Better prognostic markers for non-muscle invasive papillary urothelial carcinomas

by

Vebjørn Kvikstad

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

Faculty of Science and Technology

Department of Chemistry, Bioscience and Environmental Engineering
2022

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

©2022 Vebjørn Kvikstad

ISBN:978-82-8439-069-7
ISSN:1890-1387
PhD: Thesis UiS No. 635

Acknowledgements

First and foremost, I would like to thank Professor Emiel Janssen, my main supervisor. Emiel, I want to thank you for being positive and dedicated from the first day I announced my interest in research. Your optimism, knowledge and experience have been invaluable through this work. Your support and feedback through these years have broadened and deepened my understanding of cancer research, digital image analysis and molecular biology. You have given me not only the possibility to grow as a PhD-student, but broadened my horizons and taught me the importance of networking and connecting with others that share the same interest in bladder cancer. It is a pleasure to work with you.

I am grateful to dr. Einar Gudlaugsson, my co-supervisor, for the support and encouragement through these years. Not only have you guided me through complex questions I might have had in the research process, but also you have been an inspiration and a great mentor in diagnostic pathology.

I want to thank Melinda Lillesand for excellent cooperation on the papers, technical support and general encouragement. Without your help and positivity, this would not have been possible.

A special thanks to Kjell H. Kjellevold, former head of department of Pathology at Stavanger University Hospital. You convinced me to come to Stavanger, even if there were limited skiing possibilities, and start my training in pathology, which I shall always be thankful for. Later on, you suggested to try research, and motivated me to aim for a PhD.

I also want to thank Professor Jan Baak for invaluable help in writing the papers and general advice during the years of my research endeavours. Your experienced suggestions have been very much appreciated. I am also grateful for our inspiring talks.

I would like to thank Eliza Peixoto Albernaz, Emma Rewcastle, Bianca van Diermen-Hidle and Ivar Skaland for their contribution in this project.

Christiaan de Jong and Tahlita C. M. Zuiverloon at department of Urology, Erasmus MC Rotterdam, thank you for your cooperation and hospitality.

Thanks to Susanne Buhr-Wildhagen and the Department of Pathology at Stavanger University Hospital for giving me the opportunity to combine research with daily pathology practice.

Finally, I want to thank my wife Claudia for all her patience and support. You have always been there for me through these years. And of course my son Erik, for making me smile and focus on other things in life

Summary

Bladder cancer is a common type of cancer, especially among men in developed countries. Most cancers in the urinary bladder are papillary urothelial carcinomas. They are characterized by a high recurrence frequency (up to 70 %) after local resection. It is crucial for prognosis to discover these recurrent tumours at an early stage, especially before they become muscle-invasive. Reliable prognostic biomarkers for tumour recurrence and stage progression are lacking. This is why patients diagnosed with a non-muscle invasive bladder cancer follow extensive follow-up regimens with possible serious side effects and with high costs for the healthcare systems.

WHO grade and tumour stage are two central biomarkers currently having great impact on both treatment decisions and follow-up regimens. However, there are concerns regarding the reproducibility of WHO grading, and stage classification is challenging in small and fragmented tumour material. In Paper I, we examined the reproducibility and the prognostic value of all the individual microscopic features making up the WHO grading system. Among thirteen extracted features there was considerable variation in both reproducibility and prognostic value. The only feature being both reasonably reproducible and statistically significant prognostic was cell polarity. We concluded that further validation studies are needed on these features, and that future grading systems should be based on well-defined features with true prognostic value.

With the implementation of immunotherapy, there is increasing interest in tumour immune response and the tumour microenvironment. In a search for better prognostic biomarkers for tumour recurrence and stage progression, in Paper II, we investigated the prognostic value of tumour infiltrating immune cells (CD4, CD8, CD25 and CD138) and previously investigated cell proliferation markers (Ki-67, PPH3 and MAI). Low Ki-

67 and tumour multifocality were associated with increased recurrence risk. Recurrence risk was not affected by the composition of immune cells. For stage progression, the only prognostic immune cell marker was CD25. High values for MAI was also strongly associated with stage progression. However, in a multivariate analysis, the most prognostic feature was a combination of MAI and CD25.

BCG-instillations in the bladder are indicated in intermediate and high-risk non-muscle invasive bladder cancer patients. This old-fashion immunotherapy has proved to reduce both recurrence- and progression-risk, although it is frequently followed by unpleasant side-effects. As many as 30-50% of high-risk patients receiving BCG instillations, fail by develop high-grade recurrences. They do not only suffer from unnecessary side-effects, but will also have a delay in further treatment. Together with colleagues at three different Dutch hospitals, in Paper III, we looked at the prognostic and predictive value of T1-substaging. A T1-tumour invades the lamina propria, and we wanted to separate those with micro- from those with extensive invasion. We found that BCG-failure was more common among patients with extensive invasion. Furthermore, T1-substaging was associated with both high-grade recurrence-free and progression-free survival.

Finally, in Paper IV, we wanted to investigate the prognostic value of two classical immunohistochemical markers, p53 and CK20, and compare them with previously investigated proliferation markers. p53 is a surrogate marker for mutations in the gene *TP53*, considered to be a main characteristic for muscle-invasive tumours. CK20 is a surrogate marker for luminal tumours in the molecular classification of bladder cancer, and is frequently used to distinguish reactive urothelial changes from urothelial carcinoma in situ. We found both positivity for p53 and CK20 to be significantly associated with stage progression, although not performing better than WHO grade and stage. The proliferation marker MAI, had the highest prognostic value in our study. Any combination of

variables did not perform better in a multivariate analysis than MAI alone.

List of Publications

Paper I

Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas. Vebjørn Kvikstad*, Ok Målfrid Mangrud*, Einar Gudlaugsson, Ingvild Dalen, Hans Espeland, Jan P. A. Baak, Emiel A. M. Janssen. (2019) **BMC Diagnostic Pathology** 14:90 https://doi.org/10.1186/s13000-019-0868-3.

* Shared first authorship

Paper II

Mitotic activity index and CD25+ lymphocytes predict risk of stage progression in non-muscle invasive bladder cancer. Melinda Lillesand*, Vebjørn Kvikstad*, Ok Målfrid Mangrud, Einar Gudlaugsson, Bianca van Diermen-Hidle, Ivar Skaland, Jan P. A. Baak, Emiel A. M. Janssen. (2020) **PLoS ONE** 15(6): e0233676.

* Shared first authorship

Paper III

T1 substaging of nonmuscle invasive bladder cancer is associated with bacillus Calmette-Guerin failure and improves patient stratification at diagnosis. Florus C. de Jong, Robert F. Hoedemaeker, Vebjørn Kvikstad, Jolien T. M. Mensink, Joep J. de Jong, Egbert R. Boeve, Deric K. E. van der Schoot, Ellen C. Zwarthoff, Joost L. Boormans, Tahlita C. M. Zuiverloon. (2021) **The Journal of Urology** 205(3):701-708. doi: 10.1097/JU.0000000000001422.

Paper IV

Proliferation and immunohistochemistry for p53 and CK20 in predicting prognosis of non-muscle invasive papillary urothelial carcinomas. Vebjørn Kvikstad, Melinda Lillesand, Einar Gudlaugsson, Ok Målfrid Mangrud, Emma Rewcastle, Ivar Skaland, Jan P. A. Baak, Emiel A. M. Janssen. Manuscript.

Table of Contents

List of Figures

List of Tables

List of Appendices

Appendix 1 - *Multiclass tissue classification of whole-slide histological images using convolutional neural networks*. Wetteland R, Engan K, Eftestøl T, Kvikstad V and Janssen EAM. Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods, vol 1, pp. 320-327, 2019.

Appendix 2 – *Multiscale deep neural networks for multiclass tissue classification of histological whole-slide images*. Wetteland R, Engan K, Eftestøl T, Kvikstad V and EAM Janssen. Medical Imaging with Deep Learning (MIDL), 2019.

Appendix 3 - *A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides*. Wetteland R, Engan K, Eftestøl T, Kvikstad V and EAM Janssen. Technol. Cancer Res. Treatment, vol 19, 2020.

Appendix 4 – *Automatic diagnostic tool for predicting cancer grade in bladder cancer patients using deep learning*. Wetteland R, Kvikstad V, Eftestøl T, Tøssebro E, Lillesand M, Janssen EAM and Engan K. IEEE Access 9, pp. 109214-109223, 2021.

1 Introduction

1.1 Anatomy and physiology of the bladder

The urinary bladder is located in the pelvis, resting on the pelvic floor. It has four corners, made up by the entrance of the two ureters at the upper back, the exit from the urethra in the bottom and the apex in front where the ligamentum umbilicale medianum (a remnant of the fetal urachus) connects with the bladder. The triangular area made up of the ureteral orifices and the urethra is called trigonum.

The urinary bladder wall is composed of several well-defined layers; tunica mucosae, muscularis propria, and adventitia/serosa. The tunica mucosae includes the urothelium and the underlying lamina propria. The urothelium is a specialized epithelium, constructed to make a urine-blood barrier, and to also tolerate variable degrees of distention. The urothelium covers the urinary tract from the renal pelvis to the proximal urethra. The number of cell layers in the urothelium varies depending on location and degree of bladder distention. For the urinary bladder, it is mostly 4–6 cell layers thick. The cells lining the lumen, against the urine, are called umbrella cells. They are a bit larger than the other cells in the urothelium and have large amounts of eosinophilic cytoplasm. They are often bi-nucleated. When the bladder wall is stretched out, they are oriented along the mucosae. Under the umbrella cells, we find polygonal intermediate cells that do not have a specific orientation. At the bottom, resting on the basement membrane, 1–3 layers of cells with oval nuclei can be found; these basal cells are oriented

perpendicular to the surface.

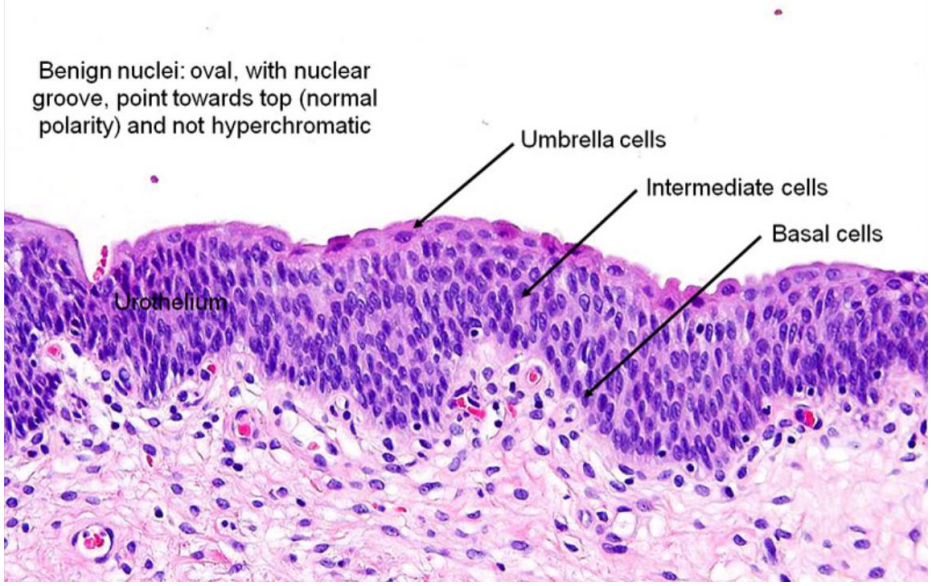

Figure 1. Normal histology of the tunica mucosae.

Image of tunica mucosae, showing the distinct layers of the urothelium. Contributed by the American Urological Association. Bolla SR, Odeluga N, Jetti R. Histology, Bladder. [Updated 2020 Apr 15]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK540963/

The lamina propria is a loose connective tissue containing blood and lymphatic vessels, and usually some discontinues bundles of smooth muscle called muscularis mucosae. The muscalaris propria, often referred to as the detrusor muscle, is a distinct layer of large smooth muscle bundles. The muscle fibres in these bundles are mostly disorganized and oriented in different directions, the exception is in the area close to the internal sphincter, where the fibres are organized in specific directions. The outermost layer is the tunica adventitia, made up of loose connective tissue. The adventitia surrounds the bladder, except at the superior surface where it is covered by serosa.

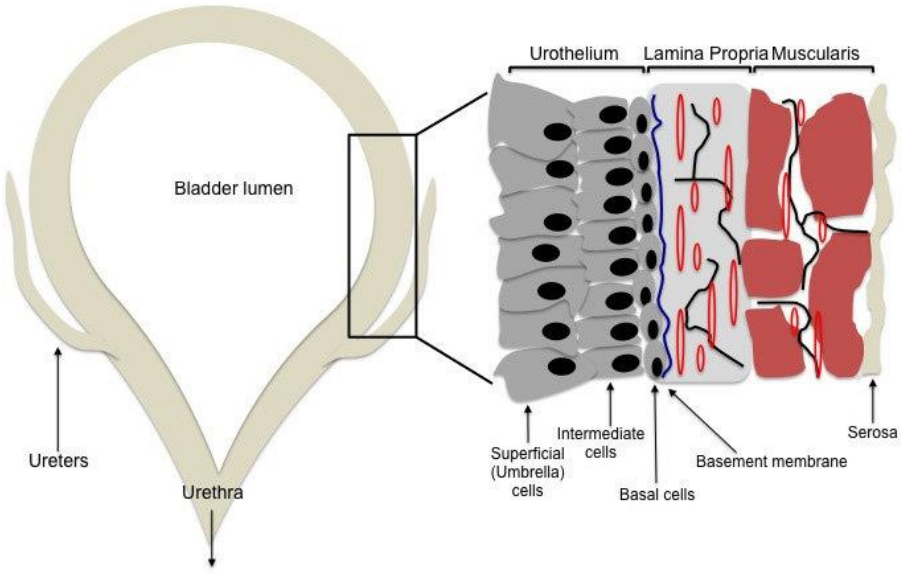

Figure 2. Composition of the urinary bladder wall.
Schematic drawing of the urinary bladder and its different layers. Chan et al. The Current Use of Stem Cells in Bladder Tissue Regeneration and Bioengineering. Biomedicines 2017. DOI 10.3390/biomedicines5010004. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

The bladder receives urine from the kidneys via the ureters. The urine is stored in the bladder until it leaves through the urethra. The bladder capacity is usually 400–500 ml. When the bladder is filled up, the walls are distended and the luminal pressure increases. When the pressure reaches around 25 mm Hg, it triggers micturition. Receptors in the bladder wall are signalling via afferent nerve fibres to the medulla spinalis, and via efferent parasympathetic nerve fibres directly back to the bladder wall for contraction. This reflex is under complex control of centres in the brain, especially areas in the cerebral cortex, the pontine micturition centre, hypothalamus, and the periaqueductal grey substance (PAG). These centres in the brain makes voiding under voluntarily regulation, partly by coordination with the internal and external urethral sphincter.

1.2 Epidemiology and etiology of urothelial bladder cancer

Bladder cancer is the 10th most common cancer globally, with an estimated 549,000 new cases in 2018. The incidence is almost four times higher in men, making bladder cancer the 6th most common cancer in men, and the ninth leading cause of cancer death in men worldwide (1). The incidence is highest in southern and western Europe, North America, and Australia. The prevalence is estimated to be six times higher in developed countries compared to developing countries (2). In men in Norway, bladder cancer is the fourth most frequent cancer type. The male to female ratio for diagnosis of bladder cancer in Norway in the year 2019 was 3.3. The median age for primary diagnosis is 73 years old (3).

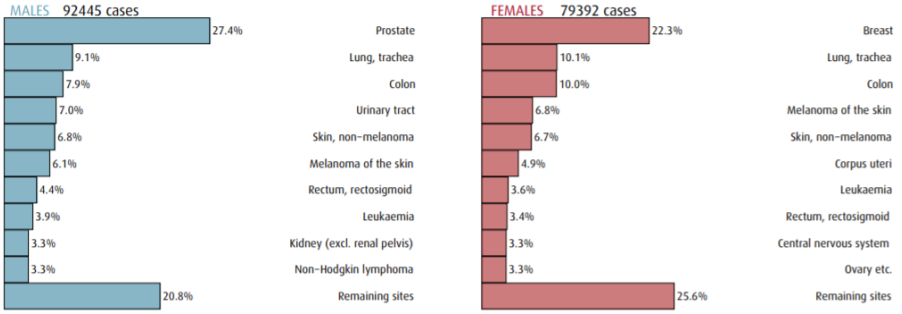

Figure 3. The most frequent types of cancer in Norway by sex, 2015 – 2019.

Cancer in the urinary tract is the fourth most frequent cancer in men. The figure is taken from the Cancer Registry of Norway. Cancer in Norway 2019 - Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2020.

Around 70–80 % of primary urothelial bladder cancer is non-muscle invasive (pTa, pT1, or pTis) at the time of diagnosis. Among these, 50–70 % will experience recurrence, and 15–25 % will progress to muscle-invasive disease (2, 4). Since non-muscle invasive bladder

cancer has a high recurrence rate, the follow-up is extensive, partly with regularly invasive procedures like cystoscopies, to detect new tumours. Detection of progression is essential for prognosis, and will often lead to even more invasive and expensive treatment. This is why bladder cancer is the most expensive cancer per patient in the United States (5). The total costs are estimated to 3.6 billion euro per year in the US and 5 billion euro per year in the European Union (6).

As described above, men are diagnosed more frequently with urothelial bladder cancer than women are. This holds true even after adjusting for differences in smoking habits. Studies are lacking regarding adjustments for occupational exposure. Differences in metabolism and detoxification of carcinogens, are postulated as explanations for the gender-based differences. For example, the sex-based difference in expression of isoforms of 5`-diphosphoglucuronosyltransferase (UGT) in the liver. UGT is involved in the metabolism of aromatic amines, which are important carcinogens in tobacco. Direct carcinogenic effects of androgens and/or protective effects of oestrogen/progesterone could also explain the sex-based differences in incidence of urothelial bladder cancer (7). Despite higher incidence among men, women often seem to have a higher stage at presentation at diagnosis. This can at least partly be explained by delayed referral to haematuria investigation, as women are more likely to be diagnosed with a urinary tract infection. After adjusting for disease stage, women still tend to have a worse prognosis than men, and the reason(s) underlying this difference remain uncertain and debated. At the molecular level, women are overrepresented with basal tumours.

The most important risk factor for bladder cancer is tobacco smoking. Tobacco contains multiple carcinogens, including aromatic amines and N-nitroso compounds, that damage DNA. One meta-analysis estimated the relative risk to be 3.47, when comparing current with never smokers (8). Quitting smoking at time of diagnosis has been shown to reduce the risk of recurrence in non-muscle invasive bladder cancer (9).

There is also an association between smoking opium and risk of bladder cancer (10). Smoking cannabis has not shown an increased risk (11).

Occupational exposure to carcinogens contributes to 5–6 % of bladder cancers (12). Workers in the rubber, metal, and dye industries are particularly high risk for developing bladder cancer. Typical chemical compounds associated with occupational exposure are 2-naphtylamine (dyes like congo red and prodan), 4-aminobiphenyl (a rubber anti-oxidant and an intermediate in dye industry), toluene (an aromatic amine), 4,4-methylenebis (2-chloroaniline) (aromatic amine used in polyurethane production), metal working fluids (liquids cooling or lubricating metal machine pieces), polyaromatic hydrocarbons (PAH), perchloroethylene and diesel exhaust. Individuals' susceptibility to develop bladder cancer when exposed to such carcinogens is variable. At least part of this variation can be attributed to genetic polymorphisms in detoxifying genes. Abnormalities in these genes cause longer exposure to the toxic agents. The three most relevant genes for bladder cancer appear to be GSTM1, UGT1A, and NAT2 (13, 14). Polymorphisms in the genes NAT2 and UGT1A have been shown to be the most important for modifying the effect of smoking (15).

Dietary factors do not appear to contribute significantly to the risk of developing bladder cancer. Neither does alcohol consumption, which does not show an obvious association with bladder cancer risk (16). Arsenic in soil and drinking water is a well-known risk factor (17), probably also in low concentrations (< 100 µg/l) (18). Arsenic in drinking water is a serious environmental problem in parts of Bangladesh, India, China, and Hungary (4).

Multiple studies have revealed increased risk of bladder cancer after radiotherapy in the treatment of other malignancies in the pelvis. A meta-analysis that examined the risk for developing bladder cancer after radiation of the prostate, found a hazard ratio of 1.67 (19). Another retrospective study found that patients with carcinoma of the prostate

receiving radiotherapy had a relative risk of 1.7 for developing a second malignancy in the bladder (20). These patients were more prone to developing non-urothelial bladder cancers and urothelial carcinoma in situ.

Lynch syndrome is a hereditary cancer syndrome causing increased risk of cancer, especially cancer of the colon and rectum. Lynch syndrome patients also have a predisposition for developing cancers in the urinary tract, and for unknown reasons, especially in the upper urinary tract (21). Costello syndrome is a very rare genetic disease caused by germline mutations in the HRAS gene. The syndrome is characterized by developmental abnormalities and a propensity to develop benign and malignant tumours—including urothelial carcinomas—at a young age (22).

The parasite, *Schistosoma haematobium,* enters the body through infested water. It is most prevalent in Africa and the Middle East. The parasite penetrates the skin and continues to live in blood vessels. Female worms lay eggs, some which may be excreted through the urine and faeces. Other eggs of this parasite may get trapped in internal organs. Eggs of *Schistosoma haematobium* may reside in lamina propria and muscularis propria of the urinary bladder. Here, they initiate chronic inflammation and fibrosis, which might further develop into urothelial hyperplasia and squamous metaplasia. Ultimately, this leads to increased risk of bladder cancer. Schistosoma-associated bladder cancer has a higher proportion of squamous- and adenocarcinoma, than traditional bladder cancer (23).

1.3 Classification of bladder cancer

Although all tissue components of the bladder can give rise to malignancies, it is notable that more than 90% of bladder cancers are

urothelial carcinomas. Most of the remaining cancers are also epithelial, like squamous cell carcinoma (3%), adenocarcinoma (0.5–2%) and neuroendocrine carcinoma (<1%). Other malignancies include mesenchymal tumours, melanocytic tumours, tumours of the Müllerian type, and hematopoietic/lymphoid tumours (4, 24).

According to the WHO Classification of Tumours of the Urinary System and Male Genital Organs from 2016, urothelial tumours can be divided into non-invasive urothelial lesions and infiltrating urothelial carcinomas. Among the non-invasive lesions, papillary urothelial carcinomas and urothelial carcinoma in situ are considered malignant, even though they do not necessarily infiltrate the basement membrane. Papillary urothelial carcinomas are defined by neoplastic urothelium covering a fibrovascular stalk. This urothelium shows variable degrees of atypia, a key factor in tumour grading. To be qualified as real papillae, one should observe secondary branches or a complexity giving rise to "detached" islands of stroma covered by urothelium. This important feature distinguishes papillary neoplasias from papillary hyperplasia/ urothelial proliferation of uncertain malignant potential. The latter can also harbour cytological atypia, but will then usually be referred to as dysplasia or urothelial carcinoma in situ with papillary formations. The papillary urothelial neoplasm of low malignant potential (PUNLMP) previously belonged to the papillary urothelial carcinomas grade 1 (WHO73), but is no longer considered a carcinoma. These neoplasias have no obvious cellular atypia other than perhaps some, slightly and homogenously enlarged nuclei. Urothelial carcinoma in situ (CIS) is defined by a flat urothelial lesion of variable thickness, devoid of papillary structures, containing cytologically malignant cells (4).

Tumours that invade the muscularis propria, have a poor prognosis. This is therefore a critical diagnostic characteristic which is important in treatment decision-making. For practical reasons, urothelial carcinomas have been traditionally divided into non-muscle invasive (NMIBC) and muscle invasive bladder cancer (MIBC). Previously the

non-muscle invasive cancers were referred to as "superficial," but this term is no longer recommended. As indicated by the title, this thesis is mainly focusing on non-muscle invasive tumours, but will also address general aspects of urothelial carcinoma of the bladder.

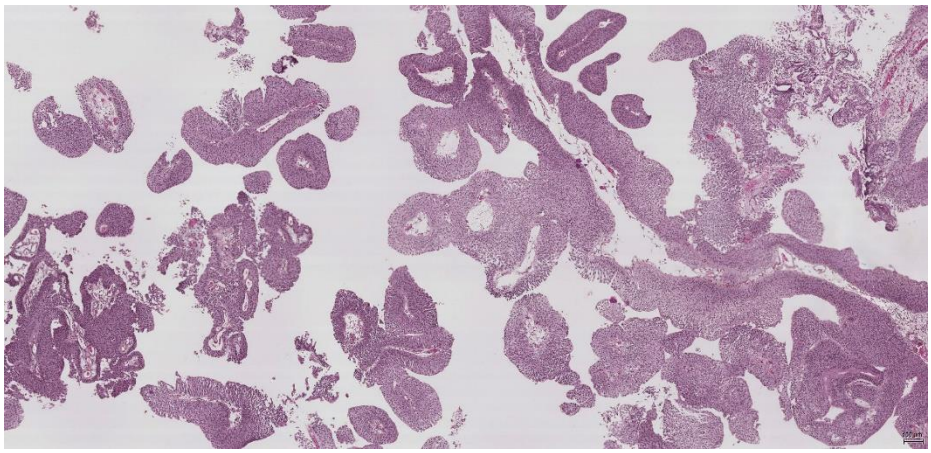

Figure 4. Histology of papillary urothelial carcinoma

Image of a low grade papillary urothelial carcinoma, from our cohort at Stavanger University Hospital. The section is stained with Haematoxylin Eosin Saffron (HES). The image illustrates fibrovascular stalks covered by neoplastic urothelium of varying thickness.

1.4 WHO Grading

The first widely accepted grading system for papillary urothelial carcinomas was introduced by the World Health Organisation (WHO) in 1973 (25). This grading system is still in clinical use, often reported together with the newer version from WHO 2004. The latter was only slightly modified in the WHO "Blue book" from 2016 (26). The WHO grading system from 2004/ 2016 is based on guidelines emerging from a WHO/ISUP (International Society of Urological Pathology) consensus conference in 1998 (27). The European Association of Urology

guidelines provide treatment and follow-up recommendations based on both grading systems (28). The Norwegian guidelines, describe both systems without recommending either of them specifically (29).

According to the WHO grading system from 1973, tumours can be classified into three groups, Grade 1, 2, and 3, based on the degree of cellular anaplasia. Grade 1 tumours show the lowest degree of cellular anaplasia compatible with malignancy. Grade 3 tumours show the highest degree of cellular anaplasia. Grade 2 tumours lie somewhere in between (25). The system relies on the subjective impression of the pathologist, as more detailed grading criteria are not provided.

