks pathologist to score at least 1,000 cells while 500 cells would be acceptable as the absolute minimum [11].

Gudlaugsson *et al.* [12] have recently compared the reproducibility and prognosis prediction accuracy of different techniques for measurement of Ki67 LI in breast cancer. Two pathologists performed global subjective impression assessment of Ki67 positivity by rapidly scanning/estimating the percentage of Ki67-positive nuclear profiles. Secondly, accurate subjective counts were performed by first identifying hot-spots of Ki67 expression on a whole section at low magnification; in the hot-spot with the subjectively highest Ki67 expression, the Ki67 LI was assessed by two pathologists independently. The third method involved computerized interactive morphometric (CIM) assessment to overcome selection bias. Finally, the DIA was performed on 2 to 10 square areas with the subjectively estimated highest Ki67 LI. The

authors concluded that Ki67 LI by DIA, but not subjective counts, was reproducible and prognostically strong. The CIM was also highly reproducible between the two pathologists, but no direct (image-based) comparison of the CIM and DIA was stated in this report.

We concur with the notions that validation of DIA tools is a multi-step process to consider all potential sources of variation. Presuming that pre- and analytical IHC variation needs to be dealt with by routine quality assurance processes, the DIA methods add unique processes of slide scanning, region of interest selection, object segmentation, characterization, enumeration and evaluation. Yet, it is hardly possible to properly address all aspects in one study. With the aim to develop a sound DIA validation and calibration methodology, we designed our experiment to test and improve the accuracy of Ki67 LI estimation by automated DIA on preselected tissue microarray (TMA) Ki67 IHC images, with the ground truth obtained by counting tumour cell profiles using a stereology test grid of systematically sampled frames. We therefore minimized the impact of the tissue heterogeneity and IHC variability, aspects to be addressed separately. In addition, we evaluated the accuracy of visual assessment (impression) of five pathologists on the same images, to simulate the widely used practices to test DIA results against visual estimates or their averaged values.

Materials and methods

Population

This study was performed on TMA images from 164 female patients with an invasive ductal carcinoma of the breast, treated at the Oncology Institute of Vilnius University and investigated at the National Center of Pathology, during the period of 2007 to 2009. The study was approved by the Lithuanian Bioethics Committee. The patients' consent to participate in the study was obtained.

Tissue preparation

The TMAs were constructed, stained and scanned as described previously [13]. Briefly, one millimetre-diameter cores were punched from tumour areas randomly selected by the pathologist and paraffin sections were cut at 3 µm-thickness.

Immunohistochemistry (IHC)

IHC for Ki67 was performed with a multimer-technology based detection system, ultraView Universal DAB (Ventana, Tucson, AZ, USA). The Ki67 antibody (clone MIB-1; DAKO, Glostrup, DK) was applied at a 1:200 dilution for 32 minutes, followed by the Ventana BenchMark XT automated immunostainer (Ventana) standard Cell Conditioner 1 (CC1, a proprietary buffer) at 95°C for 64 minutes. Finally, the sections were developed in DAB at 37°C

for eight minutes, counterstained with Mayer's hematoxylin and mounted.

Image acquisition

Digital images were captured using the Aperio ScanScope XT Slide Scanner (Aperio Technologies, Vista, CA, USA) under 20x objective magnification (0.5 µm resolution). One TMA spot image per patient was used for the study.

Quantification with stereology test grid

RV were obtained by marking Ki67-positive and negative tumour cell profiles, using a stereological method for 2D object enumeration [14,15] implemented by the Stereology module (ADCIS, Caen, France) with a test grid of systematically sampled frames (frame size - 125 pixels, spacing of frames - 250 pixels) overlaid on a spot image in ImageScope (Aperio Technologies, USA), Figure 1. The percentage of Ki67 positive tumour cell profiles established by the test grid estimation (Ki67-Count) was calculated as $100 \times \text{Ki67-positive nuclear profiles} / (\text{Ki67-positive nuclear profiles} + \text{Ki67-negative nuclear profiles})$. To test the degree of uncertainty of the RV, inter-observer variation was estimated based on Ki67-Count values produced by three observers (Ki67-Count-1, 2 and 3) independently in a subset ($n = 30$) of the TMA images. Since the inter-observer variability was found to be negligible (see Results), the RV in the whole series ($n = 164$) were established by one-observer marking (Ki67-Count), splitting the job among four observers in approximately equal proportions. Estimated time to produce cell marks on the frame grid was 30 minutes per one TMA spot image on average but varied due to variable cellularity of the tumour tissue. Also, the uncertainty of the RV was estimated through Coefficient Error (CE) computation, according to the sampling theory [16]: this uncertainty originates from the fact that the frame count is performed on the subsampled tissue and is calculated as $CE = t \cdot \sqrt{Cg \cdot (m/n^2)}$ [17] with n being the number of frames inside the tumour, m being the number of the external sides of the set of frames in the tumour. For the test grid (Figure 2) with a frame size of 125 pixels and a frame spacing of 250 pixels, the value of the grid factor Cg is 0.049. Otherwise, the value of the Student factor t is 2 for a confidence of 95% and for an event number greater than 30.

Visual evaluation (VE)

A global subjective impression for the Ki67 LI on the same images was performed by five pathologists independently and provided semi-quantitative values (Ki67-VE-1, 2, 3, 4 and 5) expressed as the percentage of Ki67-positive tumour cell profiles. Counting was not included in the procedure.

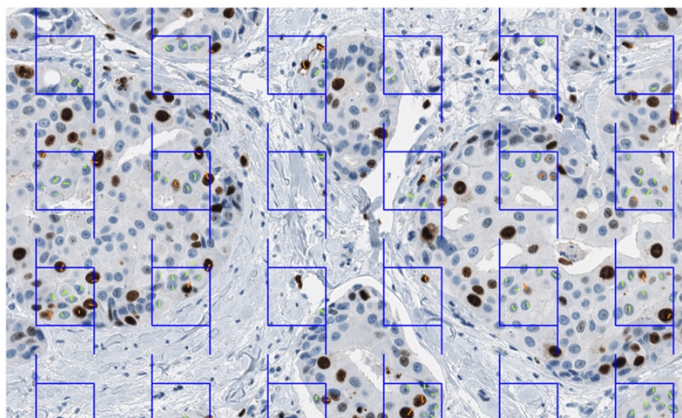


Figure 1 Test grid of frames from the stereology module overlaid on the TMA spot image. The left and bottom lines of a frame are “forbidden” - nuclear profiles intersecting them are not marked. The short line marks (orange for Ki67-positive, green for Ki67-negative tumour cell nuclear profiles) are produced manually by an observer. Total numbers and Ki67 LI are computed by the software at the end of the procedure. TMA, tissue microarray.

Digital Image Analysis

DIA was performed with Aperio Genie and Nuclear v9 algorithms enabling automated selection of the tumour tissue (the Genie Classifier was trained to recognize tumour tissue, stroma and background (glass), then combined with the Nuclear algorithm). Several calibration cycles of the DIA (named DIA-0, 1 and 2, resulting

in the percentage of Ki67-positive tumour cells - Ki67-DIA-0, 1 and 2, respectively) were performed to improve the accuracy of the tool by adjusting the settings of the Nuclear algorithm (Table 1). Ki67-DIA-0 was obtained by the default Aperio settings for the Nuclear algorithm, Ki67-DIA-1 - by “subjective” visual assessment of the quality of the DIA results on the computer monitor; Ki67-DIA-2 was fine-tuned based on the quantitative bias established by statistical analyses comparing the Ki67-DIA-1 to RV (Ki67-Count). Highly automated calibration cycles were achieved by developing software to integrate the DIA outputs and statistical analysis procedures.

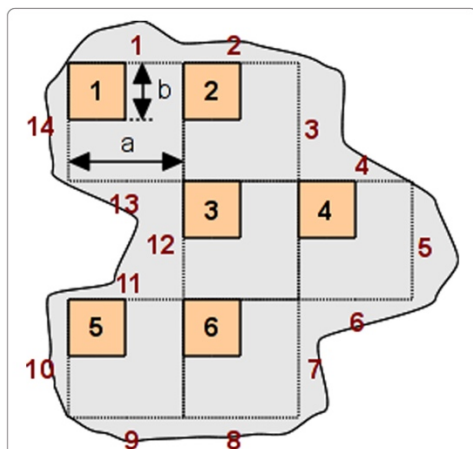


Figure 2 Tumour area (grey) and test grid of systematically sampled frames (orange) ($a = 250$ pixels, $b = 125$ pixels). For this example, the number of frames is $n = 6$ and the number of external segments is $m = 14$.

Statistical analysis

Accuracy of the DIA and VE with regard to the RV was estimated by one-way ANOVA (Duncan multiple range test was used for pairwise comparisons), Pearson correlation, single and multiple linear regression analyses, as well as orthogonal linear regression based on principal component analysis. Agreement between individual measurements was also estimated based on 95% confidence intervals calculated from the RV CE and visualized by Bland and Altman plots [18]. Dependence of RV ($n = 30$) and VE ($n = 164$) inter-observer variation on the magnitude of measurement was visualized by plots of corresponding standard deviations against the mean values of the measurements. A variable degree of right asymmetry (skewness from 0.5 to 1.6) of the parameter distribution was noted; where appropriate, statistical significance of the findings was verified, using log-transformed data. Statistical significance level was set at $P < 0.05$. Statistical analyses were performed with

Table 1 Nuclear algorithm settings for the DIA calibration after the Genie classifier

Algorithm setting	DIA-0	DIA-1	DIA-2
Averaging radius (μ)	1	1	1
Curvature threshold	2.5	2.5	2.5
Segmentation type	Cytoplasm rejection	Cytoplasm rejection	Cytoplasm rejection
Threshold type	Edge threshold	Edge threshold	Edge threshold
Lower intensity threshold	0	0	0
Upper intensity threshold	220	230	230
Min. nuclear size (μ^2)	20	45	40
Max. nuclear size (μ^2)	1,000,000	1,000	1,000
Min. roundness	0.1	0.1	0.1
Min. compactness	0	0	0.2
Min. elongation	0.1	0.1	0.2
Remove light objects	removes no nuclei	removes no nuclei	removes no nuclei
Weak (1+) threshold	210	210	229
Moderate (2+) threshold	188	188	188
Strong (3+) threshold	162	162	162
Black threshold	0	0	0
Edge trimming	Weighted	Weighted	Weighted

DIA-0 (default), DIA-1 (subjective) and DIA-2 (based on quantified bias). Modified settings are highlighted in bold.

SAS 9.3 software, Microsoft Excel software (Microsoft, Redmond, Washington, USA) and OpenOffice Calc software (Oracle, Redwood City, California, USA).

Results

Characteristics and measurement uncertainty of the reference value dataset

Summary statistics of the RV ($n = 30$) obtained by three independent observers' marking of the tumour cell profiles in the test grid are presented in Table 2, along with the results of other measurements in this dataset for reference. No significant variance between the three Ki67-Counts was revealed by one-way ANOVA ($F = 0.08$, $P = 0.9217$), while strong pairwise correlation among the values was found: $r = 0.98$, $r = 0.98$, $r = 0.97$ ($P < 0.0001$). Similarly, the total number of nuclear profiles marked did not differ significantly, although Observer 1 tended to mark less; the total number of nuclear profiles of Observer 1 correlated with that of observers 2 and 3 at $r = 0.94$, while the latter two correlated at $r = 0.98$ ($P < 0.0001$).

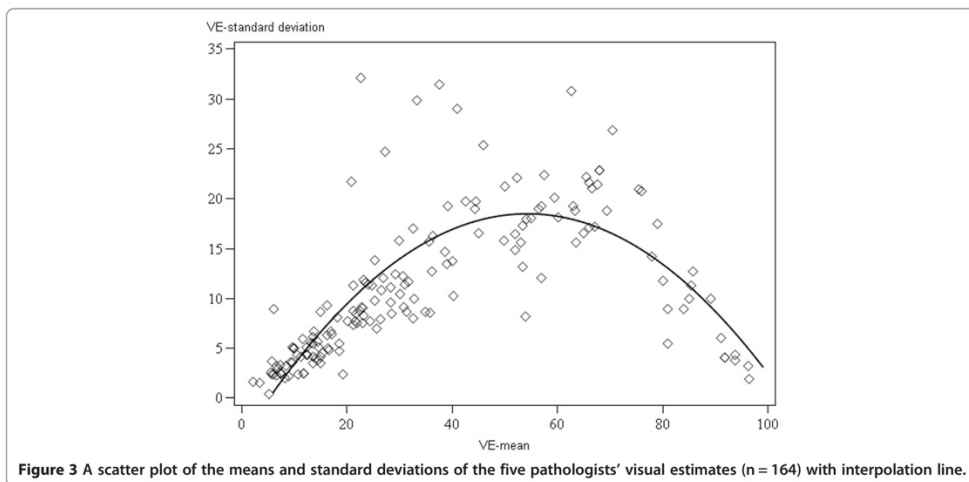
Uncertainty introduced by variance among the three counts to produce Ki67-Count for each individual spot

Table 2 Summary statistics of the reference values produced by observers, visual estimates and image analyses ($n = 30$)

Variable	Median	Mean	Std dev	Std error	Min	Max
Ki67-Count-1	21.7	28.6	20.4	3.7	0.3	72.6
Ki67-Count-2	24	29.9	19.5	3.6	0.6	69.7
Ki67-Count-3	23	28.7	18.6	3.4	1.2	69.4
Ki67-Count-median	24	29.3	19.4	3.5	0.6	67.4
Ki67-Count-mean	23.4	29.1	19.4	3.5	0.7	66.8
Total profiles Observer 1	331	425.7	273.7	50	85	1,098
Total profiles Observer 2	509	590.7	385.4	70.4	143	1,863
Total profiles Observer 3	471.5	547.2	331.9	60.6	146	1,544
Ki67-VE-1	10	18.3	15.3	2.8	5	70
Ki67-VE-2	30	40.2	29.4	5.4	2	95
Ki67-VE-3	37.5	41.4	27.7	5.1	1	90
Ki67-VE-4	20	30.2	23	4.2	4	80
Ki67-VE-5	22.5	31	24.1	4.4	1	90
Ki67-VE-median	22.5	32.5	25	4.6	2	90
Ki67-VE-mean	23.4	32.2	23.2	4.2	6.2	80
Ki67-DIA-0	16.1	19.9	12.5	2.3	2.1	50
Ki67-DIA-1	18.5	24.8	15.9	2.9	1.6	65.5
Ki67-DIA-2	22.8	29.1	15.7	2.9	9.1	68.4

was low: for the 30 spots, mean standard deviation and mean standard error were 2.6% and 1.5%, respectively. Of note, the five visual estimates (Ki67-VE), summarized for the same individual 30 spots, revealed much higher uncertainty - mean standard deviation and mean standard error were 10.9% and 4.9%, respectively. Interestingly, a scatter plot of the five VE' standard deviations against their means ($n = 164$, Figure 3) uncovered non-linear relationship reflecting higher variation in the middle of the means' scale. Meanwhile, the positive linear relationship between the standard deviations and the means for the three observers of Ki67-Count was found in the dataset available ($n = 30$, not shown).