The WHO grading system from 1973 has been criticized because of low reproducibility and the fact that many tumours ended up in the middle group, grade 2. This resulted in the development of new grading systems, finally ending up with the WHO 2004/2016 system. In this system, the malignant papillary neoplasms from the WHO 1973 grading system, were divided into three categories: "Papillary Urothelial Neoplasia of Low Malignant Potential" (PUNLMP), low grade papillary urothelial carcinoma, and high grade papillary urothelial carcinoma. Consequently, the most benign-looking tumours avoided the "carcinoma" label. PUNLMP is described as a papillary lesion covered by a thicker and usually more cellular urothelium. Polarity is not lost, and no architectural disturbances are apparent. The nuclei are homogeneous, and at most, slightly enlarged. They have a finely granular chromatin pattern, and do not present scattered hyperchromatic nuclei as might appear in low-grade urothelial carcinomas. Mitoses are rare, and when present, are basally located. Several publications show the same prognosis for PUNLMP as for low-grade urothelial carcinomas (30-32), making the future for this designation uncertain. This is reflected by decreasing use of the PUNLMP diagnostic label in the last decade. Hentschel et al. showed a dramatic reduction in the proportion of PUNLMP diagnoses in a large European/Canadian cohort, from its introduction in 1998 to 2018, falling from 31.3% to 1.1% (32).

Low grade papillary urothelial carcinomas are supposed to maintain a relatively orderly appearance at low or medium magnification (100x or 200x). Looking closer at higher magnification, some loss of polarity and mild nuclear irregularity and pleomorphism are evident. Scattered hyperchromatic nuclei can also be seen. Mitoses may be present and may occasionally be found in locations away from the basal lamina. Atypical mitoses are usually not seen.

High grade papillary urothelial carcinomas give an impression of disorder, both regarding architectural and cytologic features, at medium magnification. Architectural features include how the cells are oriented in relation to each other and to the basal lamina. Cytologic features include nuclear shape, size, and patterns of chromatin structure and distribution. Irregular and prominent nucleoli might be seen. Mitoses are typically frequent, with some of these being even atypical. High grade tumours often have fused papilla, giving an impression of greater solidity.

Tumours are graded based on the most anaplastic area. No consensus exists regarding the minimum size or proportion of such an area. Heterogeneity of morphologic grade is not uncommon. Cheng et al. investigated heterogeneity in urothelial carcinomas and prognostic correlation (33). In their cohort, 32% of tumours could be assigned both a primary and a secondary grade. In this study, combining primary and secondary grade increased the prognostic accuracy.

Although both WHO 1973 and WHO 2004/2016 guidelines include three groups, these groups cannot be readily mapped onto each other. The WHO 1973 guidelines regarding grade 1 include all PUNLMP and some low-grade carcinomas. The WHO 2004/2016 guidelines related to high grade tumours include all WHO 1973 grade 3 and most grade 2 tumours (34).

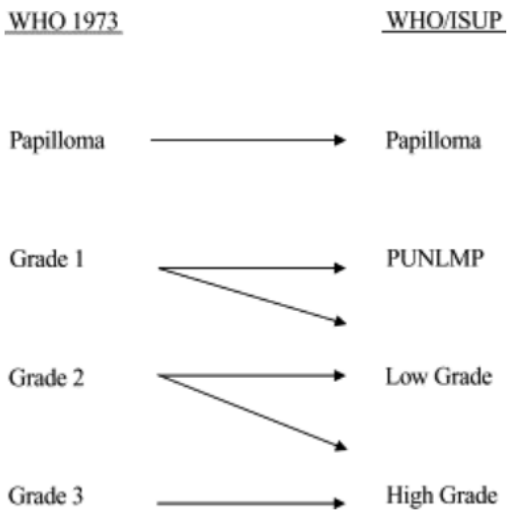

Figure 5. The relationship between WHO 1973 and WHO 2004/ 2016

Reprinted from European Urology, Lopez-Beltran A, Montironi R, Non-Invasive Urothelial Neoplasms: According to the Most Recent WHO Classification, pages 170 – 176, Copyright 2004, with permission from Elsevier.

Grading in general is considered one of the most important prognostic factors for NMIBC and has a major impact on treatment and follow-up regimes. It was previously regarded as prognostic regarding both recurrence and progression (34, 35). In a newly published, large, retrospective study including 5145 patients with pTa/pT1tumours, both grading systems were prognostic regarding disease progression, but neither of them was prognostic for recurrence (36). In this study, progression to MIBC was found in 1.4%, 5.6%, and 18.8% of grade 1, grade 2, and grade 3 patients, respectively. For low grade and high grade NMIBC, the progression rates were 2.6% and 13.7%, respectively.

Concerns regarding reproducibility of the different grading systems have been raised for decades. More detailed grading descriptions in the WHO 2004/2016 system were aimed at improving reproducibility.

However, several publications on this subject failed to show any significant improvement in reproducibility (37-39). A previous publication from our research group by Mangrud et al., confirmed the lack of significant improvement in grading reproducibility (40). In a review from 2017, Sokoup et al. concluded that the WHO 2004/2016 guidelines were marginally better concerning reproducibility, but could not confirm that this system decisively "outperformed" the WHO 1973 system (35). For the WHO 1973 classification system, the interobserver agreement according to Sokoup et al. ranged from 38%–89% (kappa values 0.003 to 0.68). For the WHO 2004/2016 system, the agreement ranged from 43%–100% (kappa values 0.17–0.70).

1.5 Variants of urothelial carcinoma

Both the papillary urothelial carcinoma and the flat urothelial carcinoma in situ can at some point start to infiltrate into the muscularis propria. Most tumours are considered conventional urothelial carcinomas, but some show a divergent differentiation, like squamous (20–40%), glandular (6–18%) and trophoblast differentiation (28–35%) (41). As long as part of the tumour has a urothelial morphology, it will be classified as a urothelial carcinoma. Divergent differentiation has no clinical consequence, as prognosis is the same as for a pure conventional urothelial carcinoma.

Infiltrating urothelial carcinomas
Conventional urothelial carcinoma
Urothelial carcinoma with divergent differentiation
Squamous differentiation
Glandular differentiation
Trophoblast differentiation
Urothelial carcinoma with variant histology
Nested urothelial carcinoma
Microcystic urothelial carcinoma
Micropapillary urothelial carcinoma

Lymphoepithelioma-like urothelial carcinoma
Plasmacytoid urothelial carcinoma
Sarcomatoid urothelial carcinoma
Giant cell urothelial carcinoma
Poorly differentiated urothelial carcinoma
Lipid-rich urothelial carcinoma
Clear cell urothelial carcinoma

Table 1. Overview of types of infiltrating urothelial carcinomas (41).

It has been recognized that some infiltrating urothelial carcinomas have other distinct morphological characteristics, which we refer to as variant histology (4, 41, 42). These are listed in Table 2. Although nested and microcystic tumours generally have low grade atypia, they often present as high-stage tumours. When corrected for TNM stage, they still have the same prognosis as conventional urothelial carcinomas. Nested carcinomas are characterized by crowded growth of tumour cells in small nests. The nests can be somewhat irregular and confluent. As they have very little cytological atypia, these tumours can be difficult to distinguish from florid von Brunn nests, although nested carcinomas often have a more infiltrating growth pattern in the invasive front. Microcystic urothelial carcinoma form oval infiltrating cysts, lined by bland urothelium with little atypia. They can be confused with cystitis cystica or cystitis glandularis.

The lymphoepithelioma-like urothelial carcinoma is rare and has a morphology reminiscent of the lymphoepithelioma of the nasopharynx. Typically, they grow in sheets with a syncytial appearance. The tumour cells have pleomorphic nuclei with prominent nucleoli. In the background one can observe a high number of mixed inflammatory cells. The prognosis of this tumour type is controversial (43).

Micropapillary, plasmacytoid, and sarcomatoid variants are associated with poor prognosis. Patients with a micropapillary urothelial carcinoma pT1, might benefit from early cystectomy (44). Microscopically, they are characterized by small nests of tumour cells

within lacunae. The nests usually do not contain fibrovascular cores. The plasmacytoid variant of urothelial carcinoma consists of infiltrating tumour cells resembling plasma cells or monocytes. The cells are typically dyscohesive, lacking immunohistochemical staining for E-cadherin. Sometimes these tumours contain vacuolated cells with the appearance of signet-ring cells. The sarcomatoid variant shows both epithelial and mesenchymal differentiation. Morphologically, they may resemble a sarcoma, still exhibiting positivity for cytokeratins via immunohistochemistry (45).

Giant cell urothelial carcinoma, lipid-rich urothelial carcinoma, and poorly differentiated carcinomas are rare, but generally associated with a poor outcome. Giant cell urothelial carcinomas are characterized by pleomorphic giant cells. Lipid-rich urothelial carcinomas are recognized by one or more lipid vacuoles in the tumour cells compressing a peripheral nucleus. Poorly differentiated tumours are tumours with mixed morphology, and include sarcomatoid, giant cell, and undifferentiated carcinomas. Finally, the clear cell urothelial carcinoma contains glycogen vacuoles in the cytoplasm, positive for staining with periodic acid Schiff (PAS), and not resistant to diastase. These tumours are also rare. Prognostic significance of these different types of histology is presently still unclear (41).

1.6 Molecular alterations in bladder cancer

Bladder cancer in general has a high mutational burden. Only melanoma and lung cancer have a higher average number of mutations per million base pairs (46). Most somatic mutations involved in tumorigenesis are in genes coding for transmembrane receptors, proteins in signalling pathways, cell cycle regulators, or proteins implicated in DNA damage repair. Other frequently occurring mutations in bladder cancer are in genes involved in chromatin regulation (47). Chromosomal alterations

are common in invasive urothelial carcinomas. These tumours are often genomically unstable, and harbor deletions that result in loss of important tumour suppressor genes.

As carcinogens are present in the urine, the complete mucosal surface is exposed. DNA damage can potentially occur at different sites, and not all of them will be repaired. One would therefore expect common molecular alterations spread along the mucosal surface, even in areas where there is no visible tumour. This effect, referred to as the "Field effect", increases cancer risk at other locations in the urinary tract once a primary cancer diagnosis is made. One such early event in carcinogenesis, also found in surrounding normal looking mucosa, is the occurrence of deletions in chromosome 9. This is a common finding in both non-invasive and muscle-invasive urothelial carcinomas. Both loss of the short and long arms of chromosome 9 have been described. On the short arm, 9p, tumour suppressor genes like *CDKN2A* are located. *CDK2NA* encodes proteins like p14 and p16, both inhibitors of the cell cycle. The long arm, 9q, harbour the TSC gene, a negative regulator of the PI3K/ AKT/ mTOR pathway, which is important in regulation of cell growth, proliferation, and survival. Alterations in chromosome 9 are regarded as among the earliest events in the development of urothelial bladder cancer (48, 49).

Traditionally, two major developmental tracks have been identified for urothelial carcinomas (47, 50, 51). They are referred to as papillary and solid (non-papillary) tumours. Some overlaps exist between these two pathways. Most of the non-invasive papillary urothelial carcinomas develop through hyperplasia. These are tumours with a high recurrence tendency. They usually do not infiltrate, but can gain additional molecular alterations and then convert to high-grade invading lesions. Tumours in the papillary group typically show activating mutations in the gene for Fibroblast Growth Factor Receptor 3 (*FGFR3*). *FGFR3* is a receptor tyrosine kinase that regulates both the MAP kinase/ERK pathway and the PI3K/AKT/mTOR pathway. Both

these pathways are involved in cell proliferation. The presence of *FGFR3* mutations is correlated with both grade and stage. One study by Hernandez et al. found *FGFR3* mutations in 77% of what they call low malignant lesions. *FGFR3* mutations were also found in 61% and 58% of pTaG1 and pTaG2 tumours, respectively. By contrast, only 17% of pT1G3 tumours had *FGFR3* mutation (52). Tumours evolving through this "papillary" track are also overrepresented with activating mutations in the *PIK3CA* gene. This gene codes for a subunit of the enzyme phosphatidyl 3-kinase (PI3K), mediating signals from the transmembrane receptor. Mutations in the *PIK3CA* gene often coexist with mutations in *FGFR3*. Mutations in *PIK3CA* are correlated with both low grade and low stage (53). Non-invasive papillary tumours are characterized by the presence of wild-type *TP53* and are usually genomically stable. Mutations in *CDKNA1, RB1, ERCC2, ERBB3,* and *FBXW7* are generally not seen in non-invasive cancers, but are observed in >10% of the muscle-invasive tumours (54).

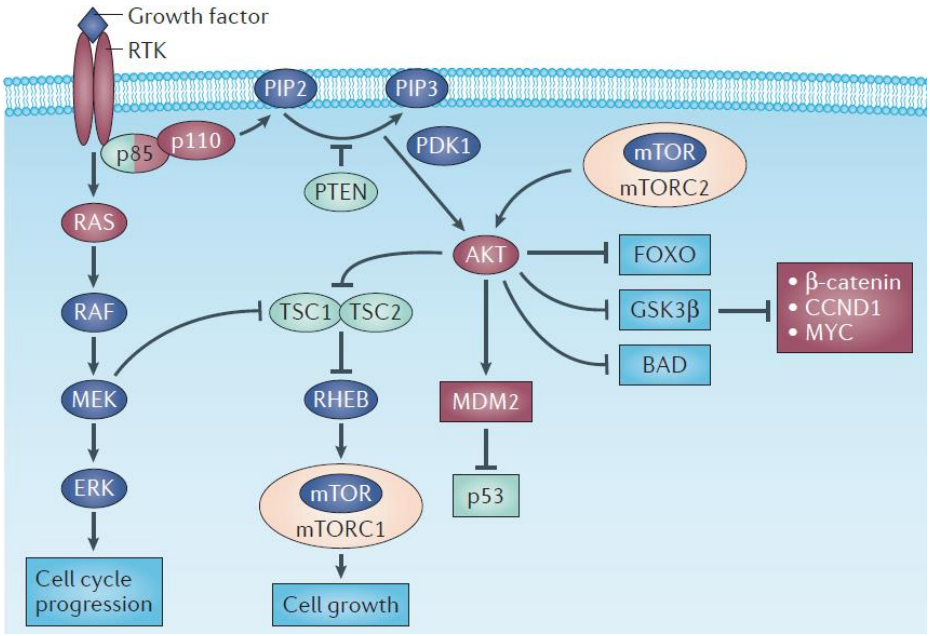

Figure 6. Schematic depicting the PI3K/AKT/mTOR and the MAP kinase/ERK pathways.

The illustration shows signalling cascades and complex interactions. Receptor tyrosine kinases (RTK), like ERBB2, ERBB3, FGFR1, FGFR3, and EGFR are often mutated in bladder cancer. FGFR3 is frequently mutated in tumours in the so-called papillary pathway. Activating mutations in the *PIK3CA* gene—the gene encoding the catalytic subunit p110—results in activation of AKT, ultimately leading to increased proliferation. Mutations in *PIK3CA* are also frequently seen in non-invasive urothelial carcinoma, often in combination with activating *FGFR3* mutations. In many muscle-invasive carcinomas, the pathway inhibitor PTEN has lost its function. The RTKs activate the MAP kinase/ERK pathway via RAS. Proteins typically activated in bladder cancer are depicted in red, those typically inactivated are depicted in green. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Nature reviews cancer. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. Knowles MA et al. Copyright 2015.

Most muscle-invasive urothelial carcinomas (solid/non-papillary) develop de novo or through carcinoma in situ (CIS). Tumours following this second track are characterized by loss of function of *TP53* or *RB1*, either because of mutations or by copy number alterations. Robertson et al. found *TP53* mutations in 48% of muscle-invasive urothelial carcinomas (55). In low grade non-invasive papillary tumours, the proportions of tumours with *TP53* mutations vary from 0 to 14% (56). *TP53* is a gene coding for the tumour suppressor protein p53. The p53 protein inhibits cell cycle progression from G1 to S, and regulates expression of other genes involved in cell cycle arrest, apoptosis, senescence, DNA repair, and changes in metabolism. Loss of heterozygosity for the gene *PTEN*, usually by a deletion of part of chromosome 10, is much more common in muscle-invasive disease (57). It is postulated that *PTEN* regulates invasion (58). *PTEN* is a tumour suppressor gene with an inhibitory effect on the PI3K/AKT/mTOR pathway. Other typical alterations along the track towards muscle-invasive disease are activating mutations in the receptor tyrosine kinases *ERBB1 (EGFR)*, *ERBB2* (*HER2*/neu), and *ERBB3*, as well as inactivating mutations in chromatin remodelling genes such as *KDM6A*, *MLL2*, and *ARID1A*. Finally, these aggressive tumours often are more genomically unstable, and deletions of chromosome arms 8p, 2q, and 5q are frequently observed (51).

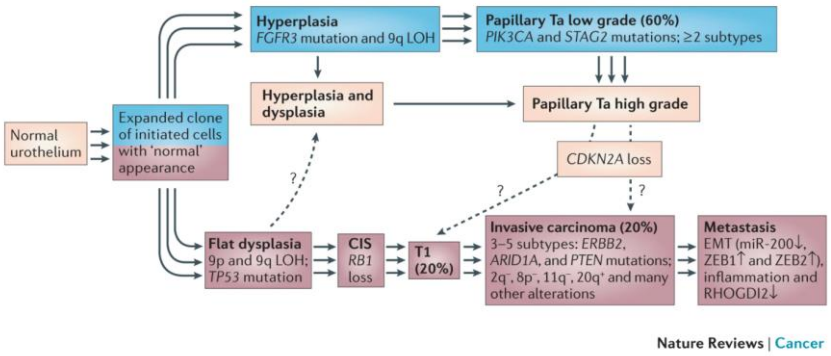

Figure 7. The "two-track model" of urothelial carcinoma development.

Illustration showing the proposed main tracks in the development of urothelial carcinoma. Blue colour indicates the "papillary" pathway and purple indicates the non-papillary/solid pathway. Recently identified molecular subtypes indicate many sub-pathways in each main track. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Nature reviews cancer. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. Knowles MA et al. Copyright 2015.

There are several known molecular alterations that are commonly found in urothelial bladder cancer, independent of the main development track followed. Mutations in the promoter of the telomerase reverse transcriptase (*TERT*) gene are the most frequently described mutations in bladder cancer (59). *TERT* promotor mutations have no prognostic significance, and probably represent an early step in tumorigenesis. TERT contributes to elongation of telomeres on the chromosome ends, thereby preventing cellular senescence. Mutations involving the MAP-kinase pathway, which promotes proliferation and survival, are found in all stages as well, but they are less common (60). Among these are mutations in *BRAF, HRAS, KRAS,* and *NRAS.*

Cancer genome sequencing in humans has revealed distinct mutational patterns, referred to as mutation signatures. Some of these

signatures are linked to specific exogenous or endogenous mutagens. The endogenous protein family of cytidine deaminases, APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like), has been shown to be such a mutagenic source (61, 62). They have a role in mRNA editing and innate immunity against viruses. The APOBEC catalytic unit converts cytosine bases to uracil. Uncontrolled activity of APOBEC seems to induce mutations in DNA. Bladder cancer, together with breast, lung, head/neck and cervical cancers, are enriched in such mutations. Shi et al. established a panel of 44 hotspot mutations associated with bladder cancer (63). One of the most frequent *FGFR3* mutations—Serine249Cystidine—is probably APOBEC-mediated (64). Hedegaard et al. showed that the level of APOBEC signature mutagenesis is correlated with the expression of APOBEC3A and APOBEC3B (65).

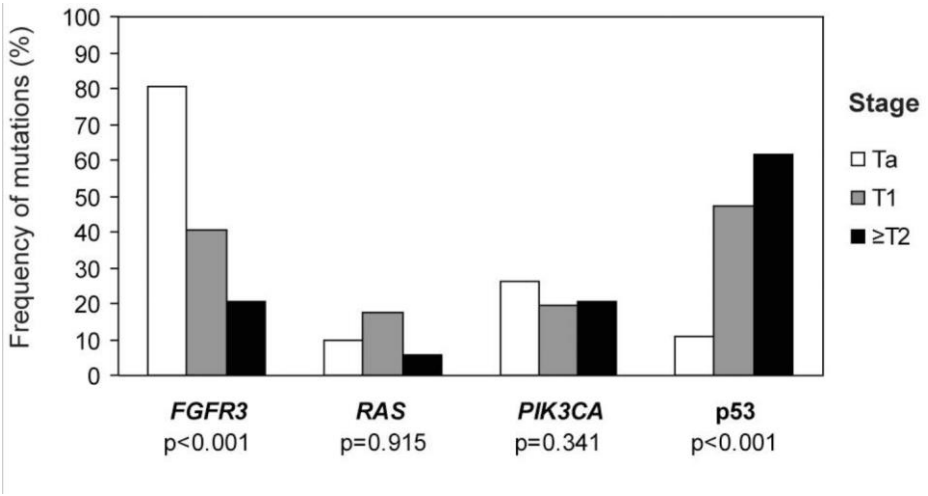

Figure 8. Frequencies of *FGFR3, RAS, PIK3CA* mutations and p53 overexpression (indicating mutation) according to stage.
FGFR3, HRAS, KRAS, NRAS, and *PIK3CA* Mutations in Bladder Cancer and Their Potential as Biomarkers for Surveillance and Therapy. Kompier et al. PLOS One 2010. DOI 10.1371/journal.pone.0013821. This article is distributed under the terms of the Creative Commons Attribution License.

1.7 Molecular classification of bladder cancer

Urothelial carcinomas have diverse clinical behaviours. Several different molecular classification proposals that attempt to categorize tumours based on similarities in molecular alterations, exist. These classification systems aim to improve the prediction of treatment response and prognosis in general. New high-throughput techniques for RNA sequencing have made extensive gene expression profiling possible. Other multi-parametric systems combine gene expression profiling with known genome alterations, protein analysis, and other clinical data. During the last decade several molecular classification systems for bladder cancer have been proposed based on different analytical techniques and approaches.

The first molecular taxonomy for urothelial carcinoma was presented by Sjödahl et al. in 2012, The Lund University Classification (66). Gene expression analysis of 308 tumours, including supervised analysis of 13,953 genes, yielded 5 major molecular subtypes: Urobasal A, Genomically unstable, Infiltrated type, Urobasal B, and squamous cell carcinoma (SCC)-like. The clusters were identified using hierarchical clustering analysis. The molecular subtypes showed different survival patterns—the Urobasal A subtype had the best prognosis, the Genomically unstable and the infiltrated type had an intermediate prognosis, while the Urobasal B and the SCC-like subtypes had the poorest prognosis.

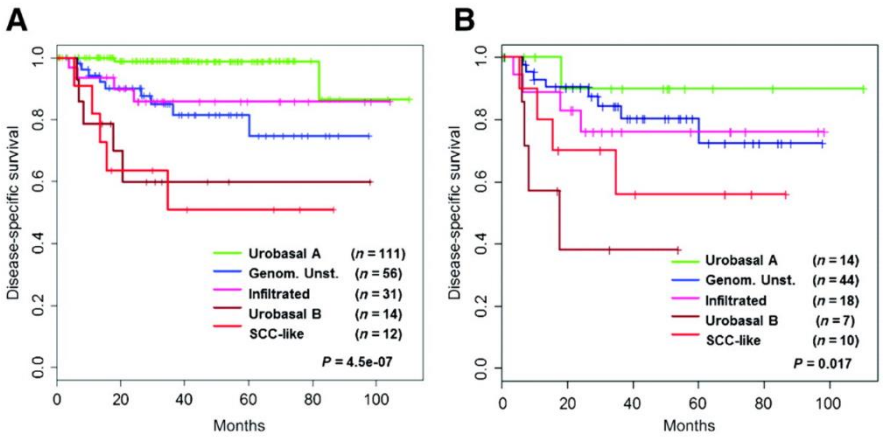

Figure 9. The Lund University Classification and disease-specific survival.

Kaplan-Meier survival curves for the different bladder cancer subtypes: (A) including patients of all grades and stages, (B) including patients with grade 3 tumours only. Reprinted by permission from Clinical Cancer Research. A Molecular Taxonomy for Urothelial Carcinoma, Sjödahl et al, Copyright 2012.

Based on genes with coordinated expression and known biologic function, an immunohistochemical validation for selected gene products was performed (66). Pronounced differences and clustering in expression of cell cycle regulatory proteins was noted. The Urobasal A tumours tended to express genes involved in early cell cycle regulation. On the other hand, the Genomically unstable and SCC-like subtypes tended to express genes involved in late cell cycle regulation. Furthermore, the different molecular subtypes can be distinguished by different cytokeratin signatures. Urobasal A and Genomically unstable tumours mostly expressed cytokeratins like CK8/18, CK7/19, and CK20. By contrast, both Urobasal B and SCC-like tumours expressed CK5/13/15/17, which are typically expressed in basal and intermediate urothelial cells. Sjödahl et al. observed that the Urobasal A tumours maintained the tissue stratification found in normal urothelium to some

extent, giving this group the designation Urobasal A. The SCC-like tumours are characterized by expression of cytokeratins associated with squamous/keratinized phenotypes—namely, CK6/14/16. Pathological re-evaluation of these tumours in most cases showed squamous differentiation—therefore the name squamous cell carcinoma-like (SCC-like). Supplementary mutation analysis revealed higher *FGFR3* mutation frequencies in Urobasal A compared to Genomically unstable tumours (55% vs. 7%). The frequency of *TP53* mutations was higher in Genomically unstable tumours compared to Urobasal A tumours (48% vs 11%). Since tumours of the former subtype tended to have rearranged genomes, they were designate "Genomically unstable." Finally, the molecular subgroups expressed different cell adhesion gene signatures, like claudins (tight junction-associated genes), with higher expression in Urobasal A and Genomically unstable tumours compared to SCC-like and Urobasal B. The Urobasal B subtype is considered to be a progressed version of Urobasal A. They have many similarities, like a high proportion of *FGFR3* mutation, but additionally show frequent mutations in *TP53* and express cytokeratins similar to those in the SCC-like subtype. The infiltrated type probably represents a heterogeneous group of tumours having in common a gene expression profile characterized by the infiltration of immune cells. In 2017, Sjödahl et al, investigated a cohort with muscle invasive urothelial bladder cancer only. Based on gene expression profiling in this cohort they slightly modified the classification system, establishing the groups Urothelial like (formerly Urobasal), Genomically unstable, epithelial-infiltrated, SCC-like/mesenchymal infiltrated, SCCL/Uro B and small-cell/ neuroendocrine-like (67). These groups did not correlate with immunohistochemical phenotyping.

In 2014, Damrauer et al, from the University of North Carolina (UNC), looked for intrinsic molecular subtypes of urothelial bladder cancer, in a cohort of muscle-invasive disease (68). By gene expression profiling, they identified two subtypes, "luminal" and "basal-like".

Basal-like tumours had significantly worse prognosis than the luminal tumours. These subtypes largely reflected the subtypes already described for breast cancer (69). In bladder cancer, luminal and basal-like tumours represented different stages of urothelial differentiation. The basal-like tumours expressed high levels of markers for basal urothelial cells, like CK14, CK5, CK6B, and CD44. By contrast, luminal tumours expressed markers associated with urothelial differentiation—like CK20 and uroplakins—that are typically expressed in urothelial umbrella cells. Furthermore, the two subtypes showed differences in genetic alterations: while luminal tumours had higher frequencies of *FGFR3* mutations, the basal-like tumours had a higher frequency of alterations in the RB1 pathway. Damrauer et al. created a panel of 47 genes (BASE47) for molecular classification of luminal and basal-like urothelial bladder cancer. They also tested it on a dataset of non-muscle invasive tumours. Some of the tumours fell into the basal-like subgroup. This indicates that the UNC classification might also work for non-muscle invasive tumours. In breast cancer, a "claudin-low" subtype has been previously identified (70). In 2016, the research group that originally described the claudin-low breast cancer subtype discovered a similar "claudin-low" subtype in muscle-invasive urothelial bladder cancer (71). In this system, 10% of urothelial bladder cancers were identified as "claudin-low." Interestingly, all these "claudin-low" tumours were previously identified as belonging to the basal-like subtype. The claudin-low tumours had the same prognosis as basal-like tumours, and expressed immune-related genes (including proinflammatory cytokines) at high levels, combined with a high level of expression of immunosuppressive genes, specifically PD-L1. Their gene expression profile suggests that these tumours are heavily infiltrated by immune cells.

At the same time Choi et al, from MD Anderson Cancer Center (MDA), Houston, Texas, presented a similar classification system (72). Choi et al. were also inspired by molecular subtyping in breast cancer, and found similar results for muscle-invasive bladder cancer. Based on

whole genome mRNA expression profiling and hierarchical clustering analyses, they reported three urothelial bladder cancer subgroups: basal, luminal, and p53-like. The basal tumours were enriched with basal biomarkers, in the same way as basal breast cancers (CD44, CK5, CK6, CK14, and P-cadherin), the luminal tumours are characterized by the same markers as for luminal breast cancer (CD24, FOXA1, GATA3, ERBB2, ERBB3, XBP1, and CK20). p53-like tumours also expressed luminal markers, but in addition, showed an activated wild-type *TP53* gene expression signature. The *TP53* mutation frequency was the same for all the three subtypes. As for the UNC classification system, the basal tumours were associated with shorter disease specific and overall survival.

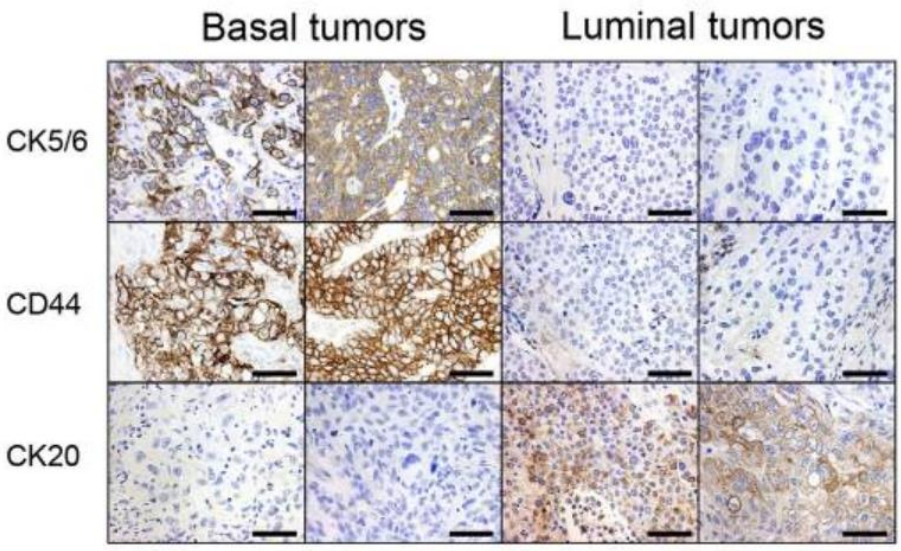

Figure 10. Basal and luminal markers.

Immunohistochemistry for basal (CK5/6, CD44) and Luminal (CK20) markers in representative basal and luminal tumours defined by gene expression profiling. Reprinted from Cancer Cell, Volume 25, Choi et al, Identification of Distinct Basal and Luminal Subtypes of Muscle-Invasive Bladder Cancer with Different Sensitivities to Frontline Chemotherapy, pages 152 – 165, Copyright (2014), with permission from Elsevier.

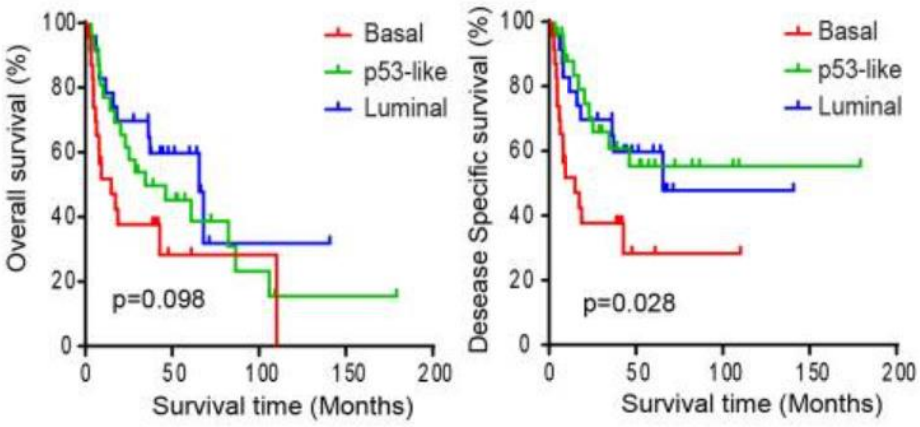

Figure 11. MD Anderson Molecular classification of urothelial bladder cancers and survival analyses.

Kaplan-Meier plots showing overall survival and disease-specific survival for the 3 tumour subtypes. Reprinted from Cancer Cell, Volume 25, Choi et al, Identification of Distinct Basal and Luminal Subtypes of Muscle-Invasive Bladder Cancer with Different Sensitivities to Frontline Chemotherapy, pages 152 – 165, Copyright (2014), with permission from Elsevier.

The basal tumours are characterized by squamous features, similar to the SCC-like tumours, from the Lund classification system. The MD Anderson tumours classified as luminal corresponded well to the Lund Urobasal A tumours. Choi et al. found that the transcription factor p63 was important in controlling the expression of basal genes (72). Correspondingly, the transcription factor PPARγ controlled the luminal gene expression signature.

The MD Anderson classification system was validated in a meta-analysis from 2016 that confirmed the intrinsic molecular subtypes luminal and basal (73). The p53-like subtype was relegated to a subtype under luminal and basal. The meta-analysis included two cohorts with a significant number of non-muscle invasive tumours, confirming the existence of luminal and basal subtypes in non-muscle invasive disease.

The gene expression subtypes were correlated with immunohistochemical phenotypes. Selected markers for basal and luminal tumours were investigated, aiming to find widely available markers for diagnostic use. Suggested markers for differential use were (a) CK5/6 and CK14 for basal tumours, and (b) GATA3, CK20, and uroplakin 2 for luminal tumours, although the markers showed some tendency to overlap. The best combination uncovered in this study was GATA3 and CK5/6, reaching an impressive accuracy of 91% (73).

The publicly available The Cancer Genome Atlas (TCGA) cohort, made it possible to investigate genomic alterations in urothelial bladder cancers, and correlate the results with the tumours' intrinsic molecular subtypes. In this cohort, the mutational burden and the overall mutational landscape were comparable for luminal and basal tumours (73). One could identify some genes that were more frequently mutated in specific molecular subtypes: *FGFR3*, *ELF3*, *CDKN1A*, and *TSC1* were more frequently mutated in luminal tumours, whereas *TP53*, *RB1*, and *NFE2L2* were more frequently mutated in basal tumours.

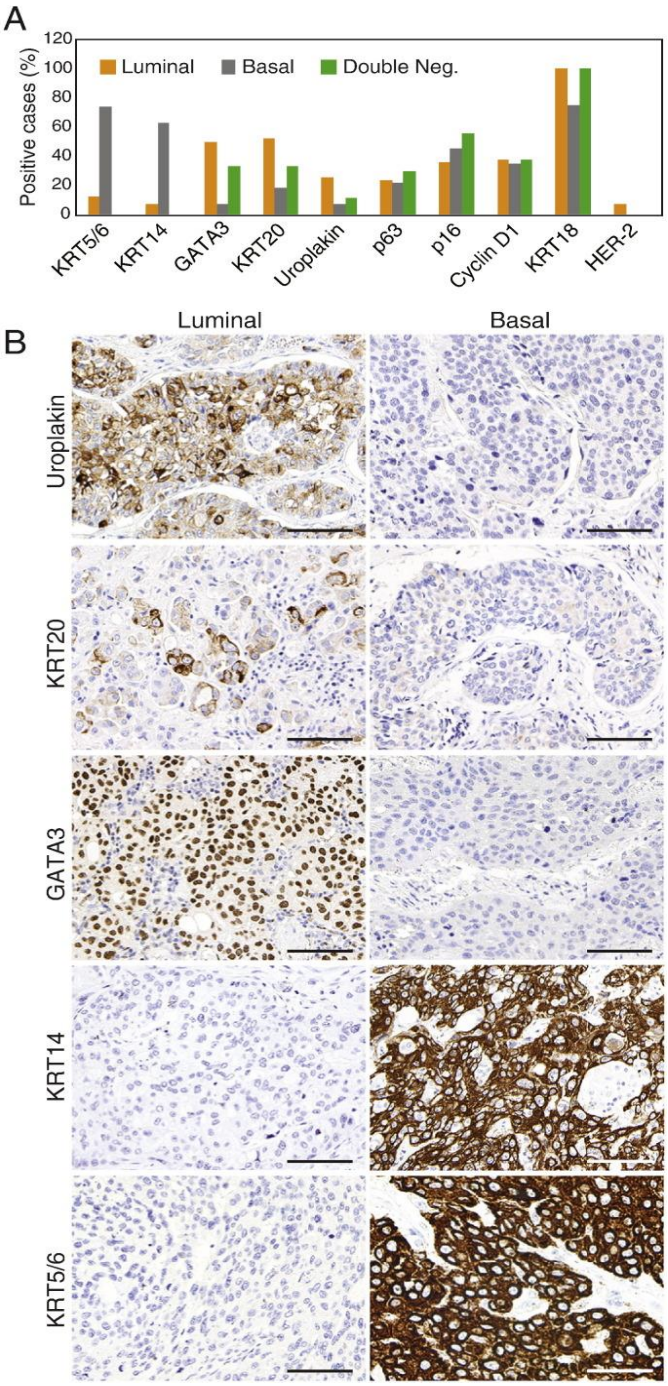
A
Positive cases (%)
120
100
80
60
40
20
0
Luminal
Basal
Double Neg.
KRT5/6
KRT14
GATA3
KRT20
Uroplakin
p63
p16
Cyclin D1
KRT18
HER-2
B
Luminal
Basal
Uroplakin
KRT20
GATA3
KRT14
KRT5/6

Figure 12. Basal and luminal markers, their proportion of positivity in basal, luminal and double negative urothelial bladder tumours.
A: Diagram showing proportion of positive basal, luminal or double negative tumours for the different well known expression markers. B: Representative luminal and basal tumours with their immunohistochemical phenotype. Dadhania et al. Meta-Analysis of the Luminal and Basal Subtypes of Bladder Cancer and the Identification of Signature Immunohistochemical Markers for Clinical Use. EBioMedicine 2016. DOI 10.1016/j.ebiom.2016.08.036. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

In 2014, TCGA Research Network also published a classification system based on mRNA and miRNA expression profiling, protein data, and cluster analysis (47). They proposed four clusters, I–IV for muscle-invasive disease.

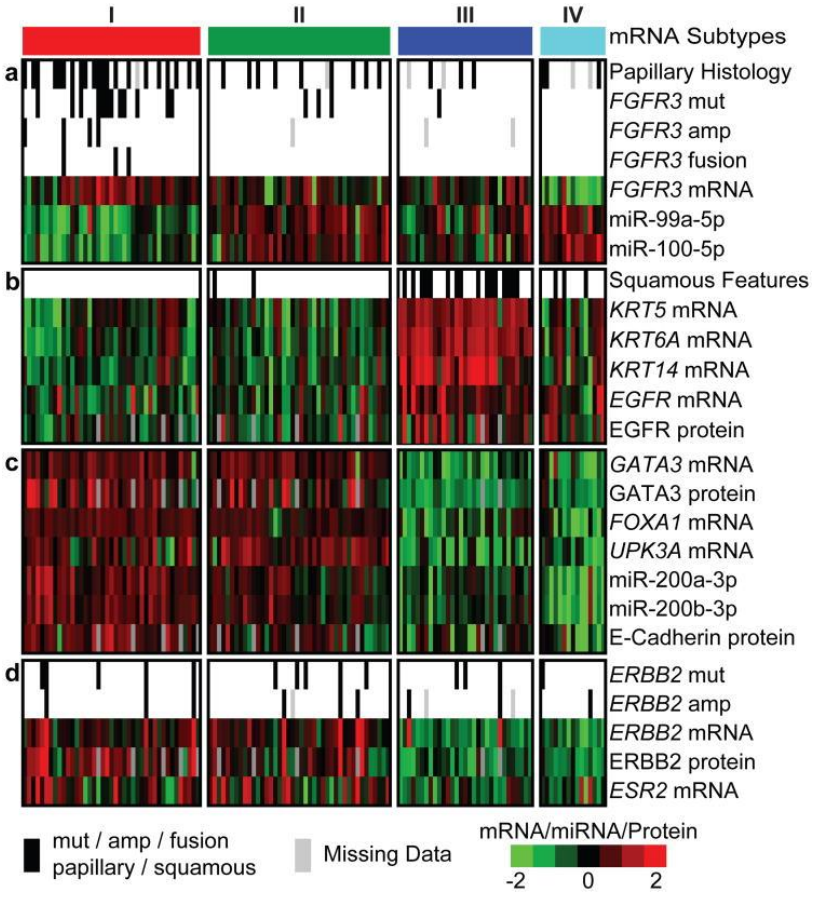

Figure 13. Expression characteristics of bladder cancer, according to TCGA.
The figure shows z-normalized data for expression characteristics. In (**a),** FGFR3 alterations are presented, together with papillary histology. The presented miRNAs downregulate FGFR3 expression. (**b)** shows basal and stem cell markers, including squamous features. (**c)** presents markers for urothelial differentiation and (**d)** shows ERBB2 characteristics and oestrogen receptor beta expression. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 2014.v DOI 10.1038/nature12965 (https://creativecommons.org/licenses/by-nc-sa/3.0/).

Cluster I was characterized by activation of FGFR3, either by mutation, amplification, or otherwise elevated expression (lower level of inhibiting miRNAs). They also had typical papillary histology. Both cluster I and II had high levels of ERBB2 and oestrogen receptor beta expression. Cluster III corresponded to the basal markers in breast cancer. Also, the

TCGA research network associated the expression of these markers with histological squamous features.

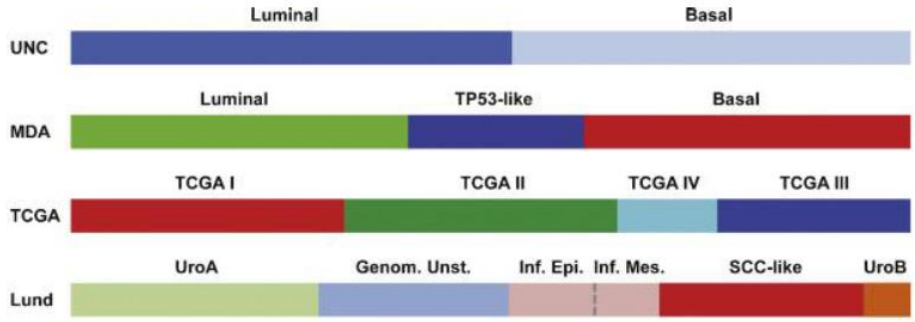

Figure 14. Interrelationship between the first proposed molecular classification systems (74).

Reprinted from European Urology, Vol 68, Aine et. al, On Molecular Classification of Bladder Cancer: Out of One, Many, pages 921 – 923, Copyright (2015), with permission from Elsevier.

With the exception of the work from Lund University (66), most research on urothelial bladder cancer has been performed on cohorts consisting of muscle-invasive disease only. In 2016, Hedegaard et al. carried out whole transcriptome RNA sequencing analysis on a cohort of non-muscle invasive urothelial carcinomas (65). They identified three classes of tumours: class 1, 2, and 3. Class 2 tumours had the poorest prognosis, with the shortest progression-free survival. Class 1 tumours were characterized by a relatively good prognosis, and class 3 tumours happened to be somewhere in between. Class 1 tumours showed high expression of early cell cycle genes, like the Urobasal A tumours in the Lund taxonomy, while class 2 tumours showed high expression of late cell cycle genes, corresponding to the genomically unstable and SCC-like tumours. Both class 1 and 2 showed high expression of uroplakins, indicating luminal differentiation. High expression of CK5, CK15, and CD44— markers of undifferentiated/basal cells—was observed in class 3 tumours. Class 1 and class 2 showed luminal characteristics, while

class 3 showed basal-like features. Although both classes (1 and 2) have luminal characteristics, class 2 tumours were more clinically aggressive than class 1 tumours. Hedegaard et al. performed whole genome mutation analysis on the same cohort and found that mutations in genes involved in DNA damage response, mutations in the MAP kinase/ERK pathway and in ERBB family genes, where associated with class 2 tumours. Furthermore, the APOBEC-related mutational signature was mostly seen in class 2 tumours. Most importantly, this study supports the use of molecular classification in non-muscle invasive bladder cancer.

The existence of different, although somewhat overlapping, molecular classification systems impede their clinical use and limit their utility. Therefore, a consensus molecular classification of muscle-invasive bladder cancer was published in 2020 (75). The transcriptomic profiles from 1750 muscle-invasive bladder cancers, from 18 different and independent datasets, were classified according to six different previously described molecular classification systems (Lund, MDA, UNC, TCGA, as described here, but also two proposals not further mentioned here, Baylor and Cartes d'Identite des tumeurs). Based on statistical algorithms six classes/subtypes are proposed; luminal papillary (LumP), luminal non-specified (LumNS), luminal unstable (LumU), stroma-rich, basal/squamous (Ba/Sq) and neuroendocrine-like (NE-like). All three luminal classes overexpressed markers for urothelial differentiation, and Ba/Sq and NE-like tumours overexpressed markers for basal/squamous and neuroendocrine differentiation, respectively. LumP typically showed papillary histology and overexpression of FGFR3. LumNS were luminal tumours with stromal and immune cell expression signatures. Histologically, they seem associated with micropapillary variant. LumU had an expression profile consistent with high cell cycle activity. They were associated with mutations in *TP53* and *ERCC2*. Stroma-rich tumours were characterized by less urothelial differentiation and a gene expression signature in the direction of smooth muscle, fibroblast and myofibroblast, in addition to immune cell

infiltration. Ba/Sq tumours, in addition to showing basal/squamous differentiation, were associated with *TP53* and *RB1* mutations. EGFR activity was increased in Ba/Sq tumours, which also expressed immune cell infiltration signature. NE-like tumours typically showed *TP53* and *RB1* inactivation at the same time. In a multivariate Cox regression model, the luminal tumours were not statistically different from each other with regard to overall survival. Patients with Ba/Sq tumours had a worse prognosis and those with NE-like tumours had the worst prognosis. Interestingly, the Ba/Sq type was overrepresented in women. Both Ba/Sq and NE-like tumours seemed to benefit from neoadjuvant chemotherapy.

The above-mentioned, consensus classification system has not been validated for non-muscle invasive tumours yet. For muscle-invasive disease however, this is a step towards further clinical trials leading to effective targeted therapy for bladder cancer.

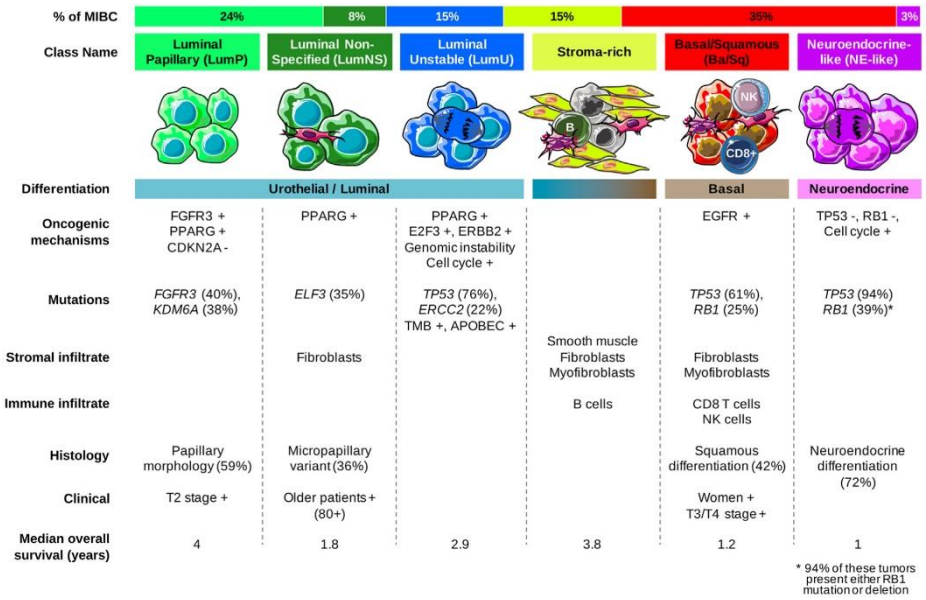

Figure 15. Summary of the main characteristics of the consensus classes of urothelial bladder cancers.
The figure is taken from Kamoun et al (75). This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

1.8 Prognosis and prognostic markers

The overall 5-year relative survival rate for bladder cancer is estimated at 77%. Tumour stage, usually according to the American Joint Committee on Cancer (AJCC) staging manual (the TNM system) (76), has great impact on choice of treatment and patient prognosis. For pTa and pTis the 5-year relative survival rate is around 90%. For pT1 and pT2 disease, the survival rates are 79% and 68%, respectively. pT4 disease has a low survival rate of only 12% (68, 77).

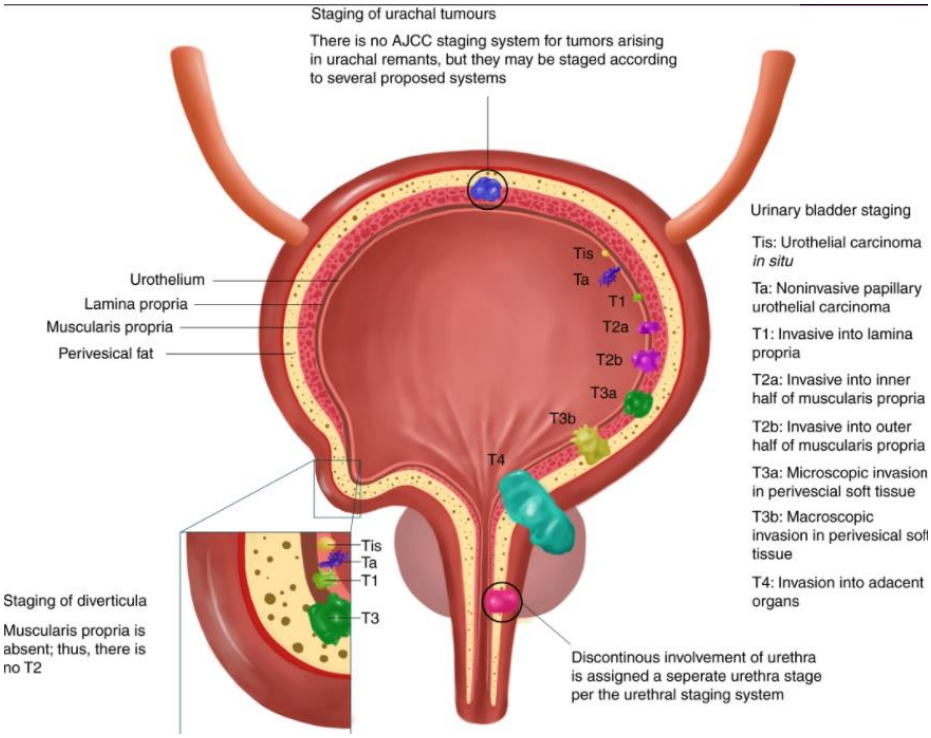

Figure 16. The current TNM staging system for bladder cancer, according to AJCC staging manual 8th edition.

Magers et al. Staging of bladder cancer. Histopathology 2018. Doi: 10.1111/his.13734. This article is distributed under the terms of the Creative Commons Attribution 3.0 International License (https://creativecommons.org/licenses/by-nc/3.0/).

Distinguishing between pTa and pT1 can be difficult. The material is often fragmented, and orientation might be difficult. Trauma artefacts, especially heating damage, makes the histopathological evaluation even harder. Tangential sectioning and benign von Brunn nests, as well as stroma reactions to tumour, can further complicate this evaluation. Bol et al. found a considerable inter-observer variability in staging Ta and T1 tumours. They found an 80% agreement among reviewers in a cohort of 130 patients with NMIBC. Of the 65 original T1 tumours, 35 were down-staged and 8 were up-staged by the reviewers (78).

There is no official consensus with regard to substaging of pT1 tumours. Despite this, separating patients having a small focus with lamina propria invasion, from those with extensive invasion, is strongly recommended, as these features are assumed to differentially impact patient prognosis (4). One suggested method is to distinguish pT1 tumours into micro-invasive (pT1m) and extensive-invasive (pT1e) categories (79). pT1m is then defined by a single focus of invasion, 0.5 mm or less in maximum extension. Tumours having more than one focus of infiltration or a single focus extending beyond 0.5 mm are defined as pT1e. van Rhijn et al. found a significantly higher progression-free survival for pT1m, with a 5-year and 10-year progression-free survival at 83% and 57%, respectively. This was compared to pT1e having corresponding 5-year and 10-year progression-free survival of 55% and 27%, respectively.