Uncertainty caused by subsampling the tissue by the test grid of frames was estimated by computation of CE providing confidence intervals for each individual Ki67-Count value. Overlap of the Ki67-Count confidence intervals for all three and each pair of the three observers was considered as agreement between the generated Ki67 LI values (Figure 4). The agreement within the same confidence interval among all three measurements was 69%; whereas the pairwise agreement varied from 83 to 86%. The uncertainty of the RV generated was therefore considered satisfactory. The RV for the whole image dataset ($n = 164$) were based on a single observer



count per spot (Ki67-Count). Yet, the subsampling uncertainty was further taken into account in the accuracy estimates.

Accuracy of the image analysis and visual estimates with regard to the reference values

Summary statistics of the RV, DIA and VE variables (n = 164) are presented in Table 3. One-way ANOVA revealed significant variance explained by the measurement method overall (Figure 5, $P < 0.0001$). Pairwise

comparisons (Table 4) revealed no significant bias among the Ki67-Count and Ki67-VE-2 and Ki67-VE-3 estimates (Duncan grouping A) or Ki67-VE-5, Ki67-VE-median and Ki67-DIA-2 (Duncan grouping B). Meanwhile, Ki67-DIA-0, Ki67-DIA-1, Ki67-VE-1 and Ki67-VE-4 produced significantly lower values.

Pairwise correlations (Table 5) were highly significant ($P < 0.0001$). Remarkably, correlation between Ki67-Count and Ki67-DIA-0, 1 and 2 improved which each calibration cycle from $r = 0.928$ to $r = 0.949$. Notably, Ki67-Count

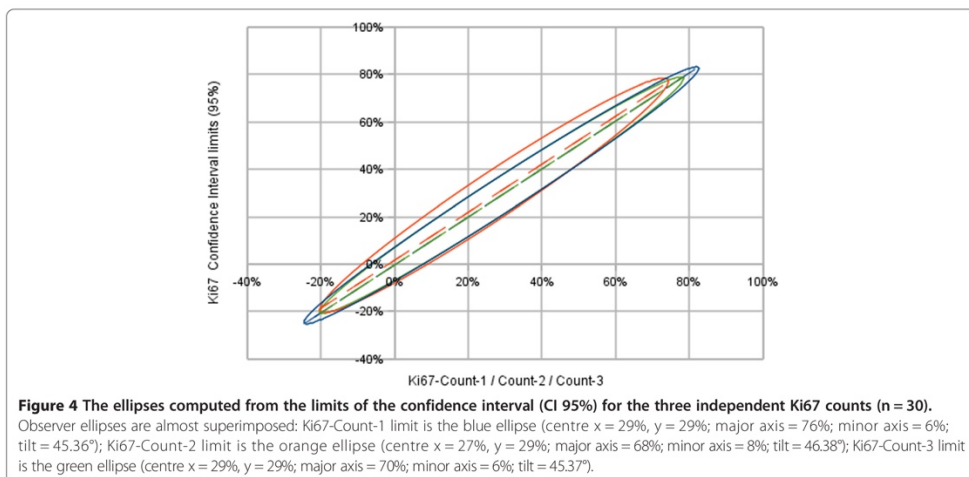


Table 3 Summary statistics of the reference values produced by three observers with the corresponding data of visual estimates and digital image analysis, n = 164

Variable	Median	Mean	Std dev	Std error	Min	Max
Ki67-Count	35.0	40.2	25.3	2.0	0.6	98.1
Ki67-DIA-2	30.1	36.5	20.2	1.6	6.4	93.0
Ki67-DIA-1	24.1	31.1	21.1	1.6	1.5	90.5
Ki67-DIA-0	20.4	25.9	18.1	1.4	2.1	85.7
Visual median	30	37.2	27.4	2.1	2	95
Visual mean	28.4	36.2	25.6	2.0	2.2	96.4
Ki67-VE-1	15	24.3	23.6	1.8	5	95
Ki67-VE-2	40	43.4	29.6	2.3	2	98
Ki67-VE-3	37.5	44.1	30.0	2.3	1	99
Ki67-VE-4	22	31.6	24.3	1.9	1	95
Ki67-VE-5	30	37.7	27.7	2.2	1	100
Total profiles observer*	2,372	2,658.7	1,390.4	108.6	464	7,452
Total profiles DIA-2	2,150.5	2,293.2	796.8	62.2	752	4,302
Total profiles DIA-1	1,920.5	2,022.7	670.1	52.3	1,012	3,788
Total profiles DIA-0	4,203.5	4,385.0	1,420.2	110.9	1,640	7,939

*Total nuclear profiles observer counts are multiplied by four in this table to be comparable to the DIA total profile numbers (the box grid used for the observer count covers ne-fourth of the image area). DIA, digital image analysis; VE, visual estimate.

correlated with the Ki67-VE-median strongest ($r = 0.930$), in comparison to the correlations with the individual VE measurements.

Single linear regression analyses for the DIA and VE results as dependent variables and the RV as explanatory variables produced highly significant ($P < 0.0001$) models in all cases (Table 6). Remarkably, determination coefficients (R-square) improved with each calibration cycle of the Ki67-DIA-0, 1, and 2 from 0.86 to 0.89 and 0.90. Notably, R-square for the VE-median (0.86) was the highest amongst the individual VE but reached only that of the Ki67-DIA-0.

The correspondence between the Ki67-DIA-2 and the RV was also tested, taking into account the uncertainty of the RV related to the subsampling of the tissue by the test grid. The confidence interval for the RV was calculated and the Ki67-DIA-2 values were tested for fitting the confidence interval (Figure 4). The R-square of the model was 0.90, the accuracy factor was 0.82. Interpretation of the plot and the slope tilt from the ellipse axis revealed a bias: underestimation of the Ki67-Count by the Ki67-DIA-2 observed at the higher end of the RV scale as well as overestimation at the low end. Similarly, Bland and Altman plots (not shown) reflected the same bidirectional bias dependent on the magnitude of the measurement. Orthogonal linear regression analysis for the DIA and RV was used to refine the accuracy value reducing the intercept factor. In this case, the ratio of the tilt of the regression line and the tilt of the ellipse axis defining the accuracy factor was 0.92 (Figure 6).

Outliers of the Ki67-DIA-2 versus RV analyses were inspected to explore potential reasons of the underestimation. In general, the tumour tissue was highly cellular

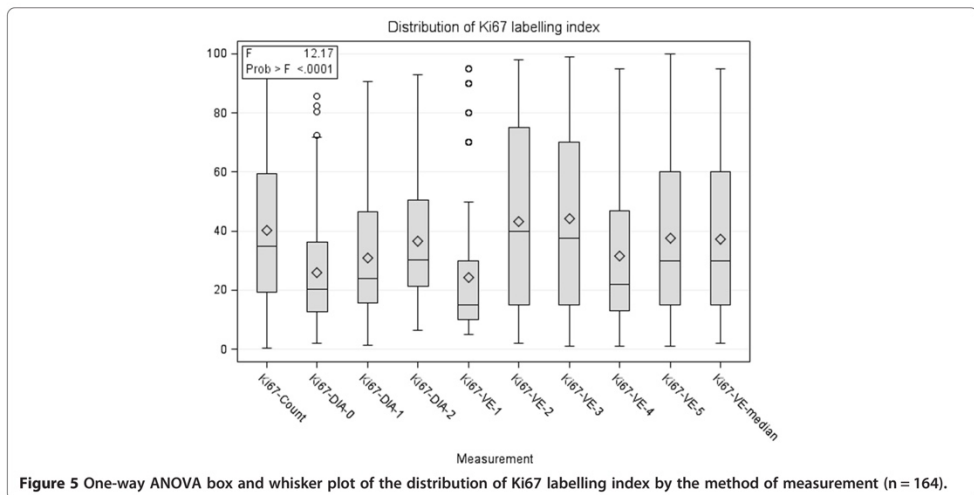


Figure 5 One-way ANOVA box and whisker plot of the distribution of Ki67 labelling index by the method of measurement (n = 164).

Table 4 Pairwise comparisons for the means of reference values, visual estimates and digital image analysis results, n = 164

Duncan grouping*	Mean	Measurement
A	44.1	Ki67-VE-3
A	43.4	Ki67-VE-2
B	40.2	Ki67-Count
B	37.7	Ki67-VE-5
B	37.2	Ki67-VE-mean
B	36.5	Ki67-DIA-2
C	31.6	Ki67-VE-4
C	31.1	Ki67-DIA-1
F	25.9	Ki67-DIA-0
F	24.3	Ki67-VE-1

*Means with the same letter are not significantly different at $P < 0.05$. DIA, digital image analysis; VE, visual estimate.

in many cases resulting in overlapping nuclei and their confluence and/or rejection by maximum size limit at the DIA. Also, in some cases, tissue artefacts, an admixture of stroma with lymphocytes and large ducts could impact the DIA results. Further fine-tuning of the Nuclear algorithm settings was attempted without notable success.

Prediction of the reference values by inverse regression and measurement error correction

Ki67-DIA-2 enabled fair accuracy and outperformed the 5 VE measurements, both individual and the median. Yet, the measurement bias for the Ki67-DIA-2 was established and enabled a measurement error correction procedure to be used to predict the ground truth in real life with maximum accuracy. Inverse regression analyses were performed to retrieve the correction criteria (Table 7). To avoid the potential impact of some non-linearity noted and to derive the most useful inverse regression model for accurate prediction of the ground

truth in the interval of clinical importance, a regression model $Ki67-DIA-2 < 40$ was produced, based on the observations with Ki67-Count values less than 40% ($n = 92$).

In addition to the single regression models, multiple regression models with inclusion of both Ki67-DIA-2 and Ki67-VE-Median gave slightly higher R-square value (0.91) than the Ki67-DIA-2 alone (0.90). Therefore, the DIA approach with calibration of the algorithm settings based quantified bias enabled most accurate measurement of the Ki67 LI, while VE of five pathologists were consistent but gave little added value in terms of accuracy, compared to the automated DIA measurement.

Effect of the prediction and measurement error correction on Ki67 dichotomisation accuracy

The effect of VE and DIA inverse regression models to predict the RV on accuracy of patient dichotomisation at RV cutoffs of clinical importance (>10, 15 and 20%) was tested (Table 8). The cutoffs used for these simulations were the ones most commonly considered as clinically relevant to test potential clinical impact of the measurement methods involved in our study. While Ki67-VE-median tended to underestimate the Ki67-Count-based class at all cutoffs, especially at 20%, the Ki67-DIA-2 prediction overestimated the classes, especially at the lower end (>10%) of the scale. Total misclassification rate at different cutoffs varied from 11 to 18% for the VE-based and 5 to 9% for the DIA-based prediction, respectively.

The effect of measurement error correction for the Ki67-DIA-2-based prediction of the RV was tested with the values obtained by the inverse regression formula $Ki67-DIA-2-corrected = 1.1878 * Ki67-DIA-2 - 3.1183$ and, to minimize potential non-linearity impact for the prediction accuracy, by the formula $Ki67-DIA-2-corrected < 40 = 1.1472 * Ki67-DIA-2 - 4.3913$ (Tables 7 and 8). The error correction for both prediction models (especially, the Ki67-DIA-2-corrected <40 model) decreased the DIA

Table 5 Pairwise correlations between the reference values, visual estimates and digital image analysis results (Pearson's coefficients, $P < 0.0001$, n = 164)

Measurement	Ki67-count	Ki67-DIA-2	Ki67-DIA-1	Ki67-DIA-0	Visual median	Ki67-VE-1	Ki67-VE-2	Ki67-VE-3	Ki67-VE-4
Ki67-DIA-2	0.949								
Ki67-DIA-1	0.945	0.989							
Ki67-DIA-0	0.928	0.974	0.976						
Ki67-VE-median	0.930	0.940	0.946	0.927					
Ki67-VE-1	0.861	0.917	0.921	0.925	0.891				
Ki67-VE-2	0.905	0.905	0.915	0.886	0.955	0.829			
Ki67-VE-3	0.921	0.921	0.931	0.900	0.969	0.857	0.972		
Ki67-VE-4	0.887	0.894	0.895	0.884	0.936	0.857	0.881	0.901	
Ki67-VE-5	0.842	0.869	0.872	0.860	0.916	0.822	0.853	0.872	0.829

DIA, digital image analysis; VE, visual estimate.