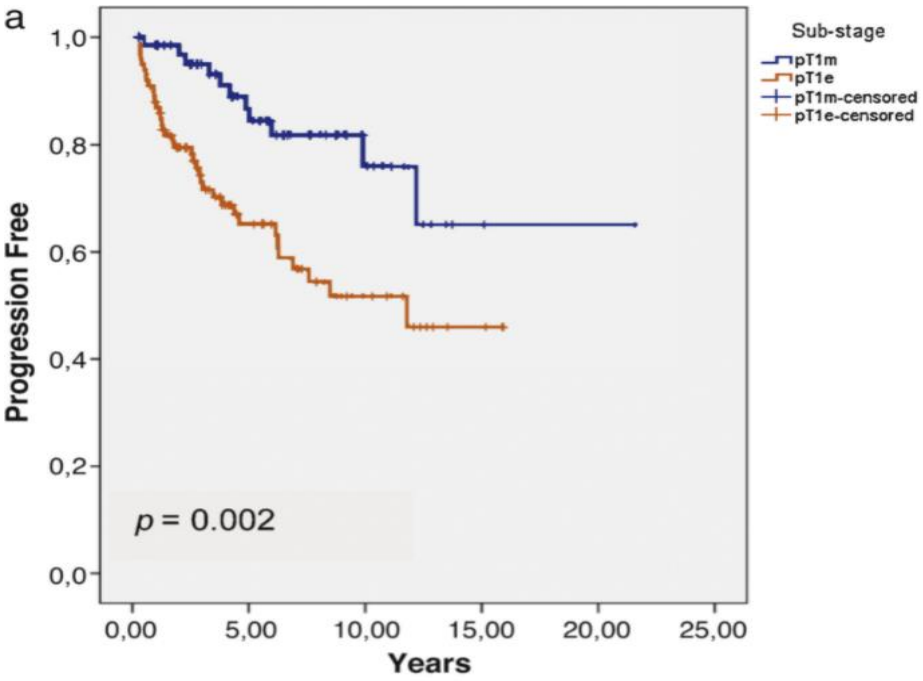

Figure 17. T1 substaging (pT1m vs pT1e) and progression-free survival.

Kaplan-Meier curves for progression-free survival comparing pT1m and pT1e in a cohort including 164 primary pT1 bladder cancer patients. Reprinted from European Urology, Vol 61, van Rhijn et al, A New and Highly Prognostic System to Discern T1 Bladder Cancer Substage, Copyright 2012, with permission from Elsevier.

Another proposal for substaging pT1 is to relate invasive border to the muscularis mucosae, and if not present, use the venous plexus as a substitute. This is challenging as the muscularis mucosae is discontinuous and sometimes absent. The location of the venous plexus varies, sometimes above and sometimes below the muscularis mucosae. Finally, measuring linear depth of invasion from the basal lamina is suggested, although different publications recommend different cut-off values (80).

As for many other malignancies, tumour grading, based on degree of anaplasia, is an important prognostic factor for urothelial

carcinomas. Grading is essential in risk stratification models (see below). For more detailed description of grading, see "1.4 WHO grading."

Several urothelial carcinomas of variant histology are associated with a worse prognosis, see section 1.5 "Variants of urothelial carcinoma" above. The micropapillary, plasmacytoid, and sarcomatoid variants in particular, are considered to be more clinically aggressive. The Giant cell, poorly differentiated and lipid-rich subtypes are also associated with poor clinical outcomes.

In 2006, the European Organization for Research and Treatment of Cancer (EORTC) presented a scoring model for determining the 1-year and 5-year risk of recurrence and progression to muscle-invasive disease, for NMIBC (81). This calculation model is based upon a retrospective analysis of 2596 NMIBC patients, and takes into account the WHO 1973 grading system, and the five most prognostic factors in the cohort regarding recurrence and progression (number of tumours, tumour size, prior recurrence rate, T-category, concomitant CIS).

On behalf of the European Association of Urology (EAU), Sylvester et al. published an updated scoring model in 2021 for risk of progression (EAU NMIBC 2021 scoring model) (31). The aim was to include the WHO 2004/2016 grading system, while still retaining the option of using the WHO 1973 grading system. In this retrospective study, 3401 patients were included. The number of tumours and tumour size were analysed as in the original scoring system from EORTC 2006. When using the WHO 2004/2016 grading system, PUNLMP and low grade carcinomas were grouped together. Age was included in the analysis and reintroduced in this scoring system. All prognostic variables were analysed using multivariable Cox regression and weighted again. The scoring model created by Sylvester et al. created four risk groups, now included in the EAU guidelines: low, intermediate, high, and very high risk.

Variable	WHO 2004/2016	WHO 1973
Age		
≤70 yr	0	0
>70 yr	55	32
Number of tumors		
Single	0	0
Multiple	50	32
Maximum diameter		
<3 cm	0	0
≥3 cm	65	43
Stage		
Ta	0	0
T1	80	52
Concomitant CIS		
No	0	0
Yes	100	58
WHO 1973 grade		
G1		0
G2		58
G3		100
WHO 2004/2016 grade		
LMP-low grade	0	
High grade	85	
Maximum total score	435	317

Risk group	Total progression score[a]	
	WHO 2004/2016	WHO 1973
Low risk	0–80	0–52
Intermediate risk[b]	85–150	58–133
High risk	165–305	142–233
Very high risk	315–435	242–317

Table 2. Weights for the prognostic factors in the EAU NMIBC 2021 scoring model.

The table shows weights for the prognostic factors used to calculate the progression score and the progression risk group. Patients with CIS ending in the intermediate risk group were reclassified as high risk. Reprinted from European Urology, Vol 79, Sylvester et al, European Association of Urology (EAU) Prognostic Factor Risk Groups for Non–muscle-invasive Bladder Cancer (NMIBC) Incorporating the WHO 2004/2016 and WHO 1973 Classification Systems for Grade: An Update from the EAU NMIBC Guidelines Panel, with permission from Elsevier.

EAU NMIBC Prognostic factor risk groups	
Low risk	➢ Primary, single Ta/T1, LG/ G1, < 3 cm and < 70 years (without CIS) ➢ Primary Ta, LG/ G1 and maximum one additional clinical risk factor (without CIS)
Intermediate risk	➢ Patients without CIS, not included in any other risk group
High risk	➢ T1, HG/G3 (without CIS), not included in the Very high risk group ➢ All CIS, not included in the Very high risk group ➢ Ta LG/ G2 or T1 G1, and 3 additional clinical risk factor (without CIS) ➢ Ta HG/ G3 or T1 LG, and 2 additional clinical risk factors (without CIS) ➢ T1 G2 and 1 additional clinical risk factor (without CIS)
Very high risk	➢ Ta HG/ G3, CIS and 3 additional clinical risk factors

	➢ T1 G2, CIS and at least 2 additional clinical risk factors ➢ T1 HG/ G3, CIS and at least 1 additional clinical risk factor ➢ T1 HG/ G3, 3 additional clinical risk factors (without CIS)

Table 3. The clinical composition of the EAU NMIBC prognostic factor risk groups

The table helps to classify patients into the right EAU NMIBC prognostic factor risk group. Additional clinical risk factors are age > 70, multiple tumours, and tumour size >3 cm. Those with CIS in prostatic urethra, some variant histology (micropapillary, plasmacytoid, sarcomatoid and small cell neuroendocrine carcinoma) as well as those with pT1 and lympho-vascular invasion, should be classified as very high risk. Patients with recurrent tumours are also not included in the table and should be classified as intermediate, high, or very high risk depending on their other prognostic factors (31).

Before the introduction of EAU NMIBC prognostic factor risk groups in 2021, the high risk group included a subgroup called "highest risk". This term is retained in the EAU guidelines, partly overlapping with the new very high risk group (28). Patients defined as having "highest risk" of progression includes:

- T1 HG/G3 and CIS in bladder and/ or urethra
- T1 HG/G3 and/ or multiple and/ or large (> 3 cm) and/ or recurrent
- T1 HG/G3 with aggressive variant histology
- T1 HG/G3 with lymphovascular invasion

Lympho-vascular invasion (LVI) is an important way of systemic spread and metastasis from malignant disease. LVI in urothelial carcinoma is not included in the most frequently used prognostic calculators, but in the EAU NMIBC 2021 scoring model, pT1 tumours with LVI are suggested to be classified as very high risk. A meta-analysis by Kim et al. found that LVI in transurethral resection of bladder tumour (TURBT) was associated with pathologic upstaging (OR 2.2) (82). They

also found significant influence on both recurrence-free and progression-free survival (HR 1.47, 95% CI 1.24-1.74, and HR 2.28, 95% CI 1.45-3.58, respectively), as well as disease specific survival (HR 1.35, 95% CI 1.01-1.81).

Several other biomarkers, although not in routine clinical use, have shown prognostic utility as well. Mitotic frequency and location are part of the standard evaluation when grading urothelial carcinomas. As noted earlier, histological grading has raised significant concerns regarding low reproducibility. This has led to a search for more objectively quantifiable and reproducible prognostic markers. Formalized and validated methods for quantification of mitosis, like Mitotic Activity Index (MAI), is such a variable. MAI is calculated by counting obvious mitosis at x 400 magnification in consecutive fields of vision in a total area of 1.59 mm^2 (83). Other proliferation markers involve the use of immunohistochemistry to detect antigens involved in cell cycle. The antigen Ki67 is a nuclear antigen present during most of the cell cycle (G1, S, G2 and M phase). It is not present in non-proliferating cells (G0). Phosphohistone H3 (PPH3) is a nuclear antigen only present in late G2 and M phase of the cell cycle. PPH3 is therefore more specific for measuring mitotic activity. Bol et al. analysed the prognostic value of proliferation in NMIBC (84). In a univariate analysis, they found the strongest prognostic predictors for progression to be Ki67, MAI, and the mean area of the ten largest nuclei (MNA10). These markers performed better than WHO grading and other more traditional prognostic markers. Ki67 and MNA10 were calculated by the QPRODIT version 6.1 image analysis system, making them reproducible. In a multivariate analysis the strongest predictors were the combinations MNA10/ Ki67 and MNA10/ MAI (Fig 18). Also van Rhijn et al. found Ki67 to be a significant prognostic marker (85). They suggested a molecular grading system combining Ki67 and *FGFR3* mutation analysis. In a cohort consisting of pT1 tumours only (n=309), Ki67 was the only marker predicting progression-free survival in a multivariable

analysis (86). Mangrud et al. investigated all the proliferation markers (MAI, Ki67, and PPH3) and compared them to WHO73 and WHO 2004/2016 grades (87). All three proliferation markers and both the WHO 1973 and WHO 2004/2016 grading systems were significantly prognostic for progression. In their study, MAI appeared to be the best prognostic marker with a HR = 16.5 (95% CI 3.6–75.3).

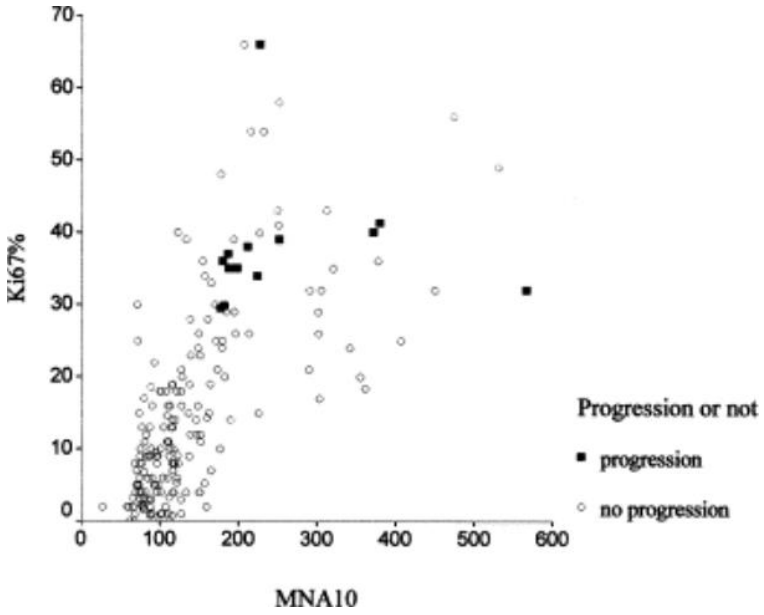

Figure 18. MNA10, Ki67 and progression in high-risk patients.

Scatter plot of MNA10 and Ki67 in high-risk patients (n=74), in a cohort from Bol et al. All progression cases had MNA10 > 170 μm^2 and Ki67 > 25 %. Reprinted from Urology, vol 60, Bol et al, Prognostic value of proliferative activity and nuclear morphometry for progression in TaT1 urothelial cell carcinomas of the urinary bladder, Copyright 2002, with permission from Elsevier.

Several other groups have investigated immunohistochemical markers and prognosis as well. Cytokeratin 20 (CK20) is a cytoskeleton-associated intermediate protein filament found in epithelial cells, mainly in the genitourinary and gastrointestinal tract. It has been used as a

marker for urothelial differentiation and lately, as a surrogate marker in molecular classification of bladder cancer. For a long time, it has been used by surgical pathologists to separate reactive urothelial changes from CIS (88). In normal urothelium, CK20 is expressed by umbrella cells. In CIS, the expression is extended to deeper layers of the urothelium (89, 90). In papillary urothelial carcinomas, abnormal CK20 expression has not been shown to predict progression. A few publications have shown an association between abnormal CK20 and recurrence (86, 91, 92). Although not found significant for prognosis, Desai et al. found a significantly higher proportion of tumours with abnormal CK20 expression among high-grade tumours compared to low-grade tumours (91). Notably, 30% of high-grade tumours did not show abnormal CK20 reaction. It is widely agreed that CK20 has no role in grading of papillary urothelial carcinomas.

As *TP53* mutations in urothelial carcinoma are associated with aggressive muscle-invasive tumours, immunohistochemistry for the p53 nuclear antigen has been investigated as a possible prognostic marker. Mutated p53 has a much longer half-life, as wild-type p53 is quite unstable. This causes mutated p53 to accumulate in the nucleus. Strong nuclear positivity for p53 on immunohistochemistry therefore indicates a possible mutation. This method is routinely used for determining the p53 mutation status in serous ovarian carcinoma. Several publications confirm the use of p53 immunohistochemical staining as an appropriate surrogate marker also in urothelial carcinomas (93, 94). There are some publications showing worse outcomes for tumours with p53 overexpression (95, 96), but several others failed to find any prognostic value (84, 86).

Even though evidence is increasing, molecular characteristics and molecular subclassification of bladder cancer are not in routine clinical use for prognostic purposes yet. Non-muscle invasive and muscle invasive tumours have different molecular characteristics (see "1.6 Molecular alterations in bladder cancer"). Different proposals for a

molecular classification system have defined tumour subgroups with very different outcomes (see "1.7 Molecular classification of urothelial carcinoma"). In the Lund University Classification system, the Urobasal A tumours had the best prognosis in disease-specific survival, while the genomically unstable and the infiltrated tumours had an intermediate prognosis, and Urobasal B and SCC-like had the worst prognosis (66). For the classification systems from UNC and MD Anderson, the basal-like and the basal tumours, respectively, were both associated with a shorter disease-specific survival (68, 72). Moreover, both these tumour subgroups showed features overlapping with the features of Urobasal B and SCC-like tumours. In the proposal from Hedegaard et al. based on NMIBC only, patients with class 1 tumours had the best prognosis, those with class 3 tumours had an intermediate prognosis, and patients with class 2 tumours had the worst prognosis (shortest progression-free survival). Both class 1 and 2 tumours had luminal characteristics, while class 3 tumours showed basal features. Class 1 tumours best matched Urobasal A, while class 2 tumours had characteristics similar to those of the Genomically unstable subgroup. In the most recent consensus classification, Ba/Sq and NE-like had the worst prognoses (75).

1.9 Immune response on tumour

In recent years, the tumour microenvironment has gained increasing interest especially after the introduction of immune checkpoint inhibitors (ICI). The immune response against tumour cells has been studied for many years, both as a prognostic and as a predictive marker. Tumour cells gain mutations and present new and not "self-recognizable" antigens on the cell surface, initiating an immune response. Tumour cells and locally situated immune cells attract lymphocytes via cytokines and chemokines. The most important immune cells that fight against tumour cells are T-cells (CD3+ cells). Several subpopulations of T-cells can be identified and the immune response depend on a finely-tuned balance

between the different subsets of cells. When studying the presence of tumour infiltrating lymphocytes (TILs) in the tumour microenvironment, this heterogeneity of immune cells has to be taken into consideration. The most frequently described immune cells in publications on TILs are cytotoxic T cells (CD8+) and T helper cells (CD4+), the latter including an important subgroup called T regulatory cells (Tregs, FOXP3+/CD25+) (97).

Cytotoxic CD8+ T cells are effector lymphocytes involved in destroying tumour cells. Their presence correlates with a better prognosis in several malignancies (98, 99). For urothelial carcinoma however, several publications show conflicting results: Lui et al. investigated TILs in both MIBC (n=49) and NMIBC (n=53) and found increased overall survival in the group of high CD8+ for NMIBC only (100). Sharma et al. found better disease-free survival in MIBC with high CD8+ expression, but lack of significance for NMIBC (101). The counting methods employed in these studies were also different: Liu et al. counted cells in the invasive front, while Sharma et al. counted both in the stroma and in the tumour nests. Others also looked at muscle-invasive cancers and found a better disease-free and overall survival in those with high CD8+ count (102, 103). Wahlin et al. found increased time to recurrence in CD8+-high MIBC after cystectomy (104). Zhang et al. analysed CD8+ lymphocytes in both organ-confined (n=75) and non-organ confined (n=51) bladder cancer (105). In organ-confined bladder cancer, CD8+ was significant associated with poorer overall survival, while in non-organ confined disease, CD8+ expression correlated with better overall survival. Finally, Horn et al. could not find any correlation between CD8+ TILs and outcome in a cohort of 149 invasive bladder cancers treated with cystectomy (106). There are also a few publications on CD4+ lymphocytes in bladder cancer. Zhang et al. presented a cohort including 131 NMIBC patients, in which CD4+ was associated with reduced overall survival (107).

Regulatory T-cells (Tregs) are important in maintaining self-tolerance in the human body. In the tumour microenvironment, these cells generally decrease the anti-tumour immune response. Their exact mechanism is still poorly understood. It seems like Tregs are a heterogenous group of cells, either changing phenotype dependent on context or being distinct determined subtypes (108). Tregs in the tumour microenvironment are thought to limit the effect of ICI. Their presence has been shown to correlate with poorer prognosis in several malignancies (109). Existing publications regarding Tregs in bladder cancer show conflicting results. Miyake et al. found an association between high Treg count and reduced recurrence-free survival in NMIBC after receiving BCG instillation (110). Winerdal et al. found longer progression-free survival in tumours with high Treg counts in a cohort of invasive bladder cancer patients who underwent cystectomy, although the included number of patients was low (n=37) (111). Another study by Parodi et al. showed reduced recurrence risk when the intra-tumoral ratio of Effector T-cell/Treg was greater than 1. This study included all stages, but also suffered from a small sample size (n=28) (112). As mentioned, Horn et al. did not find any significant association between CD8+ and prognosis, but while analysing the ratio of FOXP3+ cells/CD8+ cells, they found a significantly shorter overall survival and time to cancer-specific death in those with higher FOXP3+ cells/ CD8+ ratios (106).

A recent review by Miyake et al., regarding the prognostic value of TILs in urothelial carcinoma, states that a conclusion cannot be drawn yet (113). Existing publications in the field are difficult to compare as the counting of immune cells are performed in different locations, some in both tumour and stroma, some in tumour only, and yet others in the invasive front. Counting methods and scoring systems also differ, making head-to-head comparisons impossible. In addition, the patient cohorts differ widely in stage: some studies include NMIBC only while others include both NMIBC and MIBC. For a final conclusion regarding

this topic, more validation studies in more homogenous and standardized patient cohorts are needed.

1.10 Symptoms and diagnostics

The cardinal clinical sign of bladder cancer is hematuria, either micro- or macrohematuria. Bladder cancer can also present with irritative symptoms, like pollakisuria, dysuria, and urgency. Irritative symptoms are mostly associated with CIS. Voiding problems because of tumour blockage may also happen.

With a few exceptions, macroscopic hematuria should initiate further investigations to rule out cancer in the urinary tract (28, 29). The diagnostic approach for patients with microscopic hematuria depends on age, risk factors, and symptoms. The main investigation to look for tumours in the urinary bladder is cystoscopy. Tissue samples are taken from tumours and other suspicious lesions in the mucosa. This is the basis for a histological diagnosis. When a high grade or invasive tumour is detected, computer tomography (CT) with contrast is indicated. Such a CT should be taken from the urinary tract and thorax, because a CT can reveal tumours in the upper urinary tract in case no tumour is detected in the bladder on cystoscopy. Magnetic resonance tomography (MRI) is not routine, but can replace CT on special indications. MRI is more suitable for local tumour staging in the pelvis.

Urinary cytology can be a useful supplement to cystoscopy if no tumour is found. Cytology is also the method of choice if one suspects the presence of CIS. This method has a high sensitivity for detecting high grade lesions, but a low sensitivity for detecting low grade lesions. Different kinds of molecular tests using urine have been developed. They generally have a higher sensitivity than urinary cytology, but their

specificity is often lower. None of them are currently in routinely clinical use in Norway.

1.11 Current treatment guidelines for NMIBC

NMIBC is normally treated with trans-urethral resection of the bladder (TURB) in order to remove all the tumour tissue. In the case of uncertainty about completeness of the TURB, the presence of T1-tumors or the lack of muscularis propria in the resection, a re-TURB is recommended 2 - 6 weeks later (29). A significant proportion of patients will have residual tumour in the re-TURB (114), either because of incomplete first resection or recurrence of a new tumour. A single post-operative instillation of chemotherapy in the bladder, like Mitomycin C, is recommended for all NMIBC as this has shown to reduce the recurrence rate (OR=0.61) (115).

Because of the high recurrence rate in NMIBC, adjuvant therapy is usually necessary. Treatment decisions related to adjuvant therapy are generally based on the previously described risk-groups in the EAU guidelines. In recurrent and small low-grade tumours (intermediate risk), regular chemotherapy instillations according to protocols can be an option. Usually, six weekly instillations of Mitomycin C are administered, followed by one instillation per month. Duration of this treatment is up to one year. Such regimes can also replace Bacillus Calmette-Guerin (BCG) instillations in cases with intolerable adverse events. BCG is regarded as first option in intermediate and high-risk patients after TURB.

Radical cystectomy is suggested for NMIBC in those categorized as very high risk (see “1.8 Prognosis and prognostic markers”), or those included in the older term “highest risk”. According to the EAU guidelines radical cystectomy is also strongly recommended for BCG

unresponsive tumours, broadly defined as any recurrent high grade tumour during or after BCG therapy (28). Treatment options, regarding risks and benefits, should always be discussed with the patient.

1.12 BCG instillation in bladder cancer

The instillation of BCG into the urinary bladder is an anti-cancer immunotherapy, first published by Morales et al. in 1976 (116). An attenuated live strain of Mycobacterium bovis is suspended in 50 ml saline and injected, via a catheter, into the urinary bladder. The BCG treatment starts with an induction schedule consisting of one weekly instillation for six weeks. This is followed by a maintenance schedule, usually consisting of repeated cycles of one weekly instillation for three weeks for a period of 1–3 years. As BCG instillations often are accompanied by bothersome side effects, the duration of maintenance treatment has been debated. A randomised controlled trial by the EORTC, published in 2013 (117), compared maintenance therapy for 1 and 3 years. In the intermediate risk group there were no differences, but in the high-risk group a 3-year schedule reduced recurrence risk compared to a one-year schedule (HR=1.61, 95% CI: 1.13–2.30). The two additional years of instillations did not significantly affect risk of progression.

Several meta-analyses have shown that BCG instillations reduce the risk of both recurrence and progression in NMIBC patients treated by TUR-BT (118-121). Malmström et al. performed a meta-analysis including 2820 NMIBC patients, comparing recurrence rates in those treated with regular Mitomycin C instillations and those receiving BCG therapy with maintenance (118). They found a 32% risk reduction after BCG with maintenance compared to Mitomycin C. Two other meta-analyses investigating BCG effect on progression, found significantly reduced progression risk in those receiving BCG (OR=0.66; 95% CI:

0.47–0.94, and OR 0.73; p=0.001). However, the effect was only significant after BCG maintenance therapy for at least one year (120, 121). Several different BCG strains exist and little is known about whether there are differences in efficiency among the strains. However, a meta-analysis from 2017 by Boehm et al. found no significant superiority of any strain compared to the others (122). They also demonstrated a significantly reduced risk of recurrence compared to chemotherapy instillation.

Although the BCG-therapy has been used for decades, the exact immunological mechanism behind the treatment effect is not yet fully understood (123, 124). BCG in the bladder lumen will attach to urothelial tumour cells, probably with help of the glycoprotein fibronectin in the extracellular matrix. Normal urothelial cells are to some extent protected from BCG by a negatively charged layer of glycosaminoglycans, covering the mucosa. Following attachment, BCG enters the cells, more readily the poorly differentiated tumour cells (125). As tissue destruction ensues, tissue macrophages will also incorporate BCG. Both tumour cells and the professional antigen presenting cells will process antigens and present them through MHC molecules on the cell surface. The infected cells will release cytokines attracting neutrophils and mononuclear inflammatory cells (126). Among these are CD4+ and CD8+ T-cells. The presence of cytokines generally seen in a Th1-response, like IL-2, IL-12, IFN-γ and TNF, are associated with BCG-responsiveness, while those involved in a Th2-response are associated with non-responsiveness to BCG (127). It is not known whether the BCG induced inflammation involves a directly specific anti-tumour activity or whether the general inflammation is responsible for the anti-tumour mechanism. Both CD4+ and CD8+ T-cells are assumed to be important in the initiated immune response (128). As previously mentioned, Miyake et al. found an association between high Treg count and reduced recurrence-free survival in NMIBC after receiving BCG instillation (110). Pichler et al. found prolonged recurrence-free survival in patients

with high CD4+ count in the tumour microenvironment. Again, a high count of Tregs was inversely correlated with recurrence-free survival. They concluded that the tumour microenvironment is important for the therapeutic response to BCG treatment (129).

BCG-instillations are frequently accompanied with side effects, most of them mild irritative symptoms such as dysuria, frequency of voiding and hematuria. In a randomised controlled phase 3 trial, on behalf of EORTC, Brausi et al. registered side effects in 69.5% of the cohort (n=1316). Local side effects were reported in 62.8% while 30.6% reported systemic side effects (130). Systemic side effects included fever (8%) and general malaise (15.5%). In this publication 7.8% stopped treatment because of side effects. In another study by Lamn et al., only 16% completed the maintenance therapy schedule due to side effects (131). BCG infection is also a potential risk, as BCG consists of living attenuated Mycobacterium bovis; luckily the risk is only small [ca 1%, and mostly limited to the genitourinary tract (132)]. Among those suffering from BCG infection 30.3% had lung infections. To reduce the risk for BCG-infection, instillation should not be performed after traumatic catheterization or the first two weeks after TURB, as the urothelial barrier is broken. It should also be used with caution in immunocompromised patients. Adverse reactions to BCG is most frequent during induction and the first six months of maintenance therapy (133).

1.13 NMIBC follow-up

Patients with NMIBC have a high recurrence risk. As mentioned under "1.2 Epidemiology and etiology of urothelial bladder cancer" 50–70% will have recurrence and 15–25% will progress to MIBC. Early detection of recurrent tumours is important as a new TURB and adjuvant instillations can be sufficient treatment if tumours are discovered before

they start to invade the m. propria. Cystoscopy is the gold standard in detecting new tumours in the bladder. For all patients the first control cystoscopy is recommended 3 months after TURB (28, 29). New tumours at this point is an independent prognostic marker for future recurrence and progression risk (134). Further surveillance with repeated cystoscopies depend on EAU risk group. The extensive follow-up regime in NMIBC patients, frequently life-long, explains the high costs for the health care systems. The following table outlines the Norwegian follow-up guidelines.

EAU NMIBC prognostic factor risk group	Cystoscopy interval after TURB
Low risk	• Cystoscopy 3 months after TURB • Cystoscopy 12 months after TURB • Cystoscopy once a year for a minimum of 5 years
Intermediate risk	• Cystoscopy 3 months after TURB • Cystoscopy every 6 months for 2 years • Cystoscopy once a year from 3 to 10 years
High risk	• Cystoscopy + cytology 3 months after TURB • Cystoscopy + cytology every 3 months for 2 years • Cystoscopy + cytology every 6 months until the end of the 5. year • Cystoscopy + cytology yearly the rest of the lifespan

Table 4. EAU NMIBC prognostic factor risk group and the corresponding follow-up regimes.

The table is taken from the national Norwegian guidelines (29).

2 Aims of the thesis

Up to 70% of NMIBC will have recurrence, and up to 25% will experience progression to MIBC. Extensive follow-up, with high costs to the society and representing an additional burden to the patients, is considered necessary. Both treatment decisions and follow-up regimes are mainly decided based upon WHO grade and TNM stage. It is well known that WHO grading has issues with low reproducibility and that pathological staging can be challenging (35, 40, 78). This thesis aims to contribute to improved grading and to find new and better prognostic/ predictive histopathological biomarkers for non-muscle invasive papillary urothelial carcinomas.

Aims for Paper I

Investigate the reproducibility and the prognostic value of the individual histopathological features making up the WHO grading systems.

Aims for Paper II

To study the prognostic value of subsets of tumour infiltrating lymphocytes and plasma cells, and compare them with previously investigated proliferation markers.

Aims for Paper III

Study the prognostic and predictive value of T1 substaging in bladder cancer patients who fulfilled BCG induction treatment.

Aims for Paper IV

Investigate the prognostic value of CK20 and p53 immunohistochemistry, and compare them with previously investigated proliferation markers.

3 Methodology

3.1 Patient material

The patient cohorts for papers I, II, and IV consists of patients diagnosed with a primary non-muscle invasive papillary urothelial carcinoma in the urinary bladder (NMIBC) at department of pathology at Stavanger University Hospital. Paper I and II includes patients diagnosed between 01.01.2002 and 01.01.2007. Paper IV is based on an extended cohort, involving patients diagnosed between 01.01.2002 and 01.01.2011. Paper III includes primary high risk NMIBC patients from three different Dutch hospitals in the period 2000–2017 (Erasmus MC; Franciscus Gasthuis & Vlietland and Amphia) and Stavanger University Hospital 2002–2010.

Before the start of the studies, the research was approved by The Norwegian Regional Ethical Committee (REK Vest, #106/09). Informed consent was not obtained from the patients, as the tissues already were resected for diagnostic and treatment purposes. All patients were offered the opportunity to reserve themselves from participation before the start of the study. This was in agreement with conditions stated by REK Vest. In addition, Paper III was approved by the Erasmus MC Medical Ethics Committee (MEC-2018-1097).

Paper I:

All patients diagnosed with a primary NMIBC at Department of Pathology at Stavanger University Hospital, were identified from the beginning of 2002 to the end of 2006 (n=249). Clinical information was extracted from the medical records at Stavanger University Hospital. All patients with a history of extra-vesical urothelial carcinoma were excluded from the database. Nine patients appeared to have either MIBC or metastatic disease at time of diagnosis, and were excluded. Further 11 was lost to follow-up. Finally, 36 cases were excluded because of insufficient material, thermal damage or poor tissue quality. The final

cohort included 185 patients. Among these 13 (7%) progressed to a higher stage. All those with progression and 25 randomly selected individuals without progression were selected from the original cohort. In total 38 cases were included. The selected 25 without progression were not statistically different regarding age, sex, grade, recurrence or follow-up time, compared to the remaining 172 cases not selected.

Paper II:

In paper II we used the same cohort as for paper I, consisting of 185 cases. Two cases had insufficient material for immunohistochemical analysis and were excluded; thus, the remaining 183 cases were used for progression analysis. For analysis of recurrence, the criteria were slightly different. We wanted to investigate the tendency for a new tumour to develop locally in the urinary bladder. In the recurrence cohort the patients were followed until last registered cystoscopy or cystectomy. Minimum follow-up time was set to 3 months. Consequently 6 more cases were lost to follow-up for recurrence analysis, as they either underwent cystectomy or were not followed by cystoscopies (most of them because of comorbidity). This is why the paper distinguished between a progression cohort including 183 cases, and a recurrence cohort including 177 patients.

Paper III:

Originally, the cohort from the four different hospitals consisted of 535 primary high-risk NMIBC patients, all of them receiving at least five induction instillations of BCG treatment. After a central review, 26 cases were excluded either because of poor specimen quality or because of changed risk group. Among the remaining 509 cases, 264 were T1 disease, eligible for sub-staging analysis. From the extended Stavanger cohort (2002–2010), 63 cases met the inclusion criteria and 39 of these cases were T1 disease suitable for analysis of T1 sub-staging.

Paper IV:
After Paper I and II, our database was extended. By using the same approach as described for Paper I, we extended the database of primary NMIBC, diagnosed at Stavanger University Hospital, to the end of 2010. After excluding an additional four cases with extra-vesical urothelial carcinoma, three cases with insufficient quality, and one because of short follow-up time, we ended up with 349 cases. Additionally, in the period 2002–2010, 7 cases of primary carcinoma in situ (pTis) were identified. Among the 349 cases of papillary urothelial carcinoma, 26 (7.4%) progressed to a higher stage. There was one case with progression among the seven cases with pTis only. As for paper II, in paper IV we set stricter follow-up criteria for recurrence analysis. The recurrence cohort in paper IV was limited to 337 cases.

3.2 Histology

All analyses were performed on formalin fixed paraffin embedded (FFPE) tumour tissue, archived at Department of Pathology at Stavanger University Hospital. Tumour tissue was originally received for diagnostic purposes. FFPE blocks were cut into 4 µm thick sections and stained with Haematoxylin Eosin Saffron (HES). HES-stained sections made the basis for diagnosis, pathologic staging, and grading, as well as mitotic counting. All slides from all the specimens, both primary and recurrent tumours, were reviewed, by two pathologists, to confirm diagnosis, grade (both WHO1973 and WHO2004/2016) and stage.

Cases included in Paper I, were additionally given a grade for each of the histopathological features making up the WHO grading systems. These microscopic features were extracted from textbooks in urological pathology. Three pathologists with experience in urological pathology performed feature grading independently, without knowledge of previous diagnostics, prognostic information, or the other pathologists' evaluation. Finally, a consensus grade for each feature in

each case was created using a multi-headed microscope. The histopathological features and the descriptions for each of them are shown in Table 5.

	WHO73			WHO04	
	Grade 1	**Grade 2**	**Grade 3**	**Low grade**	**High grade**
Architecture					
Papillae	Delicate	Varies	Broad, varies	Slender	Broad
Superficial layer	Usually present	Usually present	Partially or completely lost	Usually present	Partially or completely lost
Papillary fusion	Some	Varies	Common	Some	Varies
Nuclear arrangement					
Polarity	Preserved	Moderate loss	Lost	Preserved, mod. loss	Lost
Maturation	Normal	Some	Lost	Preserved, mod. loss	Lost
Cohesion	Normal	Some	Lost	Some	Lost
Proliferation					
Mitotic figures	Rare, basal	Lower half	Common, atypical	Rare	Common
Nuclear atypia					
Nuclear enlargement	Mild	Mild	Varies	Mild	Varies
Nuclear shape	Uniform	Moderate variation	Pleomorphic	Moderate variation	Pleomorphic
Nuclear hyperchromasia	Mild	Moderate	Varies	Mild to moderate	Varies
Chromatin pattern	Finely granular	Granular	Coarse	Fine	Coarse
Nucleoli	Occasional	Occasional	Common	Occasional	Common
Giant nuclei	No	No	Yes	No	Yes

Table 5. Histopathological features and their descriptions according to WHO grade.

Showing histopathological features and the corresponding description for each WHO grade, extracted from textbooks in urological pathology.

3.3 Immunohistochemistry

Immunohistochemistry was performed to investigate the presence of specific antigens in the tumour tissue. For paper II, Ki-67, PPH3, CD4, CD8, CD25 and CD138 were investigated. For paper IV CK20 and p53 were included as well. Consecutive 4 µm thick sections for immunohistochemistry were mounted onto Superfrost Plus slides. The slides were dried overnight, deparaffinised with xylene and then rehydrated with solutions of decreasing alcohol concentrations. Then a heat-mediated antigen retrieval system, using TRIS (10 mM) - EDTA (1 mM) antigen retrieval buffer (pH 9), was applied on the sections. Endogenous peroxidase activity was inactivated by incubation in a peroxidase-blocking reagent (DAKO S2001) for 10 minutes. Immunostaining was performed using an autostainer (DAKO, Glostrup, Denmark). Antibody diluents for primary and secondary antibodies were used, and the immune-complex was visualized by 3,3`-diaminobenzidine (DAB) chromogen with haematoxylin as a counterstain.

3.4 Mitotic activity index (MAI)

Calculation of MAI was performed according to the previously published protocol (83, 84). The least differentiated area on the HES-stained slide was marked. In the selected area, x fields of vision (FOV) with a 40x objective were scanned for mitotic figures until a total area of 1.59 mm^2 was reached. All obvious mitoses were counted (prophase, metaphase, anaphase, and telophase). Each FOV included at least 75% tumour tissue.

Counting of immunohistochemically PPH3 (phosphohistone H3) positive cells was performed following the same MAI procedure. For this count, the area giving the impression of highest number of PPH3 positive cells on low magnification was marked.

3.5 Quantitative image analysis

To minimalize interobserver variability, the semi-automatic interactive computerized system QPRODIT, version 6.1, (Leica, Cambridge, UK) was used to count Ki67 and immune cell markers (CD4, CD8, CD25, and CD138). QPRODIT was also used for quantitation of MNA10 (mean nuclear area of the ten largest nuclei).

The area of interest marked by the observer is electronically demarked. For calculation of Ki67, the system randomly selects 200–300 FOV in the demarcated area. The FOVs corresponds to 400x magnification (40 x objective, numerical aperture 0.75) and is shown on a monitor. In each FOV a 2-line grid is projected, and one standard endpoint was used for counting. As this endpoint was projected over a tumour cell it was evaluated as positive or negative by the observer. Based on this procedure a percentage of Ki-67 positive cells was achieved.

For calculation of immune cell markers, 150 FOVs were randomly selected in the electronically demarcated area. For this procedure, a 6-line grid was used. Five standard endpoints were evaluated for each FOV. Each cell pointed out was evaluated as positive or negative, and a percentage of positive cells for that specific immune cell marker was achieved.

For MNA10, the demarcated area was scanned manually at low magnification for the largest nuclei. The 20 subjectively largest nuclei were projected on the monitor on high magnification (1000 x) and outlined by the mouse for automatic area measurement. The mean nuclear area of the 10 largest nuclei was calculated (83).

3.6 Digital image analysis

For counting p53-positive nuclei, a fully automatic digital image analysis software, Visiopharm®, was used. Whole-slide images with p53 immunohistochemical staining were scanned at 400 x magnification, using a 3D Histech Panoramic Scan II (3DHistech, Budapest, Hungary). The images were uploaded to the image analysis software, which performed automatic tissue and nuclei detection. A cut-off value indicating a positive nucleus was established. The pixel value for DAB was set to 80, corresponding to a strong and obvious positive nuclear staining. The application identified three hotspots of p53-positive nuclei on the whole-slide image, each with an area of 3.5 mm^2. The average percentage of p53-positive nuclei in the three hotspots was calculated. The threshold for a p53-positive tumour was set to > 15%, matching the 75-percentile in the cohort and thresholds used in other publications on the topic (95, 96).

3.7 Immunoreactive score (IRS)

Standardized scoring systems for immunohistochemical markers are lacking. There are a few semi-quantitative scoring systems, translating subjective impression by the pathologist into data amenable to statistical analysis. A common characteristic of these semi-quantitative systems is that they combine several ordinal variables into a single score (135). IRS is one of the few semi-quantitative systems widely accepted and recommended (136). In IRS, a score for percentage positive tumour cells (0–4) is multiplied with a score for staining intensity (0-3). The IRS will accordingly range from 0–12.

Percent positive tumour cells (X)	**Staining intensity (Y)**
0: No positive cells	0: No staining
1: < 10 %	1: Mild
2: 10 - 50 %	2: Moderate
3: 51 – 80 %	3: Intense
4: > 80 %	

Table 6. The immunoreactive score (IRS).

IRS is the score for percentage positive tumour cells (X) multiplied by the score for staining intensity (Y), IRS = X x Y.

In Paper IV IRS was used to assess immunohistochemistry for CK20 in tumour cells. As no standard cut-off value for positive vs. negative exists, we considered IRS > 3 as positive (based on a median IRS score = 4). Two individuals performed IRS on all the cases independently. In cases with a difference in IRS > 3, a consensus score was achieved using a multi-head microscope.

3.8 Statistical analyses

Paper I:

The statistical analyses in this paper were performed in R version 3.4.0. We wanted to investigate the reproducibility and the prognostic value of all the microscopic histopathological features behind grading. In this study we used three raters. For each feature there were two or three ordinal categories. The well-known Cohens Kappa is neither suitable for ordinal data, nor is it appropriate for more than two raters. For handling ordinal data and three raters, we found Gwet's Agreement Coefficient (AC) 1/2 most suitable (137). Gwet's AC1 was used for features with

two categories and the weighted Gwet`s AC2 was used for features with three categories. We also calculated the weighted Fleiss' Kappa for reference, but emphasized Gwet`s AC1/2 as Fleiss' Kappa is vulnerable for skewed marginal distributions. The coefficients express the proportion of agreement, after agreement by chance is removed. Several benchmark scales have been developed for guidance. They were originally meant for kappa values, but are currently used for guidance with other agreement coefficients as well. According to Altman`s kappa benchmark scale, a coefficient > 0.60 represents good agreement (138).

Kappa value	Strength of Agreement
< 0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
> 0.80	Very Good

Table 7. Altman`s Kappa benchmark scale
The scale was originally meant for kappa values, but is used for other agreement coefficients as well (138).

Prognostic value was estimated by the Area Under Curve (AUC) of the receiver operating characteristics (ROC) function. We calculated 95% confidence intervals. For a feature to be statistically significant for progression, the 95% confidence interval for the AUC could not overlap 0.5.

Paper II:

For statistical analyses, SPSS version 21 (SPSS Inc., Chicago, IL, USA), and MedCalc Statistical Software version 19.1 (MedCalc Software BV, Ostend, Belgium) were used. All continuous variables were dichotomized. For the proliferation markers (Ki67, PPH3, MAI) and MNA10, previously published thresholds were used (84, 87). For the immune cell markers (CD4, CD8, CD25 and CD138), the median value was used. Endpoints in this study were recurrence-free and progression-

free survival. Differences in survival distributions between the groups were investigated with Log Rank test, and Kaplan Meier curves were generated. Further analyses of the variables were performed using univariate and multivariate Cox regression analyses. Hazard ratio with 95% confidence interval was calculated.

Paper III:

For this paper, the statistical analyses were performed in R version 4.0 (Vienna, Austria). Endpoints in this study were High grade recurrence-free survival, progression-free survival and disease-specific survival. The non-parametric test, Chi Square test, was performed to see if T1-substaging was associated with BCG-failure. Survival distribution analyses with Log Rank test were performed, and Kaplan Meier curves were generated. Clinical and pathological parameters were further analysed with both uni- and multivariate Cox regression analyses.

Paper IV:

All the statistical analysis in this paper were performed by using IBM SPSS Statistics 26, (IBM Corp, Armonk, NY). Continuous variables were dichotomised in the same manner as in Paper II, using previously published thresholds. Recurrence-free survival and progression-free survival were the endpoints. Log Rank test investigated the differences in survival distribution for the independent categorical variables, and Kaplan Meier curves were generated. Finally we performed univariate and multivariate Cox regression analysis, including Hazard ratio with 95% confidence interval.

4 Summary of the papers

4.1 Paper I Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas

Grading is one of the main prognostic factors for papillary urothelial carcinomas, essential for decision making in follow-up and treatment of NMIBC. Currently, both the WHO grading system from 1973 (WHO73) and the WHO grading system from 2004 (WHO04) are used. WHO04 was introduced mainly to improve reproducibility. However, both grading systems are criticized for high inter-observer variability. WHO73 divides the tumour into grade 1, 2 and 3. WHO04 uses the terms "Papillary urothelial neoplasia of low malignant potential" (PUNLMP) as well as low- and high-grade carcinomas. Grading is based on degree of anaplasia. Several histopathological features are included in the evaluation. Paper I investigates the reproducibility and prognostic value of each of the underlying morphologic features evaluated by a pathologist while grading. At the Department of Pathology, Stavanger University Hospital, in the period 01.01.2002–31.12.2006, 185 primary NMIBC patients met our inclusion criteria. In this cohort, 13 patients progressed to a higher stage within 5 years. These 13 cases, together with 25 randomly selected control cases without progression, were further analysed in this study. From textbooks in urological pathology, 13 morphological microscopic features were extracted: Papillae architecture, superficial layer, papillary fusion, nuclear polarity, cell maturation, cohesion, mitoses, nuclear enlargement, nuclear shape, nuclear hyperchromasia, chromatin pattern, nucleoli, and giant nuclei. All cases were reviewed by three pathologists and scored for each of these microscopic features, according to both WHO73 and WHO04.

Finally, a consensus grade was obtained. Reproducibility was calculated by Gwet`s agreement coefficient, and prognostic value regarding progression was estimated by Area Under Curve (AUC) of the receiver operating characteristic function (ROC). The features varied considerably when it comes to both reproducibility and prognostic value. The most reproducible features, with Gwet`s agreement coefficient > 0.60, were papillary architecture, nuclear polarity, cellular maturation, nuclear enlargement and giant nuclei. The significantly prognostic features observed were: nucleoli, papillary fusion, and nuclear polarity.
Conclusion: The histopathological features behind the different WHO grading systems vary considerably when it comes to reproducibility and prognostic value. Nuclear polarity was the only morphological feature which was both reasonably reproducible and had significant prognostic value.

4.2 Paper II Mitotic activity index and CD25+ lymphocytes predict risk of stage progression in non-muscle invasive bladder cancer

WHO grade and TNM stage are currently the most emphasized prognostic factors for NMIBC, largely deciding follow-up and treatment regimes. Grading is encumbered with suboptimal reproducibility, and TNM staging can be challenging in TURB material. In a search for better prognostic markers, we wanted to investigate the independent prognostic value of proliferation markers, mean nuclear area of the ten largest nuclei (MNA10) and the composition of tumour infiltrating immune cells (CD4+, CD8+, CD25+ and CD138+). Included in the study were 183 patients diagnosed with a primary NMIBC at Stavanger University Hospital in the period 01.01.2002–31.12.2006. To investigate for recurrence, we only included patients (a) whose urinary bladder was retained for at least three months after primary diagnosis, and (b) who followed the regular protocol for cystoscopies. These inclusion criteria

yielded 177 patients for recurrence analysis. All cases were reviewed for grade (both WHO73 and WHO04) and stage. Calculation of mitotic activity index (MAI) and MNA10 were performed on original sections stained with Haematoxylin Eosin Saffron (HES). Immunohistochemical staining for the proliferation markers Ki67 and PPH3, as well as immune cell markers CD4, CD8, CD25 and CD138, were performed on consecutive sections. In the calculation of KI67, MNA10, and immune cell markers, a semi-automated interactive image analysis system (QPRODIT) was used. Estimation of MAI and PPH3 was performed according to a previously described protocol (83, 84). Recurrence of bladder tumour more than 3 months after primary diagnosis occurred in 105 patients. Progression to MIBC occurred in 13 patients. All independent variables were dichotomized and survival analysis were performed. In recurrence analysis, only multifocality and Ki67 were prognostic. Surprisingly, those with low Ki67 showed shorter recurrence-free survival. Grade, stage, and all proliferation markers were associated with increased progression risk. Among the immune cell markers only CD25 was prognostic, with a high count associated with shorter progression-free survival. In a multivariate analysis, the combination of MAI and CD25+ was the most prognostic.

Conclusion: Mitotic activity index combined with CD25+ lymphocytes are the strongest prognostic factor predicting progression in our cohort of primary non-muscle invasive bladder cancer patients.

4.3 Paper III T1 Substaging of Non-muscle Invasive Bladder Cancer is Associated with Bacillus Calmette-Guerin Failure and Improves Patient Stratification at Diagnosis

Regular BCG instillations according to protocol is first treatment option in intermediate and high-risk NMIBC, after TURB. As many as 30–50% of high-risk NMIBC recur after adjuvant BCG instillation and develop

high grade or muscle invasive disease. Delayed radical cystectomy is correlated with worse overall survival. High-risk patients classified as "highest risk" of progression, according to the EAU guidelines, are recommended for radical cystectomy, although this carries the risk of overtreatment. This publication investigates the predictive and prognostic value of T1 substaging, aiming to improve risk stratification and treatment decision making. Primary high-risk NMIBC patients receiving at least 5 induction instillations, from three Dutch and one Norwegian hospital, were included in the study. All tumours were centrally reviewed and T1 substaging was performed. T1 tumours were divided into microinvasive (T1m) and extensive invasive (T1e). Microinvasion was defined as only one invasive focus not exceeding 0.5 mm. BCG failure was defined as a biopsy-proven T1 HG recurrence after 5/6 induction instillations, HG recurrence after adequate BCG instillation (5/6 induction instillations plus 2/3 maintenance instillations), or recurring muscle invasive disease. A total number of 264 patients were T1, 27% T1m and 73% T1e. Median follow-up was 68 months. BCG failure was more frequent among T1e than T1m, 41% vs 21% (p=0.002). The 3-year high grade recurrence-free survival for T1e and T1m were 64% and 83%, respectively (p=0.004). In multivariate analysis, T1 substaging was an independent predictor of both high-grade recurrence-free and progression-free survival. Patients within the highest risk subgroup showed significantly better progression-free survival if T1m compared to T1e (p=0.038).

Conclusion: T1 substaging is predictive for BCG-failure. In patients with highest risk of progression, T1 substaging seems to improve risk stratification and might aid in therapy decision making.

4.4 Paper IV Proliferation and immunohistochemistry for p53 and CK20 in predicting prognosis of non-muscle invasive papillary urothelial carcinomas

Reliable and reproducible prognostic markers for NMIBC are lacking. We have previously shown that proliferation markers (Ki-67, PPH3, MAI) are better prognostic markers for stage progression than WHO grading. We wanted to validate our statement in an extended cohort, and also include two classical immunohistochemical markers, p53 and CK20. Immunohistochemistry for p53 is a surrogate marker for mutations in *TP53*, a molecular characteristic for MIBC. CK20, a cytoskeleton-associated intermediate protein filament, is frequently used for distinguishing urothelial carcinoma in situ from reactive urothelial changes. CK20 is also proposed as a surrogate marker for luminal urothelial carcinomas. We included 349 patients diagnosed with primary papillary NMIBC at Stavanger University hospital in the period 01.01.2002 to 01.01.2011. For recurrence analyses, we only included patients who retained their bladder for at least three months after primary diagnosis, and followed regular cystoscopies according to protocol. This yielded 337 patients for recurrence analysis. WHO grade and TNM stage were reviewed for all cases. Immunohistochemistry for Ki-67, PPH3, p53 and CK20 were performed on consecutive tissue sections. Mitotic activity index (MAI) and PPH3 were calculated according to previously described and validated protocols. Ki-67 was calculated by a computerized semi-automated image analysis system (QPRODIT). The proliferation markers were dichotomised using previously published thresholds. P53 immunohistochemically stained sections were scanned at 400x magnification and uploaded to the digital image analysis program Visiopharm®. A calculation of p53 positive cells was fully automated, after a DAB deconvolution pixel value was set. A p53 positive tumour was defined by an average >15% positive cells in three hotspots. CK20 was evaluated using the semi-quantitative

immunoreactive score (IRS). A positive tumour was defined by an IRS >3. Tumour recurrence occurred in 167 out of 337. Out of 349 patients, 26 experienced stage progression. High age and tumour multifocality were associated with shorter recurrence-free survival. High Ki-67 was barely associated with longer recurrence-free survival. All variables, except gender, were statistically significant for stage progression. Among the investigated variables, MAI had the highest hazard ratio (HR 9.1, CI 3.8 – 22.3), being the only marker exceeding the prognostic value of stage (HR 6.2, CI 2.8 – 13.7) and WHO (2004/2016) grade (HR 4.9, 2.0 – 12.2).

Conclusion: There are few reliable biomarkers for tumour recurrence in NMIBC. MAI is the strongest prognostic biomarker for stage progression in our analysis. The immunohistochemical markers p53 and CK20 are significant associated with stage progression, but they are not performing better than the currently used prognostic markers WHO grade and tumour stage.