Table 6 Single linear regression models with reference values as explanatory variable (n = 164, P <0.0001 for all models and slope estimates)

Variable	R-square	Intercept estimate	Intercept P	Slope estimate	Slope standardized estimate
Ki67-DIA-2	0.90	5.9692	<0.0001	0.7588	0.9494
Ki67-DIA-1	0.89	-0.6389	0.5324	0.7892	0.9447
Ki67-DIA-0	0.86	-0.9576	0.3389	0.6667	0.9278
Ki67-VE-median	0.86	-3.3799	0.0242	1.0093	0.9316
Ki67-VE-1	0.74	-8.1114	<0.0001	0.8057	0.8514
Ki67-VE-2	0.82	0.6733	0.7180	1.0616	0.9049
Ki67-VE-3	0.85	0.1337	0.9382	1.0926	0.9210
Ki67-VE-4	0.79	-2.7516	0.0987	0.8545	0.8545
Ki67-VE-5	0.71	0.4763	0.8294	0.9245	0.8422

DIA, digital image analysis; VE, visual estimate.

overestimation effect at the >10% cutoff. While total misclassification rate at different cutoffs for Ki67-DIA-2-corrected remained in the interval of 5 to 9%, the Ki67-DIA-2-corrected <40-based prediction enabled some improvement down to the misclassification rate of 5 to 7%.

In summary, the DIA-based prediction of the RV enabled the classification error rate half of that of the VE-based prediction, it was less than 10% at all cutoffs tested, and could be further improved by the measurement error correction attempts.

Discussion

Our experiment presents test validation, calibration and measurement error correction methodology that can be successfully applied to ensure and improve accuracy of IHC Ki67 LI estimation by DIA. In essence, in our

approach we sought to adopt the principles of analytic test validation for IHC DIA-based enumeration of Ki67 LI with quantification of the measurement bias by comparison to the Ki67 LI obtained on the same images by stereological test grid count as most direct criterion standard. Our first step of the DIA calibration (DIA-0 to DIA-1) was achieved by visual (intuitive) quality assessment of the DIA results on the computer monitor of selected images, while the second (DIA-1 to DIA-2) was based on quantified bias from the criterion standard. Our results show that only after the second (quantitative) calibration step, global bias of the DIA became not significant, while regression analyses revealed gradual improvement of the prediction of the DIA outputs with the calibration steps. Although the calibrated DIA-2 revealed the best accuracy achieved, exceeding that of the VE, nonlinearity was noted with some overestimation

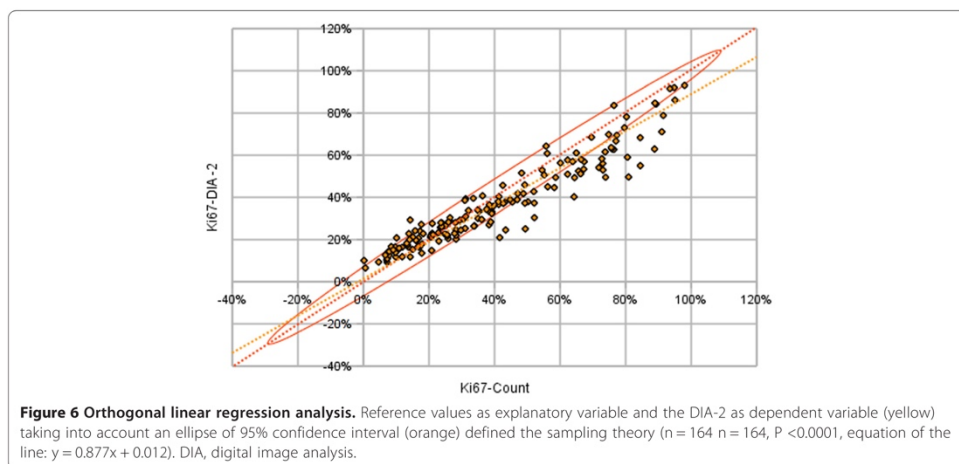


Table 7 Single and multiple linear inverse regression models to predict reference values as dependent variable (n = 164, P < 0.0001 for all models and slope estimates)

Variable	R-square	Intercept estimate	Intercept P	Slope estimate	Slope standardized estimate
Single regression models:					
Ki67-DIA-2	0.90	-3.1183	0.0165	1.1878	0.9494
Ki67-DIA-2 < 40*	0.75	-4.3913	0.0085	1.1472	0.8688
Ki67-DIA-1	0.89	5.0453	<0.0001	1.1309	0.9447
Ki67-DIA-0	0.86	6.8232	<0.0001	1.2916	0.9278
Ki67-VE-median	0.86	8.3195	<0.0001	0.8572	0.9302
Multiple regression model					
Ki67-DIA-2	0.91	-0.3245	0.8096		
Ki67-VE-median				0.8068	0.6448
				0.2985	0.3239

*Ki67-DIA-2 < 40 - represents a regression model for Ki67-DIA-2 with only Ki67-Count less than 40% cases included in the analysis (n = 92). DIA, digital image analysis; VE, visual estimate.

bias on the low and underestimation bias at the high end of the scale. We subsequently applied measurement error correction procedures by inverse regression to further enhance the DIA test applicability. Finally, we tested potential clinical impact of the accuracy achieved by applying DIA- and VE-based predictions of Ki67 LI to dichotomize patients (images) by frequently used cut off values at 10, 15 and 20% and found that DIA (after quantitative calibration and measurement error correction) enabled the classification error rate 2x less than that of the VE.

In our study we did not strictly follow the guidelines for analytical test validation [19] since the nature of the subject and the criterion standard (IHC image) used are still different from the analytical test samples used in

medicine. First, the uncertainty of our criterion standard was tested by independent measurements by three observers on a subset (n = 30) of images and was considered as satisfactory to further rely on one observer counts. Nevertheless, the inter-observer comparison was image-based but not cell-based, and we realize that human judgment/error is still involved when deciding on individual tumour/non-tumour and positive/negative cells even in this stereologically-based approach. Although some uncertainty of our criterion standard has to be taken into account, our data show that it is more reliable than that of the VE consensus of several pathologists and, therefore, should be used for DIA validation needs. Secondly, we have not tested the repeatability of the tests: it would be beyond reasonable effort to repeat

Table 8 Effect of the inverse regression-based prediction and measurement error correction on Ki67 dichotomisation accuracy at various reference value cutoffs (n = 164)

Method	Underestimated (%)	Overestimated (%)	Total misclassified (%)
Ki-67 cutoff >10%			
Ki67-VE-median	16/148 (11)	2/16 (13)	18 (11)
Ki67-DIA-2	0/148 (0)	12/16 (75)	12 (7)
Ki67-DIA-2 corrected	0/148 (0)	9/16 (56)	9 (5)
Ki67-DIA-2 corrected <40*	2/148 (1)	6/16 (38)	8 (5)
Ki-67 cutoff >15%			
Ki67-VE-median	22/136 (16)	1/28 (4)	23 (14)
Ki67-DIA-2	2/136 (1)	13/28 (46)	15 (9)
Ki67-DIA-2 corrected	3/136 (2)	11/28 (46)	14 (9)
Ki67-DIA-2 corrected <40*	5/136 (4)	6/28 (21)	11 (7)
Ki-67 cutoff >20%			
Ki67-VE-median	28/123 (23)	1/41 (2)	29 (18)
Ki67-DIA-2	2/123 (2)	9/41 (22)	11 (7)
Ki67-DIA-2 corrected	2/123 (2)	12/41 (29)	14 (9)
Ki67-DIA-2 corrected <40*	6/123 (5)	6/41 (15)	12 (7)

*Ki67-DIA-2 < 40 - represents a regression model for Ki67-DIA-2 with only Ki67-Count less than 40% cases included in the analysis (n = 92). DIA, digital image analysis; VE, visual estimate.

the stereological count manually and less important to repeat (intra-observer) VE, since it was not our main focus to investigate the VE accuracy and precision. The repeatability of the DIA in strictly the same conditions is expected to be perfect, the reproducibility of the DIA involving all phases of the Ki67 LI test as well as its robustness to the IHC staining variation was beyond the scope of this study. Third, we did not validate our DIA prediction accuracy on an independent dataset, since it requires another set of criterion standard data that is planned as an output of our next experiment. Fourth, the DIA validation tests in the present study are based on summarized data per image (Ki67 LI), while more rigid individual cell-based comparisons would provide even more granular information on the performance of the DIA tools.

Our approach uncovered a non-linear bias in the opposite directions depending on the magnitude of the measurement of the Ki67-DIA-2, which could hardly be documented by only the subjective assessment of the DIA accuracy of selected images on a computer monitor. To better understand why DIA-2 overestimated the Ki67 LI at the low and underestimated it at the high ends of the scale, we compared absolute numbers of positive and negative tumour cell profiles detected by the DIA-2 and box grid count (data not shown). We found that with increasing both Ki67 LI and the total number of cell profiles counted, DIA-2 tended to under-detect tumour cells, while this effect was more notable for positive cells. To really explore the sources of the bidirectional bias, one needs to design more sophisticated cell-based quality assurance procedures which would also allow testing the impact of cell density and other features. From our data, we can speculate that increased tumour cellularity with more cell profiles overlapping may impact nuclear segmentation quality, while the impact of the automated tumour tissue segmentation by the Genie algorithm remains to be deciphered. In general, this nonlinearity phenomenon seems to originate from "subject-measurement" interaction, where the measured subject has variable characteristics (tumour cellularity, density, texture, staining, section thickness and so on) and a specific DIA algorithm may handle them with variable success. This further highlights the complexity of the automated DIA approaches and the need of appropriate validation and quality assurance procedures.

Although the VE validation was not the main focus of our study, we observed an interesting nonlinear dependence of inter-observer variation on the magnitude of the measurement: high standard deviations in the five VE observers' mean were noted in the middle of the Ki67 LI scale. This finding is somewhat unexpected, but still consistent with the observation that IHC biomarker distribution artefacts may be generated by subjective visual

scoring [20]. Without going into extended speculations on the potential sources of this variation, we see it as additional evidence that individual VE or "eyeballing" cannot serve as reliable measurement when there is increased clinical demand for quantification accuracy. The "consensus" or median VE of five pathologists ensured better accuracy than individual VE; however, it did not reach that of the calibrated DIA. Furthermore, besides being less accurate, precise and practical for clinical use, multi-observer VE should not be used as a criterion standard method for the DIA validation purposes, because of its greater uncertainty level compared to that of DIA or count-based methods.

The deepening gap between the potential clinical utility of the Ki67 LI and availability of robust measurement methodologies is reflected by the St Gallen 2013 consensus [21]: while the cut off <14% remains in the definition of the Luminal A-like tumours, a majority voted for the threshold of $\geq 20\%$ to define "high" Ki67 status. Furthermore, a concern about the possible under-treatment of patients with luminal disease who might benefit from chemotherapy, justifies use of a lower (local laboratory specific) cut-off to define Ki-67 "high" or use of multi-gene-expression assay results. This approach would potentially require validation studies with clinical outcomes while the measurement methods remain not standardized. In the situation where one laboratory may serve different oncology units, this would become even less realistic. In addition, it is worth noting that there is a fundamental issue in defining and reproducing Ki67 LI cut-offs with the distribution pattern when the great majority of the hormone-receptor positive breast tumours fall into the Ki67 KI interval between 10 to 20%. Therefore, it is intrinsically difficult to meet the clinical demand for accuracy without measurement methods of established and controlled accuracy, preferably indicating confidence intervals for the values. Even more, combinatorial or multiple IHC biomarker systems may be needed to achieve robust prognostic and predictive indicators [13,22,23].

While manual techniques, including VE and counting, have been shown to be poorly reproducible, even at the level of decision on individual cells [24], the only viable alternative to extract most accurate Ki67 LI by IHC test is further sophistication and standardization of DIA methodologies. They enable greater capacity which also involves counting more cell profiles in more tissue samples, which in turn may lead to better accuracy at the low end of the Ki67 LI scale [25]. The success of the DIA in IHC quantification may be variable and depend both on the DIA tools used and a study design. For breast cancer Ki67 LI measurement, DIA has been shown to be comparable to the VE but of less prognostic value by one study [26] or better than VE, comparable

to CIM and of stronger prognostic accuracy by another [12]. Even if it is tempting (and useful) to validate a DIA tool to predict specific clinical outcomes, we argue that sound DIA measurement methods should be developed and maintained by meeting the “basic needs” first to quantify the measurement bias from affordable and most objectively established criterion standard. As put by Bland and Altman [18], “some lack of agreement between different methods of measurement is inevitable, what matters is the amount by which methods disagree”. We, therefore, position our experiment as the first step in DIA validation process to ensure accurate estimation of Ki67 LI in a selected tissue sample, with subsequent steps to use automated DIA to address tissue heterogeneity and sampling issues as well as prediction of clinical outcomes.