5 Discussion

5.1 Patient material and definitions

Our cohort, first from 2002–2007 and later from 2002–2011, is based on a population of around 350,000 inhabitants. The complete region is served by Stavanger University hospital (SUH). There are a few private clinics, but they all refer patients with newly diagnosed urothelial carcinomas to SUH. All NMIBC patients have their follow-up at, or in close cooperation, with SUH. All tissue removed for diagnostic or treatment purposes are therefore sent to the Department of Pathology at SUH. This makes the cohort a genuine population-based one, and ideal for research on prognostic factors. The population is representative of a western society, including patients from both urban and rural areas.

The research on bladder cancer at Department of Pathology at SUH is focused on NMIBC. As mentioned in the introduction, 70–80% of all urothelial cancer are NMIBC at the time of first diagnosis. Huge resources are spent on follow-up of these patients, aiming for early detection of recurrences and progression. Progression to MIBC marks a change in treatment protocol and prognostic view. Therefor NMIBC is a natural category for research on prognostic factors.

Our research focus has been on urothelial carcinomas of the urinary bladder. A significant number of urothelial carcinomas are extra-vesical, most of them from the upper urinary tract (ureter and renal pelvis – UTUC). Although they share some common genomic alterations, they have been shown to be a unique entity, also at the molecular level (139). A question that appeared was how to handle those patients with a recurrent extra-vesical urothelial carcinoma, or those having a previous extra-vesical urothelial carcinoma at time of diagnosis of a primary urothelial carcinoma. Since it was impossible to say if they represented the same disease or not, we decided to exclude all patients with extra-vesical urothelial carcinoma in their medical records. The only cases we

included in the cohort were those with urothelial carcinoma in the collicular area of prostatic urethra. Biopsies from this location are frequently taken, either because of tumour in the bladder neck region or clinical suspicion of CIS in urethra, or before a cystectomy as part of the preparation process. Primary carcinomas in the urethra are rare, annual age-adjusted incidence rate among men in the United States is estimated to 4.3 per million (140).

The most interesting endpoint in Paper II and IV is progression-free survival. In the general literature, the definition of progression in NMIBC varies. Some publications define "Progression" as all progression in TNM stage, others only include progression to MIBC (at least T2). The distinction between Ta/T1 vs T2 (NMIBC vs MIBC) is clinically the most important distinction. We will claim that a change from Ta to T1 is of similar importance biologically. This event marks a change in the ability of tumour tissue to invade, and possibly metastasize. In paper II we defined "Progression" as a progression to MIBC. In paper IV we decided to also include cases progressing from Ta to T1.

Another issue regarding the dataset are criteria for follow-up. While investigating "Progression", we decided to also include metastases which appear after cystectomy. For "Recurrence", we decided to investigate the tendency of tumours to recur in the bladder only. Recurrent tumours elsewhere in the body, for instance after cystectomy, were not considered interesting in this setting. Also, other differences regarding "Progression" and "recurrence" appeared. To be able to discover recurrent, small, low-grade tumours, regular follow-up with cystoscopy according to protocol is necessary. The same strict criteria do not seem necessary to discover progressing tumours, as most of them will be clinically evident at some point, independent of follow-up regime. Consequently, in our dataset, follow-up time for "Recurrence" and "Progression" are registered differently. Time to recurrence was calculated from date of primary diagnosis until last registered cystoscopy or cystectomy. Time to progression was counted

from date of primary diagnosis until death, last known contact if they moved or until our last check in the medical journal (30.06.2016). This strategy resulted in two slightly different cohorts in our publications (Paper II and IV), as a minimum follow-up time was set to 3 months.

5.2 Study design and other considerations

The study design in our research is retrospective. Information on diagnosis and prognosis was available in the medical records. For extracting data related to patients' clinical variables, we relied on detailed notes in the medical records. Variables like tumour size, and to some degree multifocality, were not sufficiently registered, making them unsuitable for statistical analyses. Also, factors like smoking habits and occupation were not sufficiently registered. Compared to prospective analysis, with the retrospective design, we have less control over factors that make the cohort heterogeneous and might have influence on outcome.

A limitation of our cohort is the limited number of patients with progression. In the final cohort that included patients from 2002–2011, we identified 26 (7.4%) with stage progression, and 21 (5.9%) with progression to MIBC. Although this progression rate is lower than that reported in most publications, we did find publications that reported comparable rates (141). We are cognizant that a higher proportion of cases with disease progression in our cohort would have strengthened our statistical analyses. As such, significant results from our papers need to be validated in several independent cohorts before adoption in routine clinical use.

Patients included in our cohort were treated according to Norwegian national guidelines. Most high-risk patients received instillation therapy with BCG, a few also received regular instillations

with chemotherapy. Many of the included patients received optimized follow-up and treatment. This may have contributed to lower number of patients with progression in our cohort. As both follow-up and treatment depended on risk group, established prognostic variables like grade and stage may have been weakened in our results compared to their "true" prognostic value. Another issue is the phenomenon re-TURB. According to national and EAU guidelines (28, 29) a second TURB is indicated during the weeks following primary surgery if uncertainty about resection completeness exists, or if there is a lack of identified m. propria in the resection material or if the resected tumour is a T1 tumour. In our data registration, TURB and re-TURB, if performed, are grouped together. In most cases tumour tissue in TURB and re-TURB represent the same tumour. The highest grade and stage were registered. Date of first diagnosis, either on biopsy or TURB, was registered. We did not adjust our analyses based on whether a re-TURB was performed. In publications involving large cohorts by Sylvester et al. (n=3401) and Gontero et al. (n=2451), a re-TURB did not affect progression risk (31, 142).

One of the main issues with WHO grading system is the lacking reproducibility, emphasized in Paper I. Our effort has been focused on establishing reproducible and better prognostic markers for NMIBC. In Paper II and IV we used the semiautomatic interactive quantitative image analysis system (QPRODIT). Although far more reproducible, some interobserver variability still persisted. For each cell pointed out, the observer needed to decide whether it was a tumour cell or not, and whether the cell was immunohistochemically positive or negative. Usually this is an easy exercise, but different shades of brown colour can make interpretation hard. In Paper IV, for estimation of p53, we used a fully automated digital image analysis system (Visiopharm®). The threshold for positivity was set subjectively, based on experience with immunohistochemistry interpretation, but all the analyses were

performed automatically. In future work, we will preferably use digital image analysis for fast, easy, and reproducible results.

In Paper II, the patient's immune response, and subsets of immune cells, were investigated in the most anaplastic area. Other publications in this field differ in analysing immune cells in urothelial tumour tissue only, stroma only, or both. We counted positive cells in tumour tissue, consisting of atypical urothelium, fibrovascular stalks, and invaded stroma. This was done for simplicity, as discerning invading urothelial tissue from the stroma embedding it, seemed difficult and arbitrary. This separation would be especially problematic in cases with invading single cells or strands of cells, like in the diffuse variant of urothelial carcinoma. In our experience, stromal areas tend to have higher density of immune cells than pure urothelial regions. Additionally, the cell density in the stroma is lower, compared to the crowded urothelium. This could bias the results, with higher proportion of immune cells in cases where the demarcated area contained a higher proportion of stroma. As the amount of urothelium in most cases is overwhelming, we believe that this issue had minimal influence on the results.

Immunohistochemical staining for CK20 and p53 were performed in Paper IV. CK20 is validated as a surrogate marker for luminal carcinomas, and p53 positivity as a surrogate marker for *TP53* mutations. Ideally, we would have implemented a basal marker as well, preferentially CK5/6. This was not done because of limited resources, but will be considered in the future.

Originally, we planned to create an image atlas for direct comparison, as a support tool for pathologists to use while grading. The idea was to improve reproducibility. Such an atlas could have been implemented in a software for digital pathology. Example images of each grade, emphasizing different histopathological features, could then be easily displayed on a screen next to an image of the case to be graded.

We spent time on finding cases representing good examples for each grade and feature mentioned in Paper I. An atlas together with a scoring system was established and tested on residents and pathologists at Department of Pathology at SUH. Unfortunately, it did not achieve any better results than current grading. We decided not to move on with this concept.

6 Future Directions

6.1 Digital pathology and artificial intelligence

Pathology is in the beginning of a digital revolution. Slides for microscopy are being scanned, digitized, and displayed as whole slide images (WSI). Virtual microscopy is on its way to replace conventional microscopy. With this change in pathology, new challenges and opportunities appear. Storing of WSIs demands enormous amounts of storing capacity. Legal and ethical aspects, as well as economic issues also make the introduction of digital pathology complicated. If we can overcome these obstacles the digitization process holds a promise of many new advantages. The WSIs can be available on several working stations, and the WSIs can be shared for second opinion over remote distances in seconds. The pathologist can annotate regions of special interest and measure structures and distances digitally (143, 144).

Digital pathology opens the opportunity for digital image analysis and computer aided diagnosis (CAD) systems. This has the potential to eliminate subjectivity and minimize variability. Applications are already available to perform tasks especially prone to high inter- and intra-observer variability. Typically, this involves estimation of proportion of tumour cells or proportion of cells with positive immunohistochemistry. We took advantage of such a CAD system in Paper IV while estimating the proportion of p53-positive cells. For instance, such systems can be routinely used for estimation of oestrogen and progesterone receptor in breast cancer, and for estimation of Ki67 in carcinoid tumours (145).

With the implementation of deep learning in pathology, even more complicated tasks can be performed automatically. These methods are criticized for reaching conclusions without outlining their way of analysing, hence the designation “black box”. With deep learning, patterns can be recognized and analysed. Regions of interest can be

pointed out, and diagnosis suggestions can be presented to the pathologist. Personally, I believe, that artificial intelligence will work together with the pathologist, increasing accuracy and consistency, rather than replacing the pathologist completely. So far, several publications on deep learning applied on WSI of bladder cancer exist. Peng-Nien Yin et al. used different machine learning techniques to differentiate Ta and T1 urothelial carcinomas on HE-stained slides (146). By combining six classic machine learning techniques, and applying them to three defined features related to invasion (desmoplastic reaction, more pink cytoplasm, retraction artefacts), an impressive accuracy of 91 -96% was achieved. Using a deep learning model (model based on convolutional neural networks) an accuracy of 84% was achieved. Jansen et al. used deep learning to automatically extract urothelium in the images of NMIBC to be presented for a classification network performing grading according to WHO04 (147). Three pathologists performed individual and consensus grading on all 328 cases. In this publication, the agreement between the pathologists ranged from fair to moderate (kappa values 0.35–0.52). The agreement between the automatic grading and the consensus grading was moderate (kappa value 0.48). The deep learning-based grading system correctly graded 76% of the low grade and 71% of high-grade tumours, compared to the consensus grading by the pathologists. The authors concluded that deep learning can be used in grading of urothelial carcinomas.

In cooperation with Department of Electrical Engineering and Computer science, at the University of Stavanger, we applied deep learning models on WSIs from our cohort of NMIBC patients from 2002–2011. Different models were developed to extract urothelial tissue in the WSI. The best models differentiated tissue types with high precision (F1 score 0.986) (Appendix 2). Furthermore, grading models based on deep learning were developed. Compared to a urological pathologist as gold standard, the best model achieved an F1 score of 0.91 for both low grade and high-grade tumours (Appendix 4). This model

creates a heat map for grading, highlighting the areas with the highest grade on each WSI. This function can aid pathologists in working more efficiently by leading attention to the most diagnostically relevant areas in the image.

Digital pathology and the introduction of artificial intelligence are promising aids that will help move towards more objective and standardized measures, and more confident diagnostics. In addition to make existing markers more reliable, the new computational techniques may aid in future research on new prognostic markers.

6.2 Next generation sequencing (NGS)

With the introduction of NGS, DNA sequencing has become time- and cost effective. Selected parts of the genome can be sequenced for mutations in hours. Mapping of clinically relevant tumour mutations make the basis for individual tailored cancer treatment. Targeted therapy can be administered based on molecular alterations, also to some extent, irrespective of organ of origin. For locally advanced or metastatic urothelial carcinomas with mutations in FGFR2/3, or those progressing on traditional platinum-based chemotherapy, treatment with a FGFR-inhibitor (erdafitinib) is FDA approved. Drugs targeting the signalling pathway PI3K/AKT/mTOR, the ERBB-receptor, as well as chromatin remodelling genes, are under investigation.

In this thesis, current knowledge regarding molecular alterations in bladder cancer are outlined in "1.6 Molecular alterations in bladder cancer". As mentioned here, urothelial carcinomas generally have a high mutational burden. Although the main pathways and common mutations for urothelial carcinomas are described, we believe that there is still clinically relevant knowledge to be discovered. In future work, we would like to perform mutational analysis on a panel of 52 known cancer genes,

using NGS (Oncomine focus assay by Thermo Fisher), and compare the mutational profile with outcome and other prognostic markers investigated in this thesis.

6.3 Imaging mass cytometry

In our investigation of bladder cancer so far, we have investigated for proteins in the tumour by performing standard immunohistochemistry, using enzyme-labelled antibodies. One section from the block was cut per protein investigated. In future work we are planning to perform multiplex immunohistochemistry on our material by using the Hyperion imaging system. This system is based on the cytometry by time of flight (CyTOF) technology, enabling the simultaneous investigation of up to 37 proteins at a subcellular level. By using this technology, we will save tumour tissue, that sometimes are limited, e.g., in small biopsies. The metal isotope labels exploited in this system make it possible to see the presence of several antigens in the same cell. This is valuable information, especially while investigating subsets of immune cells in the tumour microenvironment.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.
2. Kirkali Z, Chan T, Manoharan M, Algaba F, Busch C, Cheng L, et al. Bladder cancer: epidemiology, staging and grading, and diagnosis. Urology. 2005;66(6 Suppl 1):4-34.
3. Norway Cro. Cancer in Norway 2019 - Cancer incidence, mortality, survival and prevalence in Norway. 2020.
4. Moch H, Humphrey PA, Ulbright TM, Reuter VE. World Health Organization Classification of tumours. 2016:77 - 135.
5. Sievert KD, Amend B, Nagele U, Schilling D, Bedke J, Horstmann M, et al. Economic aspects of bladder cancer: what are the benefits and costs? World J Urol. 2009;27(3):295-300.
6. Richters A, Aben KKH, Kiemeney L. The global burden of urinary bladder cancer: an update. World J Urol. 2020;38(8):1895-904.
7. Dobruch J, Daneshmand S, Fisch M, Lotan Y, Noon AP, Resnick MJ, et al. Gender and Bladder Cancer: A Collaborative Review of Etiology, Biology, and Outcomes. Eur Urol. 2016;69(2):300-10.
8. Cumberbatch MG, Rota M, Catto JW, La Vecchia C. The Role of Tobacco Smoke in Bladder and Kidney Carcinogenesis: A Comparison of Exposures and Meta-analysis of Incidence and Mortality Risks. Eur Urol. 2016;70(3):458-66.
9. Chen CH, Shun CT, Huang KH, Huang CY, Tsai YC, Yu HJ, et al. Stopping smoking might reduce tumour recurrence in nonmuscle-invasive bladder cancer. BJU Int. 2007;100(2):281-6; discussion 6.
10. Afshari M, Janbabaei G, Bahrami MA, Moosazadeh M. Opium and bladder cancer: A systematic review and meta-analysis of the odds ratios for opium use and the risk of bladder cancer. PLoS One. 2017;12(6):e0178527.
11. Thomas AA, Wallner LP, Quinn VP, Slezak J, Van Den Eeden SK, Chien GW, et al. Association between cannabis use and the risk of bladder cancer: results from the California Men's Health Study. Urology. 2015;85(2):388-92.

12. Cumberbatch MGK, Jubber I, Black PC, Esperto F, Figueroa JD, Kamat AM, et al. Epidemiology of Bladder Cancer: A Systematic Review and Contemporary Update of Risk Factors in 2018. Eur Urol. 2018;74(6):784-95.
13. Figueroa JD, Koutros S, Colt JS, Kogevinas M, Garcia-Closas M, Real FX, et al. Modification of Occupational Exposures on Bladder Cancer Risk by Common Genetic Polymorphisms. J Natl Cancer Inst. 2015;107(11).
14. Lukas C, Selinski S, Prager HM, Blaszkewicz M, Hengstler JG, Golka K. Occupational bladder cancer: Polymorphisms of xenobiotic metabolizing enzymes, exposures, and prognosis. J Toxicol Environ Health A. 2017;80(7-8):439-52.
15. Garcia-Closas M, Rothman N, Figueroa JD, Prokunina-Olsson L, Han SS, Baris D, et al. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. Cancer Res. 2013;73(7):2211-20.
16. Botteri E, Ferrari P, Roswall N, Tjønneland A, Hjartåker A, Huerta JM, et al. Alcohol consumption and risk of urothelial cell bladder cancer in the European prospective investigation into cancer and nutrition cohort. Int J Cancer. 2017;141(10):1963-70.
17. Baris D, Waddell R, Beane Freeman LE, Schwenn M, Colt JS, Ayotte JD, et al. Elevated Bladder Cancer in Northern New England: The Role of Drinking Water and Arsenic. J Natl Cancer Inst. 2016;108(9).
18. Mendez WM, Jr., Eftim S, Cohen J, Warren I, Cowden J, Lee JS, et al. Relationships between arsenic concentrations in drinking water and lung and bladder cancer incidence in U.S. counties. J Expo Sci Environ Epidemiol. 2017;27(3):235-43.
19. Wallis CJ, Mahar AL, Choo R, Herschorn S, Kodama RT, Shah PS, et al. Second malignancies after radiotherapy for prostate cancer: systematic review and meta-analysis. BMJ. 2016;352:i851.
20. Abern MR, Dude AM, Tsivian M, Coogan CL. The characteristics of bladder cancer after radiotherapy for prostate cancer. Urol Oncol. 2013;31(8):1628-34.
21. Lim A, Rao P, Matin SF. Lynch syndrome and urologic malignancies: a contemporary review. Current opinion in urology. 2019;29(4):357-63.
22. Gripp KW, Rauen KA. Costello Syndrome. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al.,

editors. GeneReviews(®). Seattle (WA): University of Washington, Seattle

Copyright © 1993-2021, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.; 1993.

23. Zaghloul MS, Zaghloul TM, Bishr MK, Baumann BC. Urinary schistosomiasis and the associated bladder cancer: update. J Egypt Natl Canc Inst. 2020;32(1):44.
24. Pasin E, Josephson DY, Mitra AP, Cote RJ, Stein JP. Superficial bladder cancer: an update on etiology, molecular development, classification, and natural history. Rev Urol. 2008;10(1):31-43.
25. Mostofi FKS, L. H.; Torloni, H. Histologic typing of urinary bladder tumours. Geneva: World Health Organization; 1973.
26. Moch H, Humphrey P, Ulbright T, Reuter V. WHO Classification of tumours of the urinary system and male genital organs. 4 ed. Lyon: International agency for research on cancer; 2016. p. 99 - 108.
27. Epstein JI, Amin MB, Reuter VR, Mostofi FK. The World Health Organization/International Society of Urological Pathology consensus classification of urothelial (transitional cell) neoplasms of the urinary bladder. Bladder Consensus Conference Committee. Am J Surg Pathol. 1998;22(12):1435-48.
28. EAU Guidelines. Edn. presented at the EAU Annual Congress Milan 2021. 2021 [
29. Helsedirektoratet. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av blærekreft 2018 [updated 14.04.2021.
30. Kim JK, Moon KC, Jeong CW, Kwak C, Kim HH, Ku JH. Papillary Urothelial Neoplasm of Low Malignant Potential (PUNLMP) After Initial TUR-BT: Comparative Analyses with Noninvasive Low-Grade Papillary Urothelial Carcinoma (LGPUC). J Cancer. 2017;8(15):2885-91.
31. Sylvester RJ, Rodríguez O, Hernández V, Turturica D, Bauerová L, Bruins HM, et al. European Association of Urology (EAU) Prognostic Factor Risk Groups for Non-muscle-invasive Bladder Cancer (NMIBC) Incorporating the WHO 2004/2016 and WHO 1973 Classification Systems for Grade: An Update from the EAU NMIBC Guidelines Panel. Eur Urol. 2021;79(4):480-8.

32. Hentschel AE, van Rhijn BWG, Bründl J, Compérat EM, Plass K, Rodríguez O, et al. Papillary urothelial neoplasm of low malignant potential (PUN-LMP): Still a meaningful histo-pathological grade category for Ta, noninvasive bladder tumors in 2019? Urol Oncol. 2020;38(5):440-8.
33. Cheng L, Neumann RM, Nehra A, Spotts BE, Weaver AL, Bostwick DG. Cancer heterogeneity and its biologic implications in the grading of urothelial carcinoma. Cancer. 2000;88(7):1663-70.
34. Lopez-Beltran A, Montironi R. Non-invasive urothelial neoplasms: according to the most recent WHO classification. Eur Urol. 2004;46(2):170-6.
35. Soukup V, Capoun O, Cohen D, Hernandez V, Babjuk M, Burger M, et al. Prognostic Performance and Reproducibility of the 1973 and 2004/2016 World Health Organization Grading Classification Systems in Non-muscle-invasive Bladder Cancer: A European Association of Urology Non-muscle Invasive Bladder Cancer Guidelines Panel Systematic Review. Eur Urol. 2017.
36. van Rhijn BWG, Hentschel AE, Bründl J, Compérat EM, Hernández V, Čapoun O, et al. Prognostic Value of the WHO1973 and WHO2004/2016 Classification Systems for Grade in Primary Ta/T1 Non-muscle-invasive Bladder Cancer: A Multicenter European Association of Urology Non-muscle-invasive Bladder Cancer Guidelines Panel Study. European urology oncology. 2021;4(2):182-91.
37. van Rhijn BW, van Leenders GJ, Ooms BC, Kirkels WJ, Zlotta AR, Boevé ER, et al. The pathologist's mean grade is constant and individualizes the prognostic value of bladder cancer grading. Eur Urol. 2010;57(6):1052-7.
38. Yorukoglu K, Tuna B, Dikicioglu E, Duzcan E, Isisag A, Sen S, et al. Reproducibility of the 1998 World Health Organization/International Society of Urologic Pathology classification of papillary urothelial neoplasms of the urinary bladder. Virchows Arch. 2003;443(6):734-40.
39. Gonul, II, Poyraz A, Unsal C, Acar C, Alkibay T. Comparison of 1998 WHO/ISUP and 1973 WHO classifications for interobserver variability in grading of papillary urothelial neoplasms of the bladder. Pathological evaluation of 258 cases. Urol Int. 2007;78(4):338-44.
40. Mangrud OM, Waalen R, Gudlaugsson E, Dalen I, Tasdemir I, Janssen EA, et al. Reproducibility and prognostic value of WHO1973

and WHO2004 grading systems in TaT1 urothelial carcinoma of the urinary bladder. PLoS One. 2014;9(1):e83192.
41. Lopez-Beltran A, Henriques V, Montironi R, Cimadamore A, Raspollini MR, Cheng L. Variants and new entities of bladder cancer. Histopathology. 2019;74(1):77-96.
42. Wang G, McKenney JK. Urinary Bladder Pathology: World Health Organization Classification and American Joint Committee on Cancer Staging Update. Arch Pathol Lab Med. 2019;143(5):571-7.
43. Williamson SR, Zhang S, Lopez-Beltran A, Shah RB, Montironi R, Tan PH, et al. Lymphoepithelioma-like carcinoma of the urinary bladder: clinicopathologic, immunohistochemical, and molecular features. Am J Surg Pathol. 2011;35(4):474-83.
44. Willis DL, Fernandez MI, Dickstein RJ, Parikh S, Shah JB, Pisters LL, et al. Clinical outcomes of cT1 micropapillary bladder cancer. J Urol. 2015;193(4):1129-34.
45. Ikegami H, Iwasaki H, Ohjimi Y, Takeuchi T, Ariyoshi A, Kikuchi M. Sarcomatoid carcinoma of the urinary bladder: a clinicopathologic and immunohistochemical analysis of 14 patients. Hum Pathol. 2000;31(3):332-40.
46. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8.
47. Network CGAR. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315-22.
48. Obermann EC, Meyer S, Hellge D, Zaak D, Filbeck T, Stoehr R, et al. Fluorescence in situ hybridization detects frequent chromosome 9 deletions and aneuploidy in histologically normal urothelium of bladder cancer patients. Oncol Rep. 2004;11(4):745-51.
49. Chow NH, Cairns P, Eisenberger CF, Schoenberg MP, Taylor DC, Epstein JI, et al. Papillary urothelial hyperplasia is a clonal precursor to papillary transitional cell bladder cancer. Int J Cancer. 2000;89(6):514-8.
50. McConkey DJ, Lee S, Choi W, Tran M, Majewski T, Lee S, et al. Molecular genetics of bladder cancer: Emerging mechanisms of tumor initiation and progression. Urol Oncol. 2010;28(4):429-40.
51. Bertz S, Eckstein M, Stoehr R, Weyerer V, Hartmann A. Urothelial Bladder Cancer: An Update on Molecular Pathology with Clinical Implications. Eur Urol Suppl. 2017;16(12):272-94.

52. Hernández S, López-Knowles E, Lloreta J, Kogevinas M, Amorós A, Tardón A, et al. Prospective study of FGFR3 mutations as a prognostic factor in nonmuscle invasive urothelial bladder carcinomas. J Clin Oncol. 2006;24(22):3664-71.
53. López-Knowles E, Hernández S, Malats N, Kogevinas M, Lloreta J, Carrato A, et al. PIK3CA mutations are an early genetic alteration associated with FGFR3 mutations in superficial papillary bladder tumors. Cancer Res. 2006;66(15):7401-4.
54. Hurst CD, Alder O, Platt FM, Droop A, Stead LF, Burns JE, et al. Genomic Subtypes of Non-invasive Bladder Cancer with Distinct Metabolic Profile and Female Gender Bias in KDM6A Mutation Frequency. Cancer Cell. 2017;32(5):701-15.e7.
55. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017;171(3):540-56.e25.
56. Knowles MA, Hurst CD. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. Nat Rev Cancer. 2015;15(1):25-41.
57. Cappellen D, Gil Diez de Medina S, Chopin D, Thiery JP, Radvanyi F. Frequent loss of heterozygosity on chromosome 10q in muscle-invasive transitional cell carcinomas of the bladder. Oncogene. 1997;14(25):3059-66.
58. Gildea JJ, Herlevsen M, Harding MA, Gulding KM, Moskaluk CA, Frierson HF, et al. PTEN can inhibit in vitro organotypic and in vivo orthotopic invasion of human bladder cancer cells even in the absence of its lipid phosphatase activity. Oncogene. 2004;23(40):6788-97.
59. Allory Y, Beukers W, Sagrera A, Flández M, Marqués M, Márquez M, et al. Telomerase reverse transcriptase promoter mutations in bladder cancer: high frequency across stages, detection in urine, and lack of association with outcome. Eur Urol. 2014;65(2):360-6.
60. Kompier LC, Lurkin I, van der Aa MN, van Rhijn BW, van der Kwast TH, Zwarthoff EC. FGFR3, HRAS, KRAS, NRAS and PIK3CA mutations in bladder cancer and their potential as biomarkers for surveillance and therapy. PLoS One. 2010;5(11):e13821.
61. Jarvis MC, Ebrahimi D, Temiz NA, Harris RS. Mutation Signatures Including APOBEC in Cancer Cell Lines. JNCI cancer spectrum. 2018;2(1).

62. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013;45(9):970-6.
63. Shi MJ, Meng XY, Fontugne J, Chen CL, Radvanyi F, Bernard-Pierrot I. Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. Genome Med. 2020;12(1):85.
64. Shi MJ, Meng XY, Lamy P, Banday AR, Yang J, Moreno-Vega A, et al. APOBEC-mediated Mutagenesis as a Likely Cause of FGFR3 S249C Mutation Over-representation in Bladder Cancer. Eur Urol. 2019;76(1):9-13.
65. Hedegaard J, Lamy P, Nordentoft I, Algaba F, Høyer S, Ulhøi BP, et al. Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. Cancer Cell. 2016;30(1):27-42.
66. Sjödahl G, Lauss M, Lövgren K, Chebil G, Gudjonsson S, Veerla S, et al. A molecular taxonomy for urothelial carcinoma. Clin Cancer Res. 2012;18(12):3377-86.
67. Sjödahl G, Eriksson P, Liedberg F, Höglund M. Molecular classification of urothelial carcinoma: global mRNA classification versus tumour-cell phenotype classification. J Pathol. 2017;242(1):113-25.
68. Damrauer JS, Hoadley KA, Chism DD, Fan C, Tiganelli CJ, Wobker SE, et al. Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. Proc Natl Acad Sci U S A. 2014;111(8):3110-5.
69. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.
70. Prat A, Karginova O, Parker JS, Fan C, He X, Bixby L, et al. Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. Breast Cancer Res Treat. 2013;142(2):237-55.
71. Kardos J, Chai S, Mose LE, Selitsky SR, Krishnan B, Saito R, et al. Claudin-low bladder tumors are immune infiltrated and actively immune suppressed. JCI insight. 2016;1(3):e85902.
72. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, et al. Identification of distinct basal and luminal subtypes of

muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer Cell. 2014;25(2):152-65.
73. Dadhania V, Zhang M, Zhang L, Bondaruk J, Majewski T, Siefker-Radtke A, et al. Meta-Analysis of the Luminal and Basal Subtypes of Bladder Cancer and the Identification of Signature Immunohistochemical Markers for Clinical Use. EBioMedicine. 2016;12:105-17.
74. Aine M, Eriksson P, Liedberg F, Höglund M, Sjödahl G. On Molecular Classification of Bladder Cancer: Out of One, Many. Eur Urol. 2015;68(6):921-3.
75. Kamoun A, de Reyniès A, Allory Y, Sjödahl G, Robertson AG, Seiler R, et al. A Consensus Molecular Classification of Muscle-invasive Bladder Cancer. Eur Urol. 2020;77(4):420-33.
76. AJCC Cancer Staging Manual. Eight Edition ed: Springer; 2017.
77. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin. 2019;69(5):363-85.
78. Bol MG, Baak JP, Buhr-Wildhagen S, Kruse AJ, Kjellevold KH, Janssen EA, et al. Reproducibility and prognostic variability of grade and lamina propria invasion in stages Ta, T1 urothelial carcinoma of the bladder. J Urol. 2003;169(4):1291-4.
79. van Rhijn BW, van der Kwast TH, Alkhateeb SS, Fleshner NE, van Leenders GJ, Bostrom PJ, et al. A new and highly prognostic system to discern T1 bladder cancer substage. Eur Urol. 2012;61(2):378-84.
80. Magers MJ, Lopez-Beltran A, Montironi R, Williamson SR, Kaimakliotis HZ, Cheng L. Staging of bladder cancer. Histopathology. 2019;74(1):112-34.
81. Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol. 2006;49(3):466-5; discussion 75-7.
82. Kim HS, Kim M, Jeong CW, Kwak C, Kim HH, Ku JH. Presence of lymphovascular invasion in urothelial bladder cancer specimens after transurethral resections correlates with risk of upstaging and survival: a systematic review and meta-analysis. Urol Oncol. 2014;32(8):1191-9.
83. Bol MG, Baak JP, de Bruin PC, Rep S, Marx W, Bos S, et al. Improved objectivity of grading of T(A,1) transitional cell carcinomas of

the urinary bladder by quantitative nuclear and proliferation related features. J Clin Pathol. 2001;54(11):854-9.
84. Bol MG, Baak JP, Rep S, Marx WL, Kruse AJ, Bos SD, et al. Prognostic value of proliferative activity and nuclear morphometry for progression in TaT1 urothelial cell carcinomas of the urinary bladder. Urology. 2002;60(6):1124-30.
85. van Rhijn BW, Vis AN, van der Kwast TH, Kirkels WJ, Radvanyi F, Ooms EC, et al. Molecular grading of urothelial cell carcinoma with fibroblast growth factor receptor 3 and MIB-1 is superior to pathologic grade for the prediction of clinical outcome. J Clin Oncol. 2003;21(10):1912-21.
86. Bertz S, Otto W, Denzinger S, Wieland WF, Burger M, Stöhr R, et al. Combination of CK20 and Ki-67 immunostaining analysis predicts recurrence, progression, and cancer-specific survival in pT1 urothelial bladder cancer. Eur Urol. 2014;65(1):218-26.
87. Mangrud OM, Gudlaugsson E, Skaland I, Tasdemir I, Dalen I, van Diermen B, et al. Prognostic comparison of proliferation markers and World Health Organization 1973/2004 grades in urothelial carcinomas of the urinary bladder. Hum Pathol. 2014;45(7):1496-503.
88. Amin MB, Trpkov K, Lopez-Beltran A, Grignon D. Best practices recommendations in the application of immunohistochemistry in the bladder lesions: report from the International Society of Urologic Pathology consensus conference. Am J Surg Pathol. 2014;38(8):e20-34.
89. Harnden P, Eardley I, Joyce AD, Southgate J. Cytokeratin 20 as an objective marker of urothelial dysplasia. Br J Urol. 1996;78(6):870-5.
90. Alston ELJ, Zynger DL. Does the addition of AMACR to CK20 help to diagnose challenging cases of urothelial carcinoma in situ? Diagn Pathol. 2019;14(1):91.
91. Desai S, Lim SD, Jimenez RE, Chun T, Keane TE, McKenney JK, et al. Relationship of cytokeratin 20 and CD44 protein expression with WHO/ISUP grade in pTa and pT1 papillary urothelial neoplasia. Mod Pathol. 2000;13(12):1315-23.
92. Harnden P, Mahmood N, Southgate J. Expression of cytokeratin 20 redefines urothelial papillomas of the bladder. Lancet. 1999;353(9157):974-7.
93. Esrig D, Spruck CH, 3rd, Nichols PW, Chaiwun B, Steven K, Groshen S, et al. p53 nuclear protein accumulation correlates with

mutations in the p53 gene, tumor grade, and stage in bladder cancer. Am J Pathol. 1993;143(5):1389-97.
94. Kelsey KT, Hirao T, Schned A, Hirao S, Devi-Ashok T, Nelson HH, et al. A population-based study of immunohistochemical detection of p53 alteration in bladder cancer. Br J Cancer. 2004;90(8):1572-6.
95. Hodgson A, Xu B, Downes MR. p53 immunohistochemistry in high-grade urothelial carcinoma of the bladder is prognostically significant. Histopathology. 2017;71(2):296-304.
96. Esrig D, Elmajian D, Groshen S, Freeman JA, Stein JP, Chen SC, et al. Accumulation of nuclear p53 and tumor progression in bladder cancer. N Engl J Med. 1994;331(19):1259-64.
97. Gooden MJ, de Bock GH, Leffers N, Daemen T, Nijman HW. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. Br J Cancer. 2011;105(1):93-103.
98. Mahmoud SM, Paish EC, Powe DG, Macmillan RD, Grainge MJ, Lee AH, et al. Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. J Clin Oncol. 2011;29(15):1949-55.
99. Ooft ML, van Ipenburg JA, Braunius WW, Zuur CI, Koljenović S, Willems SM. Prognostic role of tumor infiltrating lymphocytes in EBV positive and EBV negative nasopharyngeal carcinoma. Oral Oncol. 2017;71:16-25.
100. Liu K, Zhao K, Wang L, Sun E. The prognostic values of tumor-infiltrating neutrophils, lymphocytes and neutrophil/lymphocyte rates in bladder urothelial cancer. Pathol Res Pract. 2018;214(8):1074-80.
101. Sharma P, Shen Y, Wen S, Yamada S, Jungbluth AA, Gnjatic S, et al. CD8 tumor-infiltrating lymphocytes are predictive of survival in muscle-invasive urothelial carcinoma. Proc Natl Acad Sci U S A. 2007;104(10):3967-72.
102. Yu A, Mansure JJ, Solanki S, Siemens DR, Koti M, Dias ABT, et al. Presence of lymphocytic infiltrate cytotoxic T lymphocyte CD3+, CD8+, and immunoscore as prognostic marker in patients after radical cystectomy. PLoS One. 2018;13(10):e0205746.
103. Faraj SF, Munari E, Guner G, Taube J, Anders R, Hicks J, et al. Assessment of tumoral PD-L1 expression and intratumoral CD8+ T cells in urothelial carcinoma. Urology. 2015;85(3):703.e1-6.
104. Wahlin S, Nodin B, Leandersson K, Boman K, Jirström K. Clinical impact of T cells, B cells and the PD-1/PD-L1 pathway in muscle invasive bladder cancer: a comparative study of transurethral

resection and cystectomy specimens. Oncoimmunology. 2019;8(11):e1644108.
105. Zhang S, Wang J, Zhang X, Zhou F. Tumor-infiltrating CD8+ lymphocytes predict different clinical outcomes in organ- and non-organ-confined urothelial carcinoma of the bladder following radical cystectomy. PeerJ. 2017;5:e3921.
106. Horn T, Laus J, Seitz AK, Maurer T, Schmid SC, Wolf P, et al. The prognostic effect of tumour-infiltrating lymphocytic subpopulations in bladder cancer. World J Urol. 2016;34(2):181-7.
107. Zhang Q, Hao C, Cheng G, Wang L, Wang X, Li C, et al. High $CD4^+$ T cell density is associated with poor prognosis in patients with non-muscle-invasive bladder cancer. Int J Clin Exp Pathol. 2015;8(9):11510-6.
108. Yano H, Andrews LP, Workman CJ, Vignali DAA. Intratumoral regulatory T cells: markers, subsets and their impact on anti-tumor immunity. Immunology. 2019;157(3):232-47.
109. Curiel TJ, Coukos G, Zou L, Alvarez X, Cheng P, Mottram P, et al. Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival. Nat Med. 2004;10(9):942-9.
110. Miyake M, Tatsumi Y, Gotoh D, Ohnishi S, Owari T, Iida K, et al. Regulatory T Cells and Tumor-Associated Macrophages in the Tumor Microenvironment in Non-Muscle Invasive Bladder Cancer Treated with Intravesical Bacille Calmette-Guérin: A Long-Term Follow-Up Study of a Japanese Cohort. International journal of molecular sciences. 2017;18(10).
111. Winerdal ME, Marits P, Winerdal M, Hasan M, Rosenblatt R, Tolf A, et al. FOXP3 and survival in urinary bladder cancer. BJU Int. 2011;108(10):1672-8.
112. Parodi A, Traverso P, Kalli F, Conteduca G, Tardito S, Curto M, et al. Residual tumor micro-foci and overwhelming regulatory T lymphocyte infiltration are the causes of bladder cancer recurrence. Oncotarget. 2016;7(6):6424-35.
113. Miyake M, Hori S, Owari T, Oda Y, Tatsumi Y, Nakai Y, et al. Clinical Impact of Tumor-Infiltrating Lymphocytes and PD-L1-Positive Cells as Prognostic and Predictive Biomarkers in Urological Malignancies and Retroperitoneal Sarcoma. Cancers (Basel). 2020;12(11).

114. Brauers A, Buettner R, Jakse G. Second resection and prognosis of primary high risk superficial bladder cancer: is cystectomy often too early? J Urol. 2001;165(3):808-10.
115. Sylvester RJ, Oosterlinck W, van der Meijden AP. A single immediate postoperative instillation of chemotherapy decreases the risk of recurrence in patients with stage Ta T1 bladder cancer: a meta-analysis of published results of randomized clinical trials. J Urol. 2004;171(6 Pt 1):2186-90, quiz 435.
116. Morales A, Eidinger D, Bruce AW. Intracavitary Bacillus Calmette-Guerin in the treatment of superficial bladder tumors. J Urol. 1976;116(2):180-3.
117. Oddens J, Brausi M, Sylvester R, Bono A, van de Beek C, van Andel G, et al. Final results of an EORTC-GU cancers group randomized study of maintenance bacillus Calmette-Guérin in intermediate- and high-risk Ta, T1 papillary carcinoma of the urinary bladder: one-third dose versus full dose and 1 year versus 3 years of maintenance. Eur Urol. 2013;63(3):462-72.
118. Malmström PU, Sylvester RJ, Crawford DE, Friedrich M, Krege S, Rintala E, et al. An individual patient data meta-analysis of the long-term outcome of randomised studies comparing intravesical mitomycin C versus bacillus Calmette-Guérin for non-muscle-invasive bladder cancer. Eur Urol. 2009;56(2):247-56.
119. Han RF, Pan JG. Can intravesical bacillus Calmette-Guérin reduce recurrence in patients with superficial bladder cancer? A meta-analysis of randomized trials. Urology. 2006;67(6):1216-23.
120. Böhle A, Bock PR. Intravesical bacille Calmette-Guérin versus mitomycin C in superficial bladder cancer: formal meta-analysis of comparative studies on tumor progression. Urology. 2004;63(4):682-6; discussion 6-7.
121. Sylvester RJ, van der MA, Lamm DL. Intravesical bacillus Calmette-Guerin reduces the risk of progression in patients with superficial bladder cancer: a meta-analysis of the published results of randomized clinical trials. J Urol. 2002;168(5):1964-70.
122. Boehm BE, Cornell JE, Wang H, Mukherjee N, Oppenheimer JS, Svatek RS. Efficacy of bacillus Calmette-Guérin Strains for Treatment of Nonmuscle Invasive Bladder Cancer: A Systematic Review and Network Meta-Analysis. J Urol. 2017;198(3):503-10.

123. Larsen ES, Joensen UN, Poulsen AM, Goletti D, Johansen IS. Bacillus Calmette-Guérin immunotherapy for bladder cancer: a review of immunological aspects, clinical effects and BCG infections. APMIS. 2020;128(2):92-103.
124. Kawai K, Miyazaki J, Joraku A, Nishiyama H, Akaza H. Bacillus Calmette-Guerin (BCG) immunotherapy for bladder cancer: current understanding and perspectives on engineered BCG vaccine. Cancer Sci. 2013;104(1):22-7.
125. Bevers RF, de Boer EC, Kurth KH, Schamhart DH. BCG-induced interleukin-6 upregulation and BCG internalization in well and poorly differentiated human bladder cancer cell lines. Eur Cytokine Netw. 1998;9(2):181-6.
126. Bisiaux A, Thiounn N, Timsit MO, Eladaoui A, Chang HH, Mapes J, et al. Molecular analyte profiling of the early events and tissue conditioning following intravesical bacillus calmette-guerin therapy in patients with superficial bladder cancer. J Urol. 2009;181(4):1571-80.
127. Luo Y. Blocking IL-10 enhances bacillus Calmette-Guérin induced T helper Type 1 immune responses and anti-bladder cancer immunity. Oncoimmunology. 2012;1(7):1183-5.
128. Ratliff TL, Ritchey JK, Yuan JJ, Andriole GL, Catalona WJ. T-cell subsets required for intravesical BCG immunotherapy for bladder cancer. J Urol. 1993;150(3):1018-23.
129. Pichler R, Fritz J, Zavadil C, Schäfer G, Culig Z, Brunner A. Tumor-infiltrating immune cell subpopulations influence the oncologic outcome after intravesical Bacillus Calmette-Guérin therapy in bladder cancer. Oncotarget. 2016;7(26):39916-30.
130. Brausi M, Oddens J, Sylvester R, Bono A, van de Beek C, van Andel G, et al. Side effects of Bacillus Calmette-Guérin (BCG) in the treatment of intermediate- and high-risk Ta, T1 papillary carcinoma of the bladder: results of the EORTC genito-urinary cancers group randomised phase 3 study comparing one-third dose with full dose and 1 year with 3 years of maintenance BCG. Eur Urol. 2014;65(1):69-76.
131. Lamm DL, Blumenstein BA, Crissman JD, Montie JE, Gottesman JE, Lowe BA, et al. Maintenance bacillus Calmette-Guerin immunotherapy for recurrent TA, T1 and carcinoma in situ transitional cell carcinoma of the bladder: a randomized Southwest Oncology Group Study. J Urol. 2000;163(4):1124-9.

132. Larsen ES, Nordholm AC, Lillebaek T, Holden IK, Johansen IS. The epidemiology of bacille Calmette-Guérin infections after bladder instillation from 2002 through 2017: a nationwide retrospective cohort study. BJU Int. 2019;124(6):910-6.
133. van der Meijden AP, Sylvester RJ, Oosterlinck W, Hoeltl W, Bono AV. Maintenance Bacillus Calmette-Guerin for Ta T1 bladder tumors is not associated with increased toxicity: results from a European Organisation for Research and Treatment of Cancer Genito-Urinary Group Phase III Trial. Eur Urol. 2003;44(4):429-34.
134. Holmang S, Johansson SL. Stage Ta-T1 bladder cancer: the relationship between findings at first followup cystoscopy and subsequent recurrence and progression. J Urol. 2002;167(4):1634-7.
135. Fedchenko N, Reifenrath J. Different approaches for interpretation and reporting of immunohistochemistry analysis results in the bone tissue - a review. Diagn Pathol. 2014;9:221.
136. Remmele W, Stegner HE. [Recommendation for uniform definition of an immunoreactive score (IRS) for immunohistochemical estrogen receptor detection (ER-ICA) in breast cancer tissue]. Pathologe. 1987;8(3):138-40.
137. Gwet KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters: Advanced Analytics, LLC; 2014.
138. DG A. Practical statistics for medical research. London: Chapman and Hall/CRC; 1991.
139. Hassler MR, Bray F, Catto JWF, Grollman AP, Hartmann A, Margulis V, et al. Molecular Characterization of Upper Tract Urothelial Carcinoma in the Era of Next-generation Sequencing: A Systematic Review of the Current Literature. Eur Urol. 2020;78(2):209-20.
140. Swartz MA, Porter MP, Lin DW, Weiss NS. Incidence of primary urethral carcinoma in the United States. Urology. 2006;68(6):1164-8.
141. van Kessel KEM, van der Keur KA, Dyrskjøt L, Algaba F, Welvaart NYC, Beukers W, et al. Molecular Markers Increase Precision of the European Association of Urology Non-Muscle-Invasive Bladder Cancer Progression Risk Groups. Clin Cancer Res. 2018;24(7):1586-93.
142. Gontero P, Sylvester R, Pisano F, Joniau S, Oderda M, Serretta V, et al. The impact of re-transurethral resection on clinical outcomes in a large multicentre cohort of patients with T1 high-grade/Grade 3 bladder

cancer treated with bacille Calmette-Guérin. BJU Int. 2016;118(1):44-52.
143. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? Histopathology. 2017;70(1):134-45.
144. Pallua JD, Brunner A, Zelger B, Schirmer M, Haybaeck J. The future of pathology is digital. Pathol Res Pract. 2020;216(9):153040.
145. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. 2019;20(5):e253-e61.
146. Yin PN, Kc K, Wei S, Yu Q, Li R, Haake AR, et al. Histopathological distinction of non-invasive and invasive bladder cancers using machine learning approaches. BMC Med Inform Decis Mak. 2020;20(1):162.
147. Jansen I, Lucas M, Bosschieter J, de Boer OJ, Meijer SL, van Leeuwen TG, et al. Automated Detection and Grading of Non-Muscle-Invasive Urothelial Cell Carcinoma of the Bladder. Am J Pathol. 2020;190(7):1483-90.

Appendices

I: Multiclass tissue classification of whole-slide histological images using convolutional neural networks.

II: Multiscale deep neural networks for multiclass tissue classification of histological whole-slide images.

III: A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides.

IV: Automatic diagnostic tool for predicting cancer grade in bladder cancer patients using deep learning.

Appendices

I: Multiclass tissue classification of whole-slide histological images using convolutional neural networks.

Multiclass Tissue Classification of Whole-Slide Histological Images using Convolutional Neural Networks

Rune Wetteland[1], Kjersti Engan[1], Trygve Eftestøl[1], Vebjørn Kvikstad[2], Emilius A.M. Janssen[2,3]
[1]*Department of Electrical Engineering and Computer Science, University of Stavanger, Norway*
[2]*Department of Pathology, Stavanger University Hospital, Norway*
[3]*Department of Mathematics and Natural Sciences, University of Stavanger, Norway*
{*rune.wetteland, kjersti.engan, trygve.eftestol*}*@uis.no,*
{*vebjorn.kvikstad, emilius.adrianus.maria.janssen*}*@sus.no*

Keywords: Histological Whole-Slide Images, Autoencoder, Deep Learning, Semi-Supervised Learning, ROI Extraction.

Abstract: Globally there has been an enormous increase in bladder cancer incidents the past decades. Correct prognosis of recurrence and progression is essential to avoid under- or over-treatment of the patient, as well as unnecessary suffering and cost. To diagnose the cancer grade and stage, pathologists study the histological images. However, this is a time-consuming process and reproducibility among pathologists is low. A first stage for an automated diagnosis system can be to identify the diagnostical relevant areas in the histological whole-slide images (WSI), segmenting cell tissue from damaged areas, blood, background, etc. In this work, a method for automatic classification of urothelial carcinoma into six different classes is proposed. The method is based on convolutional neural networks (CNN), firstly trained unsupervised using unlabelled images by utilising an autoencoder (AE). A smaller set of labelled images are used to train the final fully-connected layers from the low dimensional latent vector of the AE, providing an output as a probability score for each of the six classes, suitable for automatically defining regions of interests in WSI. For evaluation, each tile is classified as the class with the highest probability score. The model achieved an average F1-score of 93.4% over all six classes.

1 INTRODUCTION

Globally, bladder cancer resulted in 123,400 deaths in 1990, and in 2010 this number was 170,700 which is an increase of 38,3% taking population growth into consideration (Lozano et al., 2012). The majority of bladder cancer incidents are urothelial carcinoma with a representation as high as 90% in some regions (Eble et al., 2004). For patients diagnosed with bladder cancer, 50-70% will experience one or more recurrences, and 10-30% will have disease progression to a higher stage (Mangrud, 2014). Patient treatment, follow-up and calculating the risk of recurrence and disease progression depend primarily on the histological grade and stage of cancer. Correct prognosis of recurrence and progression is essential to avoid under- or over-treatment of the patient, as well as unnecessary suffering and cost.

With the introduction of digital pathology, some computer-aided tools to assist pathologists have been introduced, but still the assessment of histopathological images to diagnose, grade and stage cancer is mainly done manually. This is a time-consuming process and reproducibility among pathologists is in some cases low, for example within the prognostic classification of urinary bladder cancer. Automatic extraction of the relevant areas in large whole-slide images (WSI) would be an important first step where the results could be used in automated diagnostic and prognostic classification tools.

During the biopsy, parts of the tissue get both physical- and heating-damage, and thus can not be used as relevant diagnostic information. The WSI also contains stroma- and muscle-tissue as well as areas of blood. In this paper we consider the task of automatic classification of tiles in WSI into the six different classes; urothelium, stroma, damaged tissue, muscle, blood and background. Examples from each class are shown in Figure 1. The system uses the automatic classification tool to produce heat maps from the model's output. Such heat maps can provide useful information to help the pathologist to focus on the diagnostic important part of the large WSI during visual inspection. In addition, the heat maps are also suitable as input for automatic region of interest (ROI) extraction of relevant areas in the WSI, which can fur-

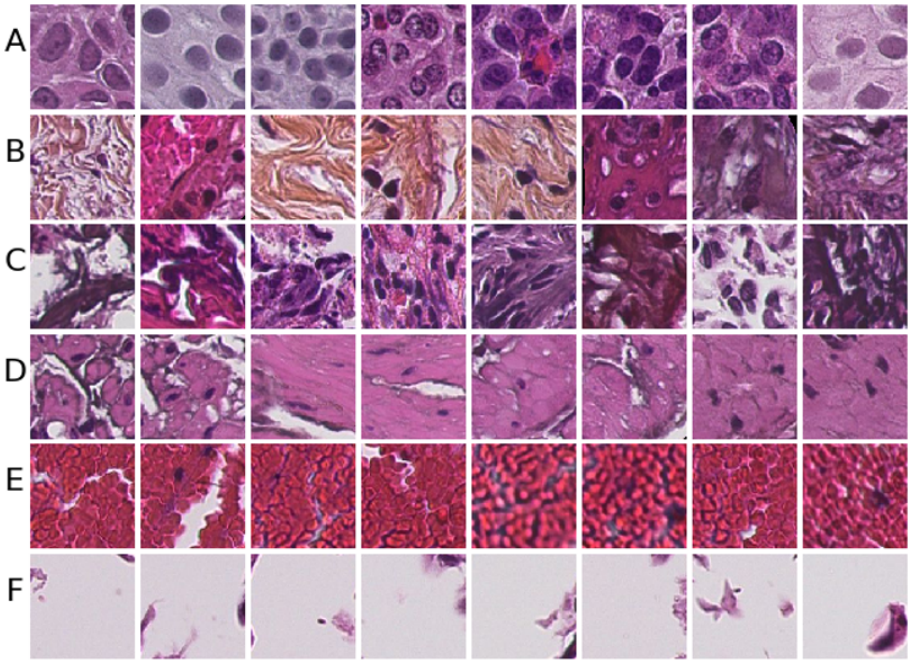

Figure 1: Example tiles from each class. A) Urothelium, B) Stroma, C) Damaged tissue, D) Muscle tissue, E) Blood, and F) Background.

ther be used in automated diagnostic and prognostic classification tools.