Conclusions

In general, we suggest that proper quantitative validation and calibration methodologies can and have to be employed to establish and ensure accuracy of Ki67 LI measurement by DIA and digital IHC. The measurement accuracy can be further improved by measurement error correction based on the quantified bias, which in our study allowed to decrease patient misclassification rate by the Ki67 LI cut offs of 10, 15 and 20% down to 5 to 7%, compared to that of the VE consensus of five pathologists at 11 to 18%. This basic validation step also opens better perspectives to use high-throughput automated DIA tools to investigate tissue heterogeneity and clinical utility aspects of Ki67 and other IHC biomarker expression.

Abbreviations

CE: Coefficient error, computed according to the sampling theory; CIM: Computerized interactive morphometric assessment; DIA: Digital image analysis; IHC: Immunohistochemistry; Ki67 LI: Ki67 labelling index; Ki67-Count: Percentage of Ki67 positive tumour cell profiles established by the stereology test grid count; Ki67-DIA: Percentage of Ki67 positive tumour cell profiles established digital image analysis; Ki67-VE: Percentage of Ki67 positive tumour cell profiles obtained by semi-quantitative visual estimate of a pathologist; RV: Reference values; TMA/TMAs: Tissue microarray/tissue microarrays; VE: Visual evaluation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ArL drafted the manuscript, performed statistical analysis, performed stereology count and participated in image analysis calibration experiments. BP drafted essential parts of the manuscript and performed statistical analysis. Ail designed and carried out the image analyses, performed and supervised stereology measurements, and edited the manuscript. PH and NE designed the stereology measurements and edited the manuscript. DD, IB, JB and RM performed stereology measurements. CB, DD, IB and RM performed visual evaluation of the Ki67-LI on the TMA images. YI produced software for TMA image analysis output conversion to streamline statistical analysis cycles and calibration process. IE performed visual evaluation of the Ki67-LI on the TMA images, edited the manuscript and assessed clinical relevance of the findings. All authors participated in conception and design of the study, reviewed the analysis results, and critically revised and approved the final manuscript.

Acknowledgement

This research is funded by European Social Fund under the Global Grant measure.

Author details

¹Department of Pathology, Forensic Medicine and Pharmacology, Faculty of Medicine, Vilnius University, Vilnius, Lithuania. ²National Center of Pathology, affiliate of Vilnius University Hospital Santariskiu Clinics, Vilnius, Lithuania. ³Path-Image/BioTiClA, University of Caen, Caen, France. ⁴Pathology Department, F. Baclesse Comprehensive Cancer Center, Caen, France. ⁵Department of Histopathology, Molecular Medical Sciences, University of Nottingham, Nottingham, UK.

Received: 19 October 2013 Accepted: 26 March 2014

Published: 6 April 2014

References

- Gu J, Ogilvie RW: Virtual microscopy and virtual slides in teaching, diagnosis and research. In *Advances in Pathology, Microscopy & Molecular Morphology*. Edited by Gu J, Hacker GW. Boca Raton, London, New York, Singapore: CRC Press, Taylor & Francis Group; 2005.
- Soenksen D: Digital pathology at the crossroads of major health care trends: corporate innovation as an engine for change. *Arch Pathol Lab Med* 2009, **133**:555–559.
- Kayser K, Borkenfeld S, Kayser G: How to introduce virtual microscopy (VM) in routine diagnostic pathology: constraints, ideas, and solutions. *Anal Cell Pathol (Amst)* 2012, **35**:3–10.
- Kayser K, Gortler J, Borkenfeld S, Kayser G: How to measure diagnosis-associated information in virtual slides. *Diagn Pathol* 2011, **6**:59.
- Madabhushi A, Agner S, Basavanahally A, Doyle S, Lee G: Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. *Comput Med Imaging Graph* 2011, **35**:506–514.
- The quest for quantitative microscopy. *Nat Methods* 2012, **9**:627. DOI: 10.1038/nmeth.2102
- Laurinavicius A, Laurinaviciene A, Dasevicius D, Elie N, Plancoulaine B, Bor C, Herlin P: Digital image analysis in pathology: benefits and obligation. *Anal Cell Pathol (Amst)* 2012, **35**:75–78.
- Tadrous PJ: On the concept of objectivity in digital image analysis in pathology. *Pathology* 2010, **42**:207–211.
- Riber-Hansen R, Vainer B, Steiniche T: Digital image analysis: a review of reproducibility, stability and basic requirements for optimal results. *APMIS* 2012, **120**:276–289.
- Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thurlimann B, Senn HJ: Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann Oncol* 2011, **22**:1736–1747.
- Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, Ellis M, Henry NL, Hugh JC, Lively T, McShane L, Paik S, Penault-Llorca F, Prudkin L, Regan M, Salter J, Sotiriou C, Smith IE, Viale G, Zujewski JA, Hayes D, International Ki67 in Breast Cancer Working Group: Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst* 2011, **103**:1656–1664.
- Gudlaugsson E, Skaland I, Janssen EA, Smaalend R, Shao Z, Malpica A, Voorhorst F, Baak JP: Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology* 2012, **61**:1134–1144.
- Laurinavicius A, Laurinaviciene A, Ostapenko V, Dasevicius D, Jarmalaite S, Lazutka J: Immunohistochemistry profiles of breast ductal carcinoma: factor analysis of digital image analysis data. *Diagn Pathol* 2012, **7**:27.
- Gundersen HJ, Bendtsen TF, Korbo L, Marcussen N, Moller A, Nielsen K, Nyengaard JR, Pakkenberg B, Sorensen FB, Vesterby A, West MJ: Some new, simple and efficient stereological methods and their use in pathological research and diagnosis. *APMIS* 1988, **96**:379–394.
- Baddeley A, Jensen EBV: *Stereology for Statisticians*. Boca Raton, FL, USA: Chapman & Hall/CRC; 2005.
- Kieu K, Mora M: Precision of stereological planar area predictors. *J Microsc* 2006, **222**:201–211.
- Kieu K, Mora M: Advances on the precision of several stereological volume estimators. In *Ecs10: The 10th European Congress of Stereology and*

- Image Analysis: June 22-26, 2009*. Edited by V Capasso et al. Bologna, Italy: The MIRIAM Project Series, ESCULAPIO Pub. Co.; 2009:17–26.
18. Bland JM, Altman DG: **Measuring agreement in method comparison studies.** *Stat Methods Med Res* 1999, **8**:135–160.
 19. Krouwer JS: *Method Comparison and Bias Estimation Using Patient Samples: Approved Guidelines*. 2nd edition. Wayne, PA, USA: Clinical and Laboratory Standards Institute; 2010.
 20. Rimm DL, Giltman JM, Moeder C, Harigopal M, Chung GG, Camp RL, Burtness B: **Bimodal population or pathologist artifact?** *J Clin Oncol* 2007, **25**:2487–2488.
 21. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, Senn HJ: **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* 2013, **24**:2206–2223.
 22. Cuzick J, Dowsett M, Pineda S, Wale C, Salter J, Quinn E, Zabaglo L, Mallon E, Green AR, Ellis IO, Howell A, Buzdar AU, Forbes JF: **Prognostic value of a combined estrogen receptor, progesterone receptor, ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic health recurrence score in early breast cancer.** *J Clin Oncol* 2011, **29**:4273–4278.
 23. Rakha EA, Reis-Filho JS, Ellis IO: **Combinatorial biomarker expression in breast cancer.** *Breast Cancer Res Treat* 2010, **120**:293–308.
 24. Varga Z, Diebold J, Dommann-Scherrer C, Frick H, Kaup D, Noske A, Obermann E, Ohlschlegel C, Padberg B, Rakozzy C, Oliver SS, Schobinger-Clement S, Schreiber-Facklam H, Singer G, Tapia C, Wagner U, Mastropasqua MG, Viale G, Lehr HA: **How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists.** *PLoS One* 2012, **7**:5.
 25. Going JJ: **Techniques of mitosis counting.** *Hum Pathol* 1993, **24**:113–114.
 26. Mohammed ZM, McMillan DC, Elsberger B, Going JJ, Orange C, Mallon E, Doughty JC, Edwards J: **Comparison of visual and automated assessment of Ki-67 proliferative activity and their impact on outcome in primary operable invasive ductal breast cancer.** *Br J Cancer* 2012, **106**:383–388.

doi:10.1186/bcr3639

Cite this article as: Laurinavicius et al.: A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. *Breast Cancer Research* 2014 **16**:R35.

Paper II

A METHODOLOGY FOR COMPREHENSIVE BREAST CANCER KI67 LABELING INDEX WITH INTRA- TUMOR HETEROGENEITY APPRAISAL BASED ON HEXAGONAL TILING OF DIGITAL IMAGE ANALYSIS DATA

Plancoulaine, B, Laurinaviciene, A, Herlin, P, Besusparis, J,
Meskauskas, R, Baltrusaityte, I, Iqbal, Y, Laurinavicius, A.,

Virchows Archiv, 2015. 467(6): p. 711–722.

A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data

Benoit Plancoulaine¹ · Aida Laurinaviciene^{2,3} · Paulette Herlin² ·
Justinas Besusparis^{2,3} · Raimundas Meskauskas^{2,3} · Indra Baltrusaityte^{2,3} ·
Yasir Iqbal² · Arvydas Laurinavicius^{2,3}

Received: 13 June 2015 / Revised: 28 September 2015 / Accepted: 5 October 2015 / Published online: 19 October 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Digital image analysis (DIA) enables higher accuracy, reproducibility, and capacity to enumerate cell populations by immunohistochemistry; however, the most unique benefits may be obtained by evaluating the spatial distribution and intra-tissue variance of markers. The proliferative activity of breast cancer tissue, estimated by the Ki67 labeling index (Ki67 LI), is a prognostic and predictive biomarker requiring robust measurement methodologies. We performed DIA on whole-slide images (WSI) of 302 surgically removed Ki67-stained breast cancer specimens; the tumour classifier algorithm was used to automatically detect tumour tissue but was not trained to distinguish between invasive and non-invasive carcinoma cells. The WSI DIA-generated data were subsampled by hexagonal tiling (HexT). Distribution and texture parameters were compared to conventional WSI DIA and pathology report data. Factor analysis of the data set, including total numbers of tumor cells, the Ki67 LI and Ki67

distribution, and texture indicators, extracted 4 factors, identified as entropy, proliferation, bimodality, and cellularity. The factor scores were further utilized in cluster analysis, outlining subcategories of heterogeneous tumors with predominant entropy, bimodality, or both at different levels of proliferative activity. The methodology also allowed the visualization of Ki67 LI heterogeneity in tumors and the automated detection and quantitative evaluation of Ki67 hotspots, based on the upper quintile of the HexT data, conceptualized as the “Pareto hotspot”. We conclude that systematic subsampling of DIA-generated data into HexT enables comprehensive Ki67 LI analysis that reflects aspects of intra-tumor heterogeneity and may serve as a methodology to improve digital immunohistochemistry in general.

Keywords Breast cancer · Immunohistochemistry · Digital pathology · Automated image analysis · Ki67 · Heterogeneity

Electronic supplementary material The online version of this article (doi:10.1007/s00428-015-1865-x) contains supplementary material, which is available to authorized users.

✉ Arvydas Laurinavicius
arvydas.laurinavicius@vpc.lt

Benoit Plancoulaine
benoit.plancoulaine@orange.fr

Aida Laurinaviciene
aida.laurinaviciene@vpc.lt

Paulette Herlin
DanHerlin@aol.com

Justinas Besusparis
justinas.besusparis@vpc.lt

Raimundas Meskauskas
raimundas.meskauskas@vpc.lt

Indra Baltrusaityte
indra.baltrusaityte@vpc.lt

Yasir Iqbal
yasirms@gmail.com

¹ Path-Image/BioTiCla, University Caen Normandy, Caen, France

² Department of Pathology, Forensic Medicine and Pharmacology, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

³ National Center of Pathology, Affiliate of Vilnius University Hospital Santariskiu Clinics, P. Baublio 5, 08406 Vilnius, Lithuania

Paper III

IMPACT OF TISSUE SAMPLING ON ACCURACY OF KI67 IMMUNOHISTOCHEMISTRY EVALUATION IN BREAST CANCER

Besusparris, J, Plancoulaine, B, Rasmusson, A, Augulis, R, Green, A.
R, Ellis, I. O, Laurinaviciene, A, Herlin, P, Laurinavicius, A.

Diagnostic Pathology, 2016. 11(1): p. 82.

RESEARCH

Open Access



Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer

Justinas Besusparis^{1,2*}, Benoit Plancoulaine³, Allan Rasmusson², Renaldas Augulis^{1,2}, Andrew R. Green⁴, Ian O. Ellis^{4,5}, Aida Laurinaviciene^{1,2}, Paulette Herlin¹ and Arvydas Laurinavicius^{1,2}

Abstract

Background: Gene expression studies have identified molecular subtypes of breast cancer with implications to chemotherapy recommendations. For distinction of these types, a combination of immunohistochemistry (IHC) markers, including proliferative activity of tumor cells, estimated by Ki67 labeling index is used. Clinical studies are frequently based on IHC performed on tissue microarrays (TMA) with variable tissue sampling. This raises the need for evidence-based sampling criteria for individual IHC biomarker studies. We present a novel tissue sampling simulation model and demonstrate its application on Ki67 assessment in breast cancer tissue taking intratumoral heterogeneity into account.

Methods: Whole slide images (WSI) of 297 breast cancer sections, immunohistochemically stained for Ki67, were subjected to digital image analysis (DIA). Percentage of tumor cells stained for Ki67 was computed for hexagonal tiles super-imposed on the WSI. From this, intratumoral Ki67 heterogeneity indicators (Haralick's entropy values) were extracted and used to dichotomize the tumors into homogeneous and heterogeneous subsets. Simulations with random selection of hexagons, equivalent to 0.75 mm circular diameter TMA cores, were performed. The tissue sampling requirements were investigated in relation to tumor heterogeneity using linear regression and extended error analysis.

Results: The sampling requirements were dependent on the heterogeneity of the biomarker expression. To achieve a coefficient error of 10 %, 5–6 cores were needed for homogeneous cases, 11–12 cores for heterogeneous cases; in mixed tumor population 8 TMA cores were required. Similarly, to achieve the same accuracy, approximately 4,000 nuclei must be counted when the intratumor heterogeneity is mixed/unknown. Tumors of low proliferative activity would require larger sampling (10–12 TMA cores, or 6,250 nuclei) to achieve the same error measurement results as for highly proliferative tumors.

Conclusions: Our data show that optimal tissue sampling for IHC biomarker evaluation is dependent on the heterogeneity of the tissue under study and needs to be determined on a per use basis. We propose a method that can be applied to determine the sampling strategy for specific biomarkers, tissues and study targets. In addition, our findings highlight the benefit of high-capacity computer-based IHC measurement techniques to improve accuracy of the testing.

Keywords: Tissue microarrays, TMA, Digital image analysis, Breast cancer, Ki67, Tumor heterogeneity, Tissue sampling

* Correspondence: justinas.besusparis@vpc.lt

¹Faculty of Medicine, Vilnius University, M.K.Ciurlionio 21, Vilnius LT-03101, Lithuania

²National Center of Pathology, affiliate of Vilnius University Hospital Santariskiu Clinics, P. Baublio 5, Vilnius LT-08406, Lithuania

Full list of author information is available at the end of the article



© 2016 The Author(s). **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

Gene expression studies have identified distinct molecular subtypes of breast cancer (Luminal A, Luminal B, HER2-enriched, basal-like and normal breast-like) with markedly different behavior and prognosis [1]. Meanwhile, clinical practice of decision making largely relies on the definition of Luminal A-like and Luminal B-like disease, based on a combination of estrogen receptor (ER), progesterone receptor (PgR) and Ki67 immunohistochemistry (IHC) [2]. Proliferative activity of tumor cells, estimated by Ki67 labeling index (Ki67 LI) is a key indicator to support this stratification and provides strong prognostic and predictive information on response to chemotherapy [3]. Clinical utility of Ki67 LI is hampered by the lack of robust measurement methodologies and widely acknowledged issue of intratumor Ki67 heterogeneity expression. Consequently, it is hard to achieve consensus on cut-off values to stratify the patients for therapeutic decisions [2]. Great effort has been made to standardize the techniques for manual and digital/automated Ki67 LI measurement, including criteria for tissue sampling, hotspot detection, and digital image analysis (DIA) tools [4–11].

Recently, Ki67 expression across distinct categories of breast cancer specimens including whole slide surgical specimens, needle core biopsies and tissue microarrays (TMA) was investigated by Knutsvik et al. [1]. They found significant differences of Ki67 LI estimates across the different sample categories and suggested that specimen-specific cut-off values should be applied for practical use. While the recommendation is logical and may compensate for the inherent differences of the tissue sampling, its implementation requires better knowledge of measurement accuracy that can be achieved by the techniques, in general. Additionally, Going [12] has previously pointed out that the counting rules depend on level of mitotic activity in tumors. This dependency has not been investigated for tumors with varying Ki67 proliferation rates.

TMA has been often applied for discovery and clinical studies of IHC biomarkers. Initially proposed by Battifora [13], it enables multiple testing on numerous tissue samples in a standardized, tissue-sparing, and high-throughput manner by assembling small core biopsies from morphologically representative areas of tissues onto a single paraffin block [14]. The approach was further refined into a precise technique by Kononen [15]. One inherent drawback of the TMA technique is related to the limited fraction of the original sample included, raising the need to achieve/be aware of adequate sampling requirements [16]. Furthermore, TMA sampling requirements may vary depending on the target, lesion, tissue, and the goal of investigation. Therefore, it is important to determine the sampling parameters on a per-use basis.

For instance, three cores of 0.6 mm diameter will have almost a similar area to one core of 1 mm diameter (0.85 mm² versus 0.78 mm²), but provide different information about the specimen as they are likely to represent multiple areas [17, 18]. To address this issue, many studies have been performed to determine the impact of size and number of TMA cores [17, 19–28]. Most commonly, the recommended number of TMA cores varied from one to four with a diameter between 0.6 mm to 2 mm.

Determining optimal TMA sampling parameters by physical sampling of the cores, is not only time-consuming, but, more importantly, it limits the options of comprehensive statistical modeling and decomposes the original tissue sample to be used as the reference standard. To overcome these limitations, the concept of a virtual TMA was explored by utilizing digital whole slide images (WSI) to extract artificial TMA cores [25]. The approach has been applied in several studies: Quintayo et al. [19] manually marked core positions on a low magnification image before acquiring images of the TMA cores at high magnification; they also matched core positions between H&E and IHC staining of the same tissue before the acquired cores were subsequently assembled to a virtual TMA of ductal carcinoma *in situ*. Pedersen et al. [28] reported a similar procedure, but used random sampling of six 1 mm diameter cores directly on 20x magnification images of both H&E and IHC slides before assembling the virtual TMA. The studies supported the principle that assembling a set of virtual TMAs by copying cores from digital images is a valuable approach in TMA-based tissue sampling modeling.

A methodology for comprehensive IHC evaluation with appraisal of intratumoral heterogeneity aspects in WSIs of Ki67-stained breast cancer tissue was recently proposed [29]. It is based on systematic subsampling of DIA-generated data into a hexagonal tiling (HexT) arrays and enables computation of a comprehensive set of texture and distribution indicators for Ki67 intratumoral variability. While the primary aim of that study was to investigate intratumoral heterogeneity of Ki67 expression, in the current study we exploit the method for modeling tissue sampling precision in homogeneous and heterogeneous tumors dichotomized by spatial entropy of Ki67 expression: The hexagons in the HexT were chosen to simulate virtual TMA cores (or corresponding fields of view in conventional microscopy), with numbers of Ki67 positive and negative cells established by DIA. Using the spatial entropy extracted from the tiling as a spatial modeling of the Ki67 expression the impact tissue and cell sample size and tumor heterogeneity has on the accuracy of Ki67 LI measurement becomes possible to investigate. We present evidence that tumors with lower Ki67 LI as well as higher spatial heterogeneity of Ki67 expression require relatively larger sampling subsets to represent the global average of

the biomarker expression in the tissue. Additionally, the results support the notion, that tumors at the low end of proliferation scale require higher cell counts [12].

Methods

Tissue and data

A data set consisting of primary breast cancer from 297 patients was used in this study. Details of the dataset are reported in [29]. Briefly, 91 % of the tumors were invasive ductal carcinoma of the breast (270/297). Tissue samples were formalin-fixed, processed with standard paraffin embedding techniques. IHC for Ki67 was performed with antibody (clone MIB-1; DAKO, Glostrup, DK) and multimer technology-based detection system (ultraView Universal DAB, Ventana, Tucson, AZ, USA). Digital WSIs were recorded using a ScanScope XT Slide Scanner (Leica Aperio Technologies, Vista, CA, USA) under 20x objective magnification (0.5- μm resolution) and subsequently subjected to DIA by the Leica Aperio Genie Classifier v.1/Nuclear v.9 algorithm. This tool was previously calibrated based on tumor versus benign tissue recognition and positive versus negative cells detection. DIA algorithm was previously validated using a criterion standard achieved by stereological counting. The research was approved by the Vilnius Regional Biomedical Research Ethics committee (reference number NR.:40, date 2007-04-26). Additional informed consent was not required for the use of archived material.

TMA simulation using hexagonal tiling

The HexT methodology forming the basis of automated texture feature extraction is described in detail in [29]. Briefly, the coordinates of positive and negative nuclei extracted by DIA were distributed into a dense HexT overlaid on each WSI. The HexT was randomly positioned within the invasive tumor area (Fig. 1, Middle). Hexagons containing no nuclear profiles by DIA were regarded as missing data; hexagons containing fewer than 100 nuclear profiles were regarded as insufficiently

sampled. A minimum requirement of 30 informative hexagons per tumor was applied. Local Ki67 LI was calculated for each hexagon to construct co-occurrence matrix used to compute Haralick texture parameters.

The individual hexagons, with local Ki67 LI, were subsequently used as TMA cores for the random sampling simulations (Fig. 1, Right) and resembled approximately a TMA core of 0.75 mm circular diameter and 0.44 mm² area. The tumors were dichotomized into homogeneous and heterogeneous groups based on the median entropy value obtained by the HexT methodology. The sampling simulations were carried out for all three tumor classes: all/mixed, homogeneous and heterogeneous.

In addition to giving insight about the minimum number of required TMA cores, the simulations can be used to infer error measurements according to how many nuclei are assessed. By dichotomizing the simulated cores by the number of nuclei contained, the error measurement can additionally be investigated as function of the nuclei count.

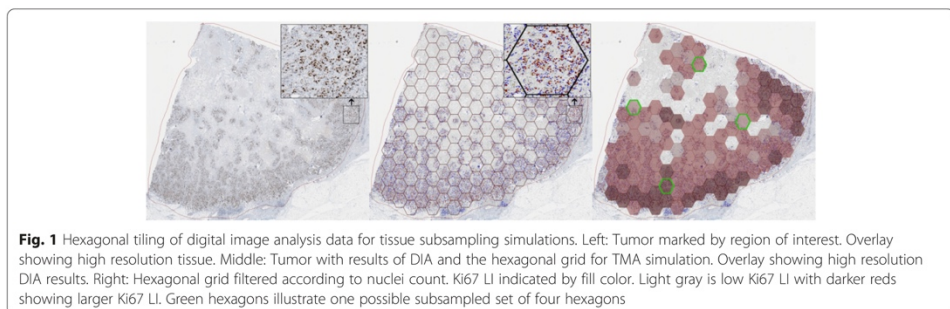
The experimental models and statistical methods

The impact caused by varying core number was investigated for a range of numbers feasible to punch out in practice. The chosen set of core numbers investigated is denoted HexN = (1, 2, ..., 15).

The practical evaluation of Ki67 LI scores from multiple cores or tissue regions is not always based on individual cell counts. Here we investigated the impact of three ways of calculating the Ki67 LI from a set of subsampled virtual cores: mean, median and by first summing total numbers of positive and negative nuclei in the subsampled hexagons, denoted sum. For a subsampled set H the Ki67 LI by sum is simply:

$$\text{sum}(H) = \frac{\sum_{i=1}^{\text{HexN}} \text{Pos}(\text{hex}_i)}{\sum_{i=1}^{\text{HexN}} \text{Pos}(\text{hex}_i) + \sum_{i=1}^{\text{HexN}} \text{Neg}(\text{hex}_i)},$$

where Pos and Neg are functions counting positive and negative nuclei in a hexagon, respectively. Note that if



TMA cores could sample the entire area of the tumors, only the evaluation by “sum” would be equivalent to the Ki67 LI determined by whole slide image analysis which extracts all nuclei before calculating Ki67 ratio.

Two different methods were used to simulate the impact of the number of hexagons/TMA cores on the precision of the sampling to represent the Ki67 LI reported by the DIA of the entire region of interest (ROI). First, the practice of “physical” TMA construction, in which a set of cores is sampled only once, was simulated by randomly sampling a subset of hexagons once. Single linear regression analysis was used to compare the data in a single random selection.

Secondly, an error analysis was conducted by simulating many samplings of TMA subsets with core numbers of sizes HexN = (1, 2, ..., 15) per case. Each subset is sampled from the set of hexagonal tiles without replacement, but all hexagons are replaced before sampling a new subset. From the resulting sampling distribution, error measurements and other statistics can be inferred. Here, the simulations were used to infer the coefficient of error (CE) of Ki67 LI predictions using subsets differing in the number of virtual cores. The CE was calculated as

$$CE = \sqrt{\frac{\text{Bias}^2 + \sigma^2}{T^2}} = \sqrt{\frac{(\mu - T)^2 + \sigma^2}{T^2}},$$

where σ is standard deviation, μ is mean of Ki67 LI inferred from the simulation distributions and T is the Ki67 LI as determined by the DIA. The interpretations of error analysis results are made according to a putative CE value of 10 % for accessible results for practical applications. The choice of this value is strongly influenced by CE dependence on Ki67 LI heterogeneity levels (Haralick entropy values). This dependence is illustrated in additional plots available as Additional file 1.

Both experiments were grouped by tumor heterogeneity and repeated for HexN = (1,2, ...,15) with hexagons resembling a 0.75 mm diameter TMA core and the simulations were performed with 50,000 iterations.

From the simulations error measurements according to how many nuclei are assessed can be inferred as follows: for one tumor case 50,000 subsets of TMA cores are sampled of size HexN = (1, 2, ..., 15). This yields a total of 750,000 subsets which are effectively grouped by HexN. By dichotomizing according to the number of nuclei sampled in each subset into bins of 250 nuclei (first bin [0;250), second bin [250;500) etc.), the error measurement can additionally be investigated as function of the nuclei count. To make it clear if CE is calculated according to hexagon area or nuclei number, the CE is denoted CE_{Area} and CE_{Nuclei} respectively.

Previously, Going [12] pointed out that to achieve the same relative error large cell counts are required for low

mitotic activity tumors while high mitotic activity requires more moderate cell counts. Specifically, it was illustrated that the relationship between the relative error in the mitotic activity can be approximated by $Relative\ Error \approx \frac{1}{\sqrt{n}} = n^{-0.5}$, where n is number of mitoses. Here we investigate if a similar relationship exists between relative error measurements CE_{Area} and CE_{Nuclei} as function of the Ki67 proliferation activity indicator by fitting CE as function of Ki67 to

$$CE = a x^{-b}.$$

This is done for each choice of HexN, for a set of bins used for dichotomizing by nuclei count and for all three classes of heterogeneity (all/mixed, homogeneous and heterogeneous).

Statistical analysis was performed using R 3.1.2, GNU GCC 5.2.1, Open Office 4.1.2 and SAS 9.4 software.

Results

Summary statistics

Extensive dataset summary statistics of the Ki67 indicators, obtained by HexT methodology, are previously reported [29]. Briefly, the global average of Ki67 LI values (in percentages) estimated by DIA of the WSIs was almost identical to the results obtained by HexT (mean: $32.5 \pm 16.9\%$, median $32.6 \pm 17.4\%$ and sum $32.7 \pm 17.3\%$). Importantly, the HexT data provided a comprehensive set of intratissue variation indicators [29].

Single subsampling – linear regression analysis

Figure 2 illustrates the linear regression analysis results (R^2 values) plotted by different types of Ki67 LI calculation methods (mean, median, and sum) and grouped by heterogeneity. All R^2 values from linear regression analysis were at $p < 0.0001$ significance level. The R^2 values for all cores are presented in Table 1. Linear regression analysis (Fig. 2) reveals that R^2 values plotted for various Ki67 LI measurement methods were nearly overlapping in the subgroup of homogeneous tumors. For the heterogeneous tumors, mean and median were less representative than the sum-based percentage. This bias was mostly apparent with small sets of cores and diminished when a larger number of cores were used.

To achieve $R^2 = 0.95$ value in the regression models, random selection of at least four, three and twelve cores were required in the mixed, homogeneous and heterogeneous tumors, respectively.

Error analysis

The mean coefficient of error (CE) for Ki67 LI estimates, calculated using the sum, is plotted for increasing TMA core numbers in the tumor subgroups (Fig. 3). To achieve the CE of 10 %, 8 cores of 0.75 mm diameter were

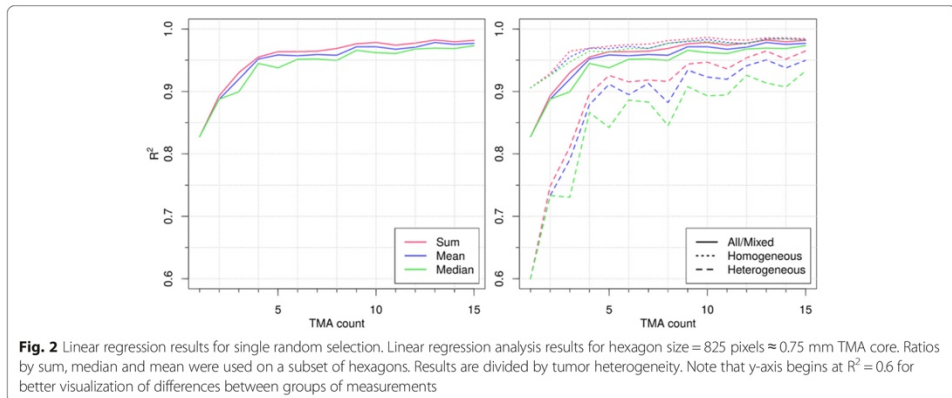


Fig. 2 Linear regression results for single random selection. Linear regression analysis results for hexagon size = 825 pixels ≈ 0.75 mm TMA core. Ratios by sum, median and mean were used on a subset of hexagons. Results are divided by tumor heterogeneity. Note that y-axis begins at $R^2 = 0.6$ for better visualization of differences between groups of measurements

required in the mixed group of tumors. Respectively, 5–6 or 11–12 cores were required in the subgroups of homogeneous and heterogeneous tumors.

To achieve a CE of 10 %, approximately 4,000 nuclei were required in the mixed group of tumors as depicted in Fig. 4. For the subgroups of homogeneous and heterogeneous tumors to reach the same error, 3,000 and 7,000 nuclei were necessary, respectively.

An inverse relationship between CE_Area and proliferation activity is clearly seen in Fig. 5 for any choice of TMA count. Furthermore, Table 2 reveals that the fitted

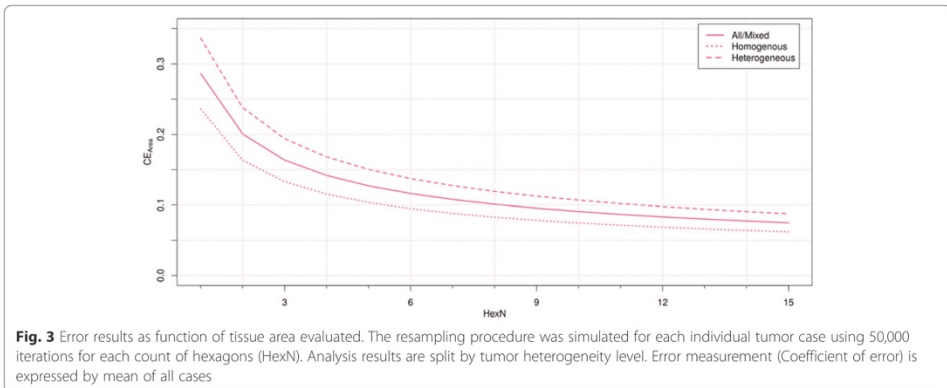
parameter b is close to the value of 0.5 as reported in [12] confirming the same dependence. The close-up around the critical point of 10 % CE and 20 % Ki67 LI in Fig. 5 shows that for mixed tumor population to achieve the CE of 10 %, approximately 10–11 TMA cores were required at the level of 20 % Ki67 LI. Respectively, 7–8 or 13–14 cores were required in the subgroups of homogeneous and heterogeneous tumors.

Similarly, a high CE_Nuclei is also observed for low proliferation rates as depicted in Fig. 6, which graphically confirms the need for counting more nuclei for low

Table 1 Linear regression analysis results for hexagon size = 825 pixels (≈0.75 mm TMA core)

HexN	All tumor cases			Homogeneous cases			Heterogeneous cases		
	Sum	Mean	Median	Sum	Mean	Median	Sum	Mean	Median
1	0.827	0.827	0.827	0.906	0.906	0.906	0.6	0.6	0.6
2	0.893	0.888	0.888	0.929	0.926	0.926	0.749	0.733	0.733
3	0.93	0.92	0.9	0.964	0.955	0.947	0.81	0.79	0.731
4	0.955	0.952	0.945	0.969	0.969	0.965	0.897	0.879	0.866
5	0.964	0.958	0.938	0.972	0.969	0.963	0.926	0.912	0.842
6	0.964	0.957	0.951	0.975	0.972	0.969	0.916	0.895	0.886
7	0.964	0.959	0.952	0.976	0.969	0.969	0.918	0.913	0.883
8	0.969	0.958	0.95	0.981	0.977	0.977	0.916	0.882	0.845
9	0.976	0.971	0.966	0.984	0.98	0.981	0.944	0.934	0.908
10	0.978	0.971	0.962	0.987	0.983	0.98	0.947	0.923	0.893
11	0.974	0.968	0.961	0.983	0.978	0.977	0.936	0.92	0.894
12	0.977	0.971	0.968	0.982	0.976	0.977	0.954	0.941	0.926
13	0.982	0.978	0.969	0.986	0.984	0.984	0.965	0.95	0.914
14	0.98	0.975	0.969	0.986	0.984	0.984	0.952	0.938	0.907
15	0.982	0.977	0.973	0.984	0.983	0.983	0.965	0.95	0.933

In each data set Ki-67 LI was calculated by counting mean, median and sum of positive and negative cells. All linear regression analysis results were statistically significant, $p < 0.0001$



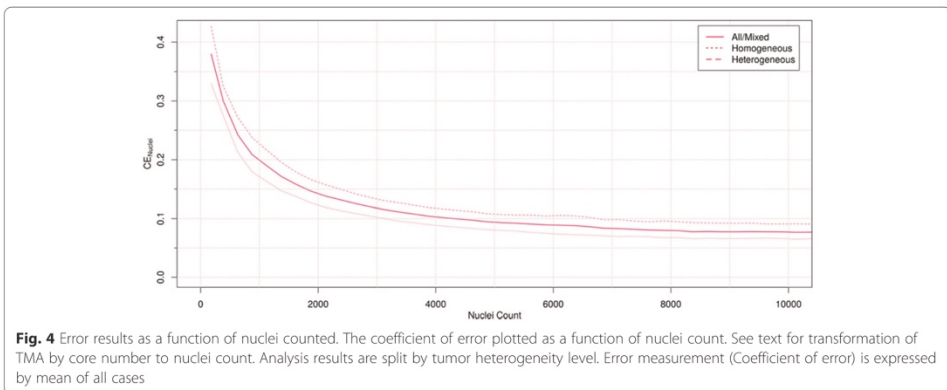
proliferation tumors. Also here the fitting parameter b is close to the value 0.5 reported [12], see Table 3. The close-ups in Fig. 6 reveal that at the level of 20 % Ki67 LI, to achieve the CE of 10 %, approximately 6,250 nuclei were required in the mixed group of tumors. For the subgroups of homogeneous and heterogeneous tumors to reach the same error 5,000 and 10,000 nuclei were necessary, respectively.

Discussion

This study has exploited novel opportunities that digital microscopy images offers for virtual TMA modeling with incorporation of DIA results. Firstly, the virtual TMAs were modeled after the HexT methodology extracted both global texture information and local feature information from the WSI. Secondly, simulation of the TMA cores using the HexT dataset enabled multiple random sampling iterations bypassing the digital

assembly of the virtual TMAs. This gave a much greater flexibility in investigating a wider range of sampling methodologies, parameters and error measurements. The added benefits do not impose any new limitations: if cores are needed for several stainings of the same tissue, cores can be sampled at the exact location in different images by applying mapping techniques similar to the ones reported by Quintayo et al. [19].

Previously, a similar approach was tested by Heus et al. [30], who utilized a dense grid of rectangular frames instead of hexagons. From each subsampled frame, a core was simulated by the largest circle contained within. This has a side-effect that tissue located at the corners of the frames will never be sampled; this effect is not independent of size of the simulated cores. The use of hexagons for virtual core simulation does not suffer from this: the dense HexT ensures that all parts of the tissue are considered with the same probability. Sampling without replacement



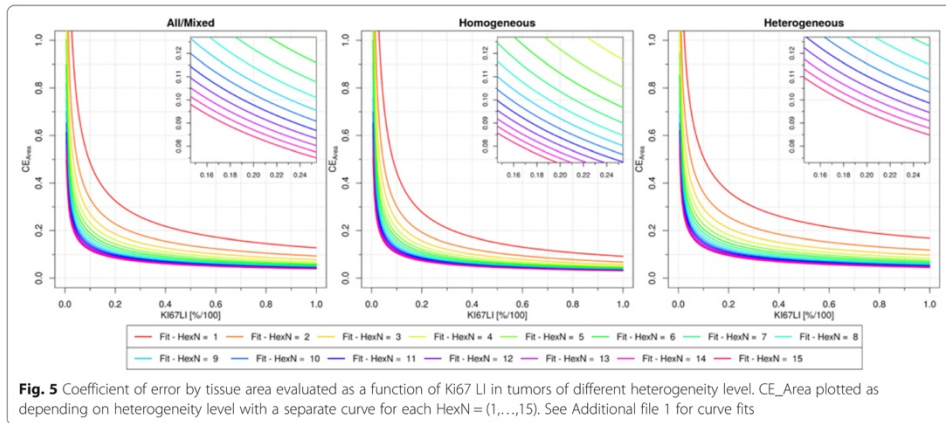


Fig. 5 Coefficient of error by tissue area evaluated as a function of Ki67 LI in tumors of different heterogeneity level. CE_Area plotted as depending on heterogeneity level with a separate curve for each HexN = (1,...,15). See Additional file 1 for curve fits

further ensures that the same area is not represented by multiple cores in the subsets used in the simulations.

The analysis of core/cell sampling requirements in this study was made possible to group according to Haralick entropy texture feature extracted by the HexT methodology for each WSI. It must be noted that the Haralick entropy threshold value is not clearly defined. Therefore, the optimal method to split the dataset it into equal parts by median was chosen. In a similar study, the variance of the local Ki67 LI was used as entropy measurement, but without a complete error analysis for the entire dataset [30].

Table 2 Fit parameters for relative error CE_Area fitted to proliferation index for all three heterogeneity classes

HexN	All/Mixed		Homogenous		Heterogeneous	
	a	b	a	b	a	b
1	0.128	0.58	0.092	0.684	0.168	0.481
2	0.093	0.553	0.068	0.647	0.119	0.482
3	0.077	0.542	0.057	0.63	0.097	0.48
4	0.068	0.533	0.051	0.614	0.084	0.478
5	0.062	0.527	0.046	0.604	0.076	0.477
6	0.057	0.521	0.043	0.595	0.069	0.474
7	0.053	0.516	0.04	0.587	0.064	0.473
8	0.05	0.512	0.038	0.579	0.06	0.473
9	0.048	0.507	0.037	0.572	0.057	0.471
10	0.046	0.502	0.035	0.564	0.054	0.469
11	0.044	0.498	0.034	0.557	0.052	0.469
12	0.042	0.493	0.033	0.549	0.05	0.467
13	0.041	0.489	0.032	0.543	0.048	0.463
14	0.04	0.486	0.031	0.538	0.047	0.463
15	0.039	0.482	0.031	0.531	0.045	0.464

Combining Ki67 LI (or any other biomarker) from several cores is often needed in TMA studies. This introduces a risk of bias which involves assessing the number of positive and negative nuclei for the observed cores before recalculating the Ki67 LI. We evaluated this potential bias by comparing results from combined Ki67 LI from a set of simulated cores using the mean, the median and Ki67 LI calculated by using sum of nuclei in the sampled hexagons. We found that when larger set of cores were used, any bias with regard to the Ki67 LI calculation methods was negligible (Table 1), while calculation of the combined Ki67 LI, by assessing the core data first, is strongly advised where only a few TMA cores are used from heterogeneous tumors (Fig. 2, right).

The practical TMA construction, where cores were randomly chosen only once, was investigated using linear regression. This allowed comparison of the hexagonal simulation data to previous studies. For a tumor set with mixed heterogeneity, we found a number of cores to achieve $R^2 = 0.95$, to be four, in line with the previous reports [17, 25, 26, 31]. For homogeneous tumors, the optimal number of cores was three, depending whether the sum or mean calculations were used, respectively. A rather dramatic increase to the requirement of 12 cores was found in the heterogeneous tumors.

The single sampling brings some eventuality, because for each sampling, it is possible to obtain cores containing different tissue representation and thus biomarker expression level. The error is particularly important when considering tissue samples with varying degrees of heterogeneity, as it influences the representativeness of TMAs [27]. A number of studies have shown that more cores will improve the agreement level and reduce the limitations due to the heterogeneity in various types of tumors and IHC biomarkers [25–27, 32, 33]. However,

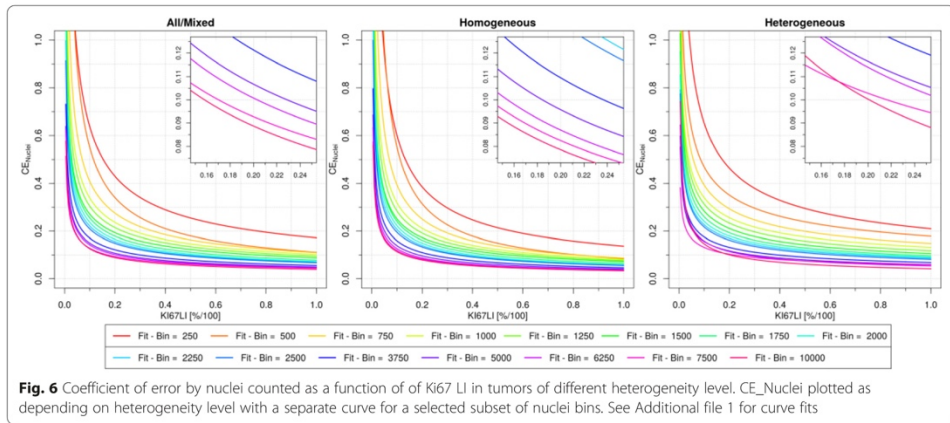


Fig. 6 Coefficient of error by nuclei counted as a function of Ki67 LI in tumors of different heterogeneity level. CE_{Nuclei} plotted as depending on heterogeneity level with a separate curve for a selected subset of nuclei bins. See Additional file 1 for curve fits

only the introduced method allows inference about the relative error caused by different TMA sampling parameters in combination with tumor heterogeneity. Our results show that to obtain a CE of 10 % the necessary number of cores in the dataset with mixed heterogeneity is 8; five cores hold sufficient information for Ki67 LI determination in homogeneous tumors, while heterogeneous tumors need at least 11–12 cores to be sampled.

In practical TMA applications, intratumoral heterogeneity of biomarker expression is usually unknown in advance; therefore, a more conservative approach would

assume that all tumors in the study population are heterogeneous. On the other hand, Ki67 LI expression in breast cancer tissue is known for its spatial heterogeneity and may serve a reference standard for other biomarkers and tumors. In that sense, our study reveals that 11–12 random TMA cores of 0.75 mm diameter would sufficiently represent IHC biomarker expression in heterogeneous tumors. Our simulations also indicate that disagreements between different studies of TMA core numbers may in fact be due to unestablished differences in heterogeneity aspects. In general, our findings support the notion that heterogeneity information is crucial for optimizing TMA studies. Ideally, the presented method could be used in pilot studies to validate the optimal number of cores, or at least heterogeneity should be investigated from a larger set of cores, for instance by measuring a range of Ki67 LI between several TMA cores taken from the tissue.

Our study also provides evidence for minimum cell counting requirements to achieve robust Ki67 LI measurement, especially with regard to the limited capacity of manual counting procedures. Current clinical guidelines on the minimal number of cells to be counted are quite arbitrary, mostly set in the range of 500 and 2000 tumor cells [9]. While small samples (e.g., needle core biopsies) may allow counting all the invasive tumor cells, it becomes impractical in larger samples. Therefore, to achieve adequate precision, it is recommended for the interpreting pathologist to score at least 1,000 cells, while 500 cells would be acceptable as the absolute minimum [9]. Importantly, our findings reveal that to achieve 10 % CE approximately 4,000 nuclei must be counted when the intratumor heterogeneity is mixed/unknown (Fig. 4). These cell counts are rather large to accomplish in clinical practice for all breast carcinomas, but could be feasible for cases considered as

Table 3 Fit parameters for relative error CE_{Nuclei} fitted to proliferation index for all three heterogeneity classes

Nuclei bin	Proliferation fit to relative error					
	All/Mixed		Homogenous		Heterogeneous	
	a	b	a	b	a	b
250	0.171	0.571	0.137	0.644	0.209	0.494
500	0.111	0.697	0.085	0.815	0.178	0.421
750	0.109	0.573	0.087	0.658	0.148	0.425
1000	0.102	0.522	0.081	0.595	0.132	0.412
1250	0.094	0.509	0.076	0.572	0.117	0.426
1500	0.087	0.491	0.072	0.544	0.105	0.43
1750	0.078	0.516	0.064	0.575	0.096	0.434
2000	0.072	0.516	0.058	0.58	0.09	0.425
2250	0.069	0.504	0.056	0.567	0.086	0.419
2500	0.067	0.493	0.055	0.547	0.081	0.426
3750	0.055	0.487	0.046	0.538	0.068	0.411
5000	0.049	0.485	0.041	0.534	0.059	0.422
6250	0.046	0.494	0.037	0.539	0.054	0.47
7500	0.044	0.464	0.035	0.526	0.058	0.356
10000	0.039	0.507	0.033	0.532	0.042	0.543

“grey zone”, e.g. in the range of Ki67 LI 10-30 % [3]. A visual scoring methodology proposed by Hida et al., might be used as method of choice for “low” (Ki67 LI <10 %) or “high” (Ki67 LI >30 %) proliferatively active cases, leaving behind “grey zone” cases, which requires more precise methodologies [34].

The inverse relationship between relative estimation error and mitotic activity previously highlighted by Going [12] was confirmed to also exist between each of the two error estimates (CE_{Area} , CE_{Nuclei}) and the Ki67 LI proliferation activity indicator (Figs. 5 and 6). This dependency of CE on Ki67 LI shows that tumor cases with low proliferation rate contribute most of the CE in Fig. 4 which is a set of “mixed” proliferation rate. Consequently, when scoring a single case with unknown Ki67 LI one may need to evaluate a higher cell count or larger TMA sample to ensure a 10 % CE at a specific grey zone. Specifically, the tumors at the lower scale of proliferative activity (Ki67 LI <20 %, Fig. 5, left) will for a mixed/unknown heterogeneity case require larger sampling (at least 10–11 TMA cores) to achieve the same error measurement (10 % CE_{Area}) results as for highly proliferative tumors (4–6 TMA cores). Similarly, for cases with Ki67 LI <20 % (Fig. 5, right) at least 6,250 nuclei are necessary (for 10 % CE_{Nuclei}). As such, Figs. 5 and 6 may aid determining practical sampling requirements of individual cases for acceptable CE at specific grey zones.

In general, the results of our study suggest that adequate accuracy levels of Ki67 LI measurement can hardly be achieved by manual counts and argue in favor of DIA-based techniques to benefit from the high-capacity methods. In addition, automated hotspot detection with standard definitions by DIA, which was out of scope in the present study, would provide another advantage compared to the visual evaluation by conventional microscopy or inspection of WSI.

Conclusion

Several aspects raised in this study relate to the evaluation of Ki67 immunohistochemistry in breast cancer in clinical research and practice. Firstly, obtaining an optimal number of TMA cores/cell number needed for biomarker research studies depends on the tissue, especially its intratissue heterogeneity and level of expression. For Ki67 LI in breast cancer, we found 5–6 cores sufficient for homogeneous expression in the tissue, 8 cores for tumors with mixed heterogeneity and at least 11 cores for heterogeneous tumors. Secondly, our findings reveal that to achieve low error estimates when evaluating by cell counting, approximately 4,000 nuclei must be evaluated when the intratumor heterogeneity is mixed/unknown. In breast cancer cases of the lower proliferative activity (Ki67 LI <20 %) larger sampling is required to achieve the same error measurement results as for highly

proliferative tumors. The presented data may aid in defining practical sampling requirements of individual cases and specific grey zones.

The wide range of the number of cores/nuclei needed supports the notion that optimal sampling requirements must be determined on a peruse basis and that heterogeneity information must be assessed in the study. The method presented can be applied for individual pilot study measurements. In addition, our findings highlight the importance of high-capacity computer-based IHC measurement techniques to improve accuracy of the testing.

Additional files

Additional file 1: Fits curves of proliferation index to CE_{Area} and CE_{Nuclei} (depending on heterogeneity levels). Graphs for CE plotted as a function of Area (Hexes in case). (DOCX 12 mb)

Additional file 2: Initial dataset of the study. (XLSX 2 mb)

Abbreviations

CE, coefficient of error; CE_{Area} , coefficient of error calculated according to hexagon area; CE_{Nuclei} , coefficient of error calculated according to nuclei number; DIA, digital image analysis; ER, estrogen receptor; H&E, hematoxylin and eosin staining; HexN, set of hexagonal cores; HexT, hexagonal tiling; IHC, immunohistochemistry; Ki67LI, Ki67 labeling index; PgR, progesterone receptor; TMA, tissue microarray; TMAs, tissue microarrays; WSI, whole slide images

Acknowledgements

None.

Funding

This research is fully funded by the European Social Fund under the Global Grant measure, Grant #VP1-3.1-SMM-07-K-03-051.

Availability of data and materials

The dataset supporting the conclusions of this article is included within the article and its Additional file 2.

Authors' contributions

BP, AR constructed and performed TMA simulation procedures and statistical analysis systems. AL, AR, BP, RA, and PH performed statistical analyses. JB, in collaboration with AR and AL, drafted essential parts of the manuscript. AG, Ail, AR, IE reviewed and edited the manuscript. All authors participated in conception and design of the study, reviewing the analysis results, critically revised and approved the final manuscript.

Authors' information

None.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The research was approved by the Vilnius Regional Biomedical Research Ethics committee (reference number NR.40, date 2007-04-26). Additional informed consent was not required for the use of archived material.

Author details

¹Faculty of Medicine, Vilnius University, M.K.Giurlionio 21, Vilnius LT-03101, Lithuania. ²National Center of Pathology, affiliate of Vilnius University Hospital Santariskiu Clinics, P. Baublio 5, Vilnius LT-08406, Lithuania. ³PathImage/BioTICLA, Inserm (UMR 1199), University Caen Normandy, Cancer Center F. Baclesse, Caen, France. ⁴Division of Cancer and Stem Cells, School of

Medicine, University of Nottingham, Nottingham, UK. ⁵Histopathology, Nottingham City Hospital University of Nottingham, Nottingham, UK.

Received: 26 February 2016 Accepted: 31 July 2016

Published online: 30 August 2016

References

- Knutsvik G, Stefansson IM, Aziz S, Arnes J, Eide J, Collett K, Akslen LA. Evaluation of Ki67 expression across distinct categories of breast cancer specimens: a population-based study of matched surgical specimens, core needle biopsies and tissue microarrays. *PLoS One*. 2014;9(11), e112121. doi:10.1371/journal.pone.0112121.
- Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, Senn HJ. 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*. 2013;24(9):2206–23. doi:10.1093/annonc/mdt303.
- Untch M, Harbeck N, Huober J, von Minckwitz G, Gerber B, Kreipe HH, Liedtke C, Marschner N, Mobus V, Scheithauer H, Schneeweiss A, Thomssen C, Jackisch C, Beckmann MW, Blohmer JU, Costa SD, Decker T, Diel I, Fasching PA, Fehm T, Janni W, Luck HJ, Maass N, Scharl A, Loibl S. Primary Therapy of Patients with Early Breast Cancer: Evidence, Controversies, Consensus: Opinions of German Specialists to the 14th St. Gallen International Breast Cancer Conference 2015 (Vienna 2015). *Geburtshilfe Frauenheilkd*. 2015;75(6):556–65. doi:10.1055/s-0035-1546120.
- Romero Q, Bendahl PO, Ferno M, Grabau D, Borgquist S. A novel model for Ki67 assessment in breast cancer. *Diagn Pathol*. 2014;9:118. doi:10.1186/1746-1596-9-118.
- Lu H, Papatomas TG, van Zessen D, Palli I, de Krijger RR, van der Spek PJ, Dinjens W, Stubbs AP. Automated Selection of Hotspots (ASH): enhanced automated segmentation and adaptive step finding for Ki67 hotspot detection in adrenal cortical cancer. *Diagn Pathol*. 2014;9(1):216. doi:10.1186/s13000-014-0216-6.
- Potts SJ, Krueger JS, Landis ND, Eberhard DA, Young GD, Schmechel SC, Lange H. Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. *Lab Invest*. 2012;92(9):1342–57. doi:10.1038/labinvest.2012.91.
- Haroske G, Dimmer V, Steindorf D, Schilling U, Theissig F, Kunze KD. Cellular sociology of proliferating tumor cells in invasive ductal breast cancer. *Anal Quant Cytol Histol*. 1996;18(3):191–8.
- Gudlaugsson E, Skaland I, Janssen EA, Smaaland R, Shao Z, Malpica A, Voorhorst F, Baak JP. Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology*. 2012;61(6):1134–44. doi:10.1111/j.1365-2559.2012.04329.x.
- Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, Ellis M, Henry NL, Hugh JC, Lively T, McShane L, Paik S, Penault-Llorca F, Prudkin L, Regan M, Salter J, Sotiriou C, Smith IE, Viale G, Zujewski JA, Hayes DF. International Ki-67 in Breast Cancer Working Group. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst*. 2011;103(22):1656–64. doi:10.1093/jnci/djr393.
- Laurinavicius A, Plancoulaine B, Laurinaviciene A, Herlin P, Meskauskas R, Baltrusaityte I, Besusparis J, Dasevicius D, Elie N, Iqbal Y, Bor C. A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. *Breast Cancer Res*. 2014;16(2):R35. doi:10.1186/bcr3639.
- Laurinavicius A, Plancoulaine B, Laurinaviciene A, Herlin P, Meskauskas R, Baltrusaityte I, Besusparis J, Elie N, Bellhomme P, Iqbal Y, Bor-Angelier C (2013) A methodology to ensure and improve accuracy of Ki67 digital immunohistochemistry analysis in breast cancer tissue. *Molecular Cancer Research* 11. doi:10.1158/1557-3125.ADVBC-B116
- Going JJ. Techniques of mitosis counting. *Hum Pathol*. 1993;24(1):113–4.
- Battifora H. The multitumor (sausage) tissue block: novel method for immunohistochemical antibody testing. *Lab Invest*. 1986;55(2):244–8.
- Pertuth-Wey J, Boulware D, Valkov N, Livingston S, Nicosia S, Lee JH, Sumpfen R, Schildkraut J, Narod S, Parker A, Coppola D, Sellers T, Pal T. Sampling Strategies for Tissue Microarrays to evaluate biomarkers in Ovarian Cancer. *Cancer Epidemiol Biomarkers Prev*. 2009;18(1):28–34.
- Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med*. 1998;4(7):844–7.
- Kallioniemi O-P, Wagner U, Kononen J, Sauter G. Tissue microarray technology for high-throughput molecular profiling of cancer. *Hum Mol Genet*. 2001;10(7):657–62. doi:10.1093/hmg/10.7.657.
- Ilyas M, Grabsch H, Ellis IO, Womack C, Brown R, Berney D, Fennell D, Salto-Tellez M, Jenkins M, Landberg G, Byers R, Treanor D, Harrison D, Green AR, Ball G, Hamilton P. Guidelines and considerations for conducting experiments using tissue microarrays. *Histopathology*. 2013; 62(6):827–39. doi:10.1111/his.12118.
- Anagnostou VK, Lowery FJ, Syrigos KN, Cagle PT, Rimm DL. Quantitative evaluation of protein expression as a function of tissue microarray core diameter: is a large (1.5 mm) core better than a small (0.6 mm) core? *Arch Pathol Lab Med*. 2010;134(4):613–9. doi:10.1043/1543-2165-134.4.613.
- Quintayo MA, Starczynski J, Yan FJ, Wedad H, Nofech-Mozes S, Rakovitch E, Bartlett JM. Virtual tissue microarrays: a novel and viable approach to optimizing tissue microarrays for biomarker research applied to ductal carcinoma *in situ*. *Histopathology*. 2013. doi:10.1111/his.12336.
- Schmidt LH, Biesterfeld S, Kummel A, Faldum A, Sebastian M, Taube C, Buhll R, Wiewrodt R. Tissue microarrays are reliable tools for the clinicopathological characterization of lung cancer tissue. *Anticancer Res*. 2009;29(1):201–9.
- Zhang D, Salto-Tellez M, Putti TC, Do E, Koay ES. Reliability of tissue microarrays in detecting protein expression and gene amplification in breast cancer. *Mod Pathol*. 2003;16(1):79–84. doi:10.1097/01.MP.0000047307.96344.93.
- Torhorst J, Bucher C, Kononen J, Haas P, Zuber M, Kochli OR, Mross F, Dieterich H, Moch H, Mihatsch M, Kallioniemi OP, Sauter G. Tissue microarrays for rapid linking of molecular changes to clinical endpoints. *Am J Pathol*. 2001;159(6):2249–56. doi:10.1016/S0002-9440(1)063075-1.
- Mucci NR, Akdas G, Manely S, Rubin MA. Neuroendocrine expression in metastatic prostate cancer: evaluation of high throughput tissue microarrays to detect heterogeneous protein expression. *Hum Pathol*. 2000;31(4):406–14. doi:10.1053/hp.2000.7295.
- Camp RL, Charette LA, Rimm DL. Validation of tissue microarray technology in breast carcinoma. *Lab Invest*. 2000;80(12):1943–9.
- Goethals L, Perneel C, Debucquoy A, De Schutter H, Borghys D, Ectors N, Geboes K, McBride WH, Haustermans KM. A new approach to the validation of tissue microarrays. *J Pathol*. 2006;208(5):607–14. doi:10.1002/path.1934.
- Gulmann C, Butler D, Kay E, Grace A, Leader M. Biopsy of a biopsy: validation of immunoprofiling in gastric cancer biopsy tissue microarrays. *Histopathology*. 2003;42(1):70–6.
- Alkushi A. Validation of tissue microarray biomarker expression of breast carcinomas in Saudi women. *Hematol Oncol Stem Cell Ther*. 2009;2(3):394–8.
- Pedersen MB, Riber-Hansen R, Nielsen PS, Bendix K, Hamilton-Dutoit SJ, D'Amore F, Steiniche T. Digital pathology for the validation of tissue microarrays in peripheral T-cell lymphomas. *Appl Immunohistochem Mol Morphol*. 2014;22(8):577–84. doi:10.1097/PAI.0b013e3182a7d16d.
- Plancoulaine B, Laurinaviciene A, Herlin P, Besusparis J, Meskauskas R, Baltrusaityte I, Iqbal Y, Laurinavicius A. A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data. *Virchows Arch*. 2015;467(6):711–22. doi:10.1007/s00428-015-1865-x.
- Heus R (2009) Approches virtuelles dédiées à la technologie des puces à tissu "Tissue MicroArrays" TMA: Application à l'étude de la transformation tumorale du tissu colorectal. PhD thesis, Life Sciences, Université Joseph-Fourier - Grenoble I French <https://tel.archives-ouvertes.fr/tel-00429056>
- Rubin MA, Dunn R, Strawderman M, Pienta KJ. Tissue microarray sampling strategy for prostate cancer biomarker analysis. *Am J Surg Pathol*. 2002; 26(3):312–9.
- Lax SF, Pizer ES, Ronnett BM, Kurman RJ. Comparison of estrogen and progesterone receptor, Ki-67, and p53 immunoreactivity in uterine endometrioid carcinoma and endometrioid carcinoma with squamous, mucinous, serous, and ciliated cell differentiation. *Hum Pathol*. 1998;29(9):924–31.
- Ruiz C, Selb S, Al KK, Siraj AK, Mirfischer M, Schraml P, Maurer R, Spichtin H, Torhorst J, Popovska S, Simon R, Sauter G. Tissue microarrays for comparing molecular features with proliferation activity in breast cancer. *Int J Cancer*. 2006;118(9):2190–4.
- Hida AI, Oshiro Y, Inoue H, Kawaguchi H, Yamashita N, Moriya T. Visual assessment of Ki67 at a glance is an easy method to exclude many luminal-type breast cancers from counting 1000 cells. *Breast Cancer*. 2015;22(2):129–34. doi:10.1007/s1282-013-0460-8.

Paper IV

BIMODALITY OF INTRATUMOR KI67 EXPRESSION IS AN INDEPENDENT PROGNOSTIC FACTOR OF OVERALL SURVIVAL IN PATIENTS WITH INVASIVE BREAST CARCINOMA

Laurinavicius, A, Plancoulaine, B, Rasmusson, A, Besusparis, J,
Augulis, R, Meskauskas, R, Herlin, P, Laurinaviciene, A, Abdelhadi
Muftah, A. A, Miligy, I, Aleskandarany, M, Rakha, E. A, Green, A.
R, Ellis, I. O.

Virchows Archiv, 2016. 468(4): p. 493–502.

Bimodality of intratumor Ki67 expression is an independent prognostic factor of overall survival in patients with invasive breast carcinoma

Arvydas Laurinavicius^{1,2} · Benoit Plancoulaine³ · Allan Rasmusson² ·
Justinas Besusparis^{1,2} · Renaldas Augulis^{1,2} · Raimundas Meskauskas² ·
Paulette Herlin¹ · Aida Laurinaviciene^{1,2} · Abir A. Abdelhadi Muftah⁴ · Islam Miligy⁴ ·
Mohammed Aleskandarany⁴ · Emad A. Rakha^{4,5} · Andrew R. Green⁴ · Ian O. Ellis^{4,5}

Received: 10 October 2015 / Revised: 15 November 2015 / Accepted: 14 January 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Proliferative activity, assessed by Ki67 immunohistochemistry (IHC), is an established prognostic and predictive biomarker of breast cancer (BC). However, it remains underutilized due to lack of standardized robust measurement methodologies and significant intratumor heterogeneity of expression. A recently proposed methodology for IHC biomarker assessment in whole slide images (WSI), based on systematic

subsampling of tissue information extracted by digital image analysis (DIA) into hexagonal tiling arrays, enables computation of a comprehensive set of Ki67 indicators, including intratumor variability. In this study, the tiling methodology was applied to assess Ki67 expression in WSI of 152 surgically removed Ki67-stained (on full-face sections) BC specimens and to test which, if any, Ki67 indicators can predict

Electronic supplementary material The online version of this article (doi:10.1007/s00428-016-1907-z) contains supplementary material, which is available to authorized users.

✉ Arvydas Laurinavicius
arvydas.laurinavicius@vpc.lt

Benoit Plancoulaine
benoit.plancoulaine@orange.fr

Allan Rasmusson
allan.rasmusson@vpc.lt

Justinas Besusparis
justinas.besusparis@vpc.lt

Renaldas Augulis
renaldas.augulis@vpc.lt

Raimundas Meskauskas
raimundas.meskauskas@vpc.lt

Paulette Herlin
DanHerlin@aol.com

Aida Laurinaviciene
aida.laurinaviciene@vpc.lt

Abir A. Abdelhadi Muftah
mrxam21@nottingham.ac.uk

Islam Miligy
msxima@nottingham.ac.uk

Mohammed Aleskandarany
mohammed.aleskandarany@nottingham.ac.uk

Emad A. Rakha
emadrakha@yahoo.com

Andrew R. Green
andrew.green@nottingham.ac.uk

Ian O. Ellis
ian.ellis@nottingham.ac.uk

¹ Department of Pathology, Forensic Medicine and Pharmacology, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

² National Center of Pathology, Affiliate of Vilnius University Hospital Santariskiu Clinics, P. Baublio 5, LT-08406 Vilnius, Lithuania

³ PathImage/BioTICLA, Inserm (UMR 1199), University Caen Normandy, Cancer Center F. Baclesse, Caen, France

⁴ Division of Cancer and Stem Cells, School of Medicine, University of Nottingham, Nottingham, UK

⁵ Histopathology, Nottingham City Hospital University of Nottingham, Nottingham, UK

Vilniaus universiteto leidykla
Universiteto g. 1, LT-01513 Vilnius
El. p. info@leidykla.vu.lt,
www.leidykla.vu.lt
Tiražas 100 egz.