1.1 Previous Work

In recent literature, some methods for automatic tissue classification have been suggested. However, most previous works have focused on classifying only two classes, a binary problem set to differentiate between cancer-patches and non-cancer patches.

Recent literature shows good results for binary tissue classification using convolutional neural networks (CNN). Wang et al. (2016) won both competitions of the Camelyon16 grand challenge for automated detection of metastatic breast cancer in WSI. As part of their model, GoogLeNet was utilised to do patch classification. The model was trained to discriminate between positive and negative patches and achieved an accuracy of 98.4%.

Some attempts of multiclass tissue classification can be found in recent years. Araujo *et al.* classified patches of breast cancer into four classes using convolutional neural networks (Araújo et al., 2017). The best patch-wise accuracy for four classes was 66.7%. When the task was simplified as a two-classes problem, non-carcinoma vs carcinoma, the accuracy was improved to 77.6%. The work of Kather et al. (2016) uses a combination of several hand-crafted feature methods to classify different types of tissue in colorectal cancer, performing tests on both a two-class and eight-class problem. They achieved the best result on the two-class problem with a tumour-stroma separation accuracy of 98.6%, while the multiclass problem achieved an accuracy of 87.4%.

To the author's knowledge, there are no published results on multiclass classification on WSI of bladder cancer.

Some few and recent work on ROI detection can be found. ROI detection has been done by multi-scale real-time coarse-to-fine topology preserving segmentation (CTFTPS) by utilising superpixel clustering technique (Li and Huang, 2015; Yao et al., 2015). A RAPID (Regular and Adaptive Prediction-Induced Detection) segmentation method for ROI detection in large WSI is presented by Sulimowicz and Ahmad (2017) while using the multi-scale CTFTPS technique as a baseline. An SVM was utilised to classify the detected regions as ROI vs non-ROI. For this task, the classifier achieved an F1-score of 89.8% for the RAPID method, and 91.2% for the optimised multi-scale CTFTPS method.

Deep CNN has shown to provide state of the art results in many computer vision tasks in recent years (LeCun et al., 2015) and has also found its way into medical image assessment tasks. In this work, a method for automatic classification of WSI from urothelial carcinoma into six different classes is proposed. The method is based on CNN, firstly trained

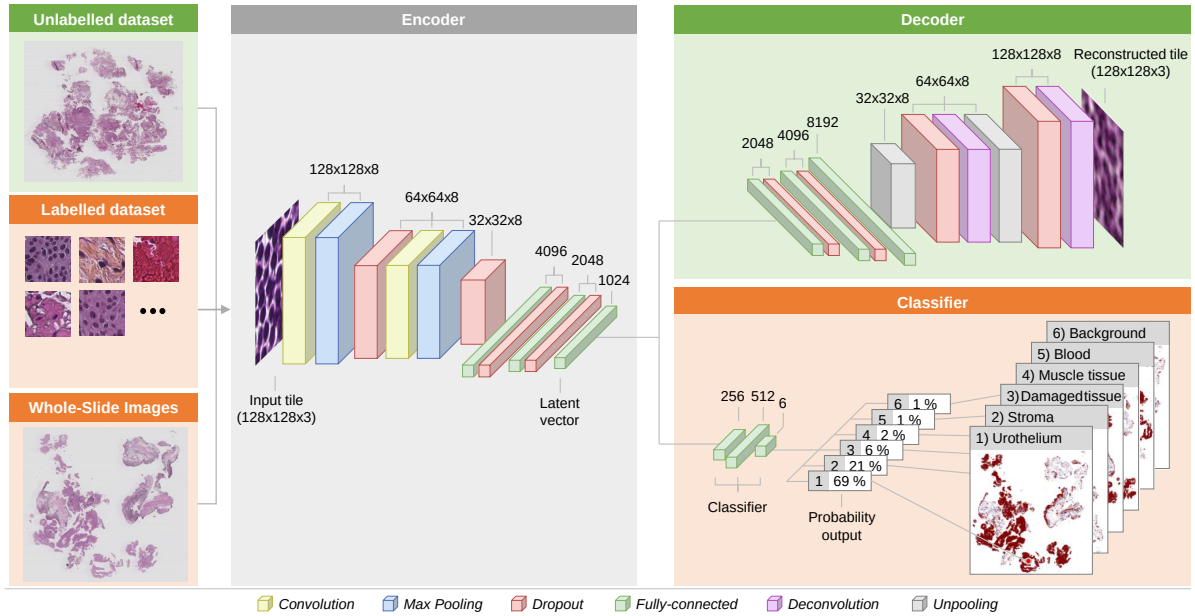

Figure 2: Overview of the CNN-model. First, the unlabelled dataset is used to train the encoder-decoder model. Then the labelled dataset is used to train the encoder-classifier model. Finally, the trained encoder-classifier model are used to classify new WSI into probability maps. These probability maps are further postprocessed to produce the heat maps.

unsupervised, using large unlabelled image sets by utilising an autoencoder (AE). A set of labelled images are used to train the final fully-connected layers from the low dimensional latent vector of the AE, providing an output as a probability score for each of the six classes, suitable for automatically defining ROI in WSI. A visualisation of the system is depicted in Figure 2.

The novelty of the work lies both in the specific application of urinary bladder WSI and in the method development, more specifically in a combination of using CNN, learned in a semi-supervised way, for the application of automatic region of interest extraction in WSI by *multiclass* tissue classification, tested on urinary bladder cancer.

2 DATA MATERIAL

The data material used in this paper consists of histopathological images from patients with primary bladder cancer, collected in the period 2002-2011 at the University Hospital of Stavanger in Norway. The biopsies are formalin fixed and paraffin embedded, 4 μm slides are cut and stained with Hematoxylin Eosin Saffron (HES). All slides are diagnosed and graded according to WHO73 and WHO04, cancer stage (Tis, Ta or T1) and follow-up data on recurrence and disease progression are recorded.

The slides are then scanned using a Leica SCN400 histological slide scanner to produce a digital histological image. The images are in Leicas data format called SCN and to be able to process these images the Vips library (Martinez and Cupitt, 2005) has been used, which is specially designed for image processing of large images.

3 PROPOSED METHOD

An overview of the proposed method can be seen in Figure 2. The different parts will be explained in this section.

3.1 Preprocessing

Each WSI is sliced into smaller non-overlapping tiles of size 128x128 pixels, extracted at 400x magnification level. The background takes up as much as 70-80% of the WSI and is detected and discarded automatically by computing the histogram of the tile and setting a fixed threshold value. This removes tiles consisting of grey background, however, if the background tile contains small parts of debris, tissue or similar it is not discarded. Examples of tiles belonging to this class are illustrated in Figure 1-F.

The histological images are split into three datasets. First, an unlabelled dataset is created in the manner explained above where the extracted tiles have no label associated with it. In total 48 WSI all

from different patients were preprocessed resulting in 7,130,527 unlabelled tiles after the pure background tiles are excluded. This set, called *train-ae*, is utilised as training data for the AE-model.

Secondly, a labelled training dataset is created. A pathologist has manually annotated carefully selected regions in the WSI. The tiles in the regions are preprocessed by evaluating the histogram to be sure not to include background or boundaries and given a label corresponding to its class. The number of patients and tiles produced are listed as *train-set1* in Table 1.

Lastly, a labelled test set is created to assess the performance of the classifier. The set is created in the same manner as the labelled training set, but on separate WSI which has not been used in either the unlabelled or labelled datasets to avoid cross-contamination between training and test data. The dataset is listed as *test-set* in Table 1.

The texture of urothelium tissue will change for the different cancer grades, and thus it is vital to include a wide variety of samples for this class. The other five classes, however, will not change as a function of cancer grade and may include fewer samples. Another issue is that the occurrence of some classes is more sparse in the WSI, making it difficult to extract a large amount of it. A disadvantage of these two issues is a significant deviation in the number of samples in two of the classes, stroma and muscle tissue, as seen in *train-set1* in Table 1.

To compensate for the class-imbalance in *train-set1*, data augmentation techniques have been utilised. Tiles in the muscle and stroma class are extracted with 50% overlap, to produce more data from the same regions. These extracted tiles are further augmented by randomly flipping and rotating them to create new data. These techniques result in a more balanced dataset, which is listed in Table 1 as *train-set2*. This dataset is used to train the classifier in the presented experiments. The augmentation techniques were not performed on the *test-set*, resulting in an unbalanced test set. In this case, accuracy as a performance metric could be misleading. Instead, precision, recall and F1-score are used to evaluate the performance.

Table 1: The resulting labelled datasets after preprocessing. Results show the total number of tiles extracted for each class, and the number of WSI used are shown in parenthese.

	Train-set1	**Train-set2**	**Test-set**
Urothelium	25,635 (25)	25,635 (25)	3,612 (3)
Stroma	4,329 (4)	25,974 (4)	505 (1)
Damaged	30,714 (8)	30,714 (8)	2,679 (1)
Muscle	2,002 (3)	23,949 (3)	475 (1)
Blood	19,071 (4)	19,071 (4)	692 (1)
Background	20,000 (2)	20,000 (2)	500 (1)

3.2 CNN-Model

The system consists of an autoencoder model which is trained on the unlabelled dataset *train-ae*. The autoencoder consists of two main parts; the encoder and the decoder. The encoder will transform the input tile into a latent vector of much lower dimension. A small latent space is chosen which will force the network to extract the essential features of the image and preserve these in the vector. The decoder will use the features stored in the latent vector and reconstruct the input. During training, the network compares the reduced mean of the squared difference between the input image and reconstructed output image as given by the loss function $\sum(input - output)^2$. The AE function is described in details in (Baldi, 2012). The encoder consists of two convolutional-, two max-pooling- and four dropout-layers, as well as three fully-connected layers as seen in Figure 2. The decoder consists of the same layers, but in reverse order and uses unpooling and deconvolutional layers instead.

After training, the encoder has learned to extract the features of the input tile, which are now stored in the latent vector. To do classification, the decoder part is discarded and exchanged with a classifier. The classifier consists of three fully-connected layers connected to the output of the encoder. This encoder-classifier model constitutes the proposed CNN-model and is trained on the labelled training dataset *train-set2* and evaluated on the *test-set*.

For initialisation of the system, the bias is set to zero, and the weights are taken from a truncated normal distribution. The convolutional layers use a filter kernel of 3x3 and a stride of 1, whereas the max-pooling layers use a filter kernel of 2x2 with a stride of 2. The number of feature maps is used to control the size of the latent vector space and is experimented on as described in section 4. The parameters of the network are optimised using the Adam optimiser with a mini-batch of size 128. For the activation function between layers, the Rectified linear unit (ReLU) activation function is used. For the last layer, the Softmax activation function is utilised. This will output a true probability distribution, meaning each output lays in the interval 0 to 1 and all outputs combined sums up to one. Dropout is a technique where randomly selected nodes are set to zero during training to provide regularisation to the network. The portion of nodes set to zero is specified by the dropout rate as a percentage. During evaluation of the network, dropout is disabled.

The histological images are in Leicas data format called SCN and to be able to process these images

the Vips library (Martinez and Cupitt, 2005) has been used. This is a library specially designed for image processing of large images. The model is written in Python 3.5 using the Tensorflow 1.7 machine learning library (Abadi et al., 2016). For evaluation of the model, the Scikit-learn metric package (Pedregosa et al., 2011) is used which computes precision, recall and F1-score of each class in addition to an average total score.

The model is used to predict the class of each tile in a WSI. The probability for each class provided by the model can be rearranged as probability maps, one for each class, and will visualise the location in the histological image where each class is present. An overview of this process is presented in Figure 2.

4 EXPERIMENTS AND RESULTS

Two experiments were conducted, the first to find the best combination of architecture and hyperparameters and the second to verify its performance and use the final model on WSI.

4.1 Experiment 1: Architecture and Hyperparameters

To find a suitable architecture and appropriate hyperparameters, a large grid search was conducted. To reduce both computational time and search space, a preliminary search was set up with some limitations. A reduced version of the *train-ae* dataset was used to decrease the processing time, and each model was only trained for 50 epochs.

The encoder-decoder model was tested with two different sizes of the latent vector, which was altered by changing the number of feature maps in the convolutional layers. Latent vectors of size 512 and 1024 were tested. A learning rate of 10^{-3} and 10^{-4} was tested as well as dropout rates of 0%, 10% and 20%. Each of these combinations was tested on network configuration consisting of two, four and six convolutional layers in the autoencoder.

In the encoder-classifier model, the classifier consists of three dense layers. The first layer after the encoder was tested with 256, 512 and 1024 neurons, and the second layer with 128, 256 and 512 neurons. The number of neurons in the output layer is bounded to the number of classes. This results in 9 different configurations for the classifier layers. Each of these configurations was tested with a learning rate of 10^{-3}, 10^{-4} and 10^{-5}. There are no dropout layers in the classifier itself, but changing the dropout rate will affect how the encoder codes the input tile into the latent vector. The encoder-classifier was therefore also tested with the same dropout rates as above. The model was tested both with and without freezing the pre-trained encoder-layers to see how it affected the result.

The prediction accuracy on the *test-set* was used to compare the performance of the different hyperparameter combinations. Hyperparameters that showed poor performance on several models were excluded to narrow down the search space.

The experiments showed an overall best result using an encoder-decoder structure with two convolutional layers with a latent vector of 1024 neurons trained with 10^{-4} learning rate and 10% dropout rate. The results further showed best performance while not freezing the encoder part of the encoder-classifier model. A classifier with 256 neurons in the first layer and 512 in the second layer was favourable, trained using a learning rate of 10^{-5} and 10% dropout rate. These hyperparameters and settings will be used as the resulting model of this experiment. The model is depicted in Figure 2.

4.2 Experiment 2: Training, Testing and using the Resulting Model

The resulting architecture after the first experiment was trained once more, this time on the full dataset. First, the autoencoder was trained on the unlabelled dataset *train-ae* for 100 epochs, then the encoder-classifier was fine-tuned on the augmented labelled dataset *train-set2* for another 600 epochs. Since experiment 1 showed best results when the encoder was not frozen during fine-tuning, both the encoder and classifier was trained during this step. Evaluation using the Scikit-learn metric package on the *test-set* was performed every 5th epoch. The model achieved the best result after 540 epochs of training with an average F1-score of 93.4% over all six classes. The precision, recall and F1-score of each class is shown in Table 2.

Table 2: Detailed classification results from the model trained using 10% dropout rate.

Class	Precision	Recall	F1-Score
Urothelium	0.924	0.952	0.938
Stroma	0.897	0.929	0.913
Damaged	0.925	0.927	0.926
Muscle	0.980	0.714	0.826
Blood	0.996	0.991	0.994
Background	0.990	0.988	0.989
Average total	0.936	0.935	0.934

The overall results in Table 2 are good. However,

there are some observations.

In *train-set2*, which is used to train the classifier, the classes of blood and background have the fewest number of samples. However, these are the classes which perform best. This is probably because these classes have the least within-class variance, e.g. most of the tiles have a similar visual appearance.

Urothelium and damaged tissue both perform well, even though these classes have a substantial visual variance in the form of colour and texture in the tiles. The dataset for these classes contains the most number of patients (25 and 8 patients, respectively), and therefore contains the most diverse samples in the dataset, contributing to the good results.

The precision of stroma and recall of muscle is not performing as good as the rest. The dataset for these classes contains few patients and are also the two classes which needed augmentation due to small amounts of available data. The low recall of muscle tissue indicates that a large proportion of the muscle tiles are misclassified as other classes, most probably urothelium, stroma and damaged tissue (due to the high precision of blood and background, these are not likely to include many misclassified tiles). It is important to note that the muscle class achieves a very good precision score, and stroma has an acceptable good recall score.

4.3 Heat Maps

The resulting model was utilised to classify entire whole-slide images. Each tile in the WSI was classified and the percentage for each class recorded. These were then combined to create the probability maps. These maps were then post-processed in MATLAB by applying a Gaussian filter kernel with a standard deviation of $\sigma = 0.6$ to smooth the images. After filtering, a thresholding operation was performed on the image with a limit of 0.8, setting all predictions below this threshold to zero. This ensures that only predictions of 0.8 or higher are visible in the final heat maps.

Figure 3 shows three example WSI with their corresponding heat maps. By visual inspection performed by pathologists, this is considered to look very promising. However, a quantitative measure for the WSI ROI extraction is lacking since we do not have complete WSI manually labelled into the six classes at the current time.

5 CONCLUSION

This paper proposes a method for automatic classification of tile-segments of histopathological WSI of urinary bladder cancer into six different classes using a CNN-based model. An encoder-decoder structure is trained on a large set of unlabelled data. After training, the encoder part of the autoencoder acts as a feature extractor making low dimensional latent vectors. An encoder-classifier structure is then fine-tuned on a set of labelled tiles. The finished model is able to classify input tiles from the WSI into the classes urothelium, stroma, damaged tissue, muscle, blood and background. The best model achieved an average F1-score of 93.4% over all the six classes, an overall good result. However, future work will include an effort to improve the classifier. Other methods such as a multiscale approach are considered.

The model is further used to classify entire WSI to produce heat maps, which visualises each of the classes and their location in the image. These maps can provide useful information to the pathologist during visual inspection. Future work consists of using the above model as an ROI extractor of relevant tissue in the WSI to make a dataset suitable as training data for a diagnostic and prognostic classification model.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., and Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6):e0177544.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49.

Eble, J. N., Sauter, G., Epstein, J. I., and Sesterhenn, I. A. (2004). World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs. *IARC Press: Lyon*.

Kather, J. N., Weis, C.-A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., Marx, A., and Zöllner, F. G. (2016). Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Li, R. and Huang, J. (2015). Fast regions-of-interest detection in whole slide histopathology images. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 120–127. Springer.

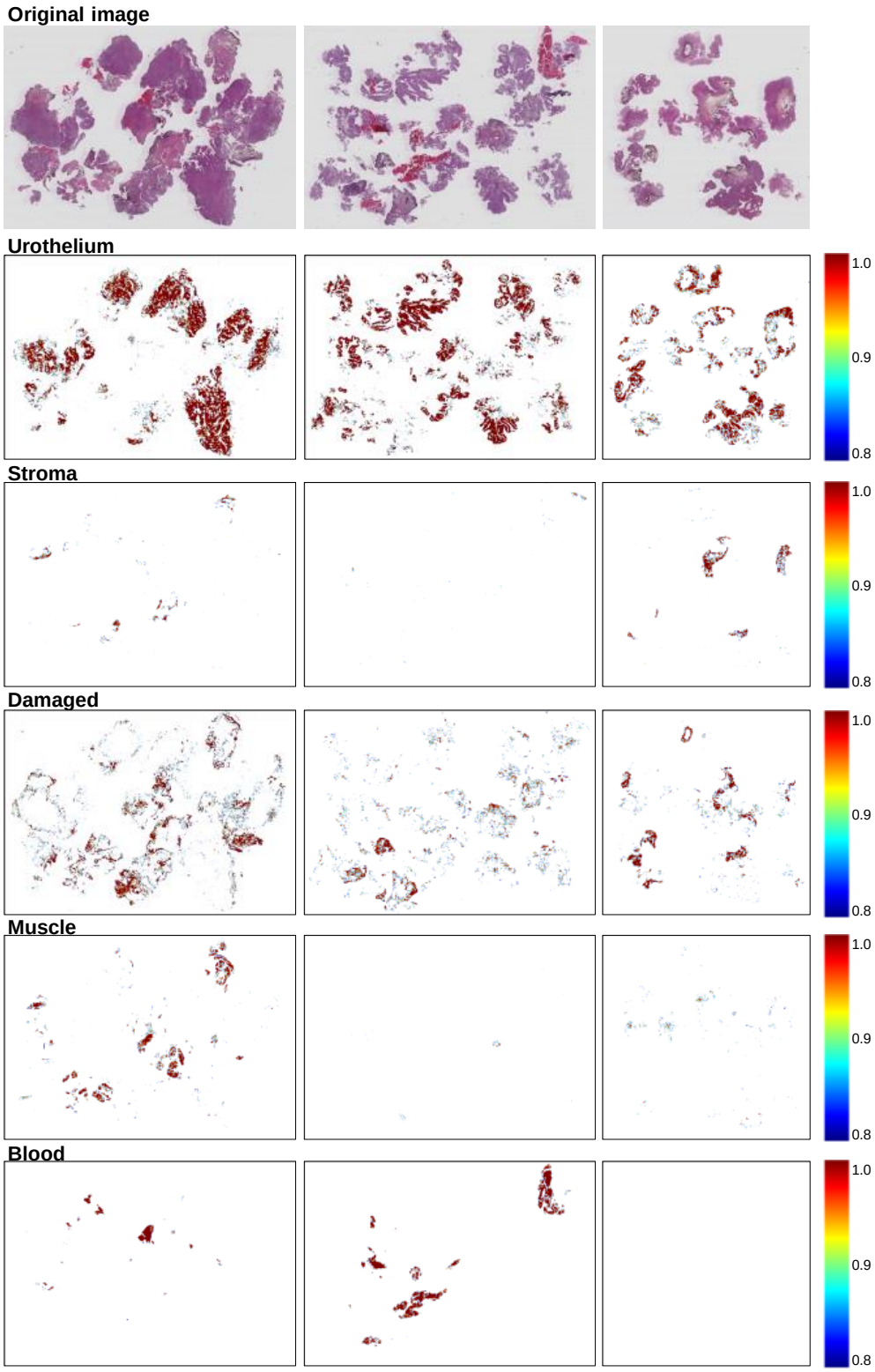

Figure 3: The original WSI together with the corresponding heat maps. The scale in the rightmost column shows the confidence level given by the model. The background heat maps are performing very good, but has been omitted from the heat map visualisation since it is just removing the borders between background and tissue. The heat maps have been smoothed with a Gaussian filter and thresholded to only contain predictions of 0.8 and higher.

Lozano, R., Naghavi, M., Foreman, K., and Lim, S. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859):2095–2128.

Mangrud, O. (2014). *Identification of patients with high and low risk of progresson of urothelial carcinoma of the urinary bladder stage Ta and T1*. PhD thesis, Ph. D. dissertation, University of Bergen.

Martinez, K. and Cupitt, J. (2005). Vips-a highly tuned image processing software architecture. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–574. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Sulimowicz, L. and Ahmad, I. (2017). rapid regions-of-interest detection in big histopathological images. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 595–600. IEEE.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Yao, J., Boben, M., Fidler, S., and Urtasun, R. (2015). Real-time coarse-to-fine topologically preserving segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2947–2955.

Appendices

I: Multiclass tissue classification of whole-slide histological images using convolutional neural networks.

II: Multiscale deep neural networks for multiclass tissue classification of histological whole-slide images.

Multiscale Deep Neural Networks for Multiclass Tissue Classification of Histological Whole-Slide Images

Rune Wetteland[1] RUNE.WETTELAND@UIS.NO
Kjersti Engan[1] KJERSTI.ENGAN@UIS.NO
Trygve Eftestøl[1] TRYGVE.EFTESTOL@UIS.NO
Vebjørn Kvikstad[2] VEBJORN.KVIKSTAD@SUS.NO
Emilius A.M. Janssen[2,3] EMILIUS.ADRIANUS.MARIA.JANSSEN@SUS.NO
[1] *Department of Electrical Engineering and Computer Science, University of Stavanger, Norway*
[2] *Department of Pathology, Stavanger University Hospital, Norway*
[3] *Department of Mathematics and Natural Sciences, University of Stavanger, Norway*

arXiv:1909.01178v1 [eess.IV] 3 Sep 2019

Abstract

Correct treatment of urothelial carcinoma patients is dependent on accurate grading and staging of the cancer tumour. This is determined manually by a pathologist by examining the histological whole-slide images (WSI). The large size of these images makes this a time-consuming and challenging task. The WSI contain a variety of tissue types, and a method for defining diagnostic relevant regions would have several advantages for visualization as well as further input to automated diagnosis systems. We propose an automatic multiscale method for classification of tiles from WSI of urothelial carcinoma patients into six classes. Three architectures based on convolutional neural network (CNN) were tested: MONO-CNN (400x), DI-CNN (100x/400x) and TRI-CNN (25x/100x/400x). The preliminary results show that the two multiscale models performed significantly better than the mono-scale model, achieving an F1-score of 0.986, substantiating that utilising multiple scales in the model aids the classification accuracy.

1. Introduction

Bladder cancer is the 10th most common cancer type worldwide (Bray et al., 2018). More than 90% of bladder cancer cases are urothelial carcinomas which has a particular high recurrence (50-70%) and progression rate (10-30%), making correct treatment and follow-up vital for survivability. Treatment is dependent on the cancer grade and stage, determined manually by an expert pathologist examining the histological whole-slide images (WSI). This is a time-consuming and challenging task, and studies have shown that it may have a low reproducibility in some cases, such as grading of urothelial carcinoma (Mangrud, 2014).

Examination of the WSI is challenging because of the large size of the image, which contains several different tissue types, where only some are useful for diagnostic information. An automatic tool for identification of such regions would be beneficial for both guiding a pathologist to the useful areas of the large WSI during examination, and for ROI extraction of useful tissue for a computer aided diagnostic solution. In this paper we present an automatic method for classification of tiles from WSI of urothelial carcinoma patients into the classes: urothelium, stroma, muscle, damaged tissue, blood and background. The tiles

© 2019 R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad & E. A.M. Janssen.

are extracted at different magnification levels, to combine and utilise information at different scales in a similar fashion to that of a pathologist.

Multiscale approaches to tile-based classification have previously been done on other cancer types. In the work of Li et al. (2017) a multiscale U-Net was proposed for segmentation of histological images from radical prostatectomies to classify tiles into four classes. Tiles of size 100x100, 200x200 and 400x400 pixels were all extracted from histological images at 200x magnification. Features from the different tiles were then concatenated and used as input to the multiscale U-net. The model achieved a mean Jaccard index of 65.8% over the four classes. In Sirinukunwattana et al. (2018) a comparison of five single-scale and five multiscale architectures were tested on two datasets. Their best model (G) was a multiscale model which achieved an average F1-score of 0.782±0.07 across four classes of prostate cancer and 0.538±0.08 across four classes of breast cancer. Their result supports the claim that incorporating a larger visual context improves the results. In Wetteland et al. (2019) we presented a method based on deep convolutional neural networks (CNN) for classifying tiles of urothelial carcinoma WSI into the six classes mentioned above. This was a single-scale approach where all tiles were extracted from the full resolution image of 400x magnification. The method got an F1-score of 0.934±0.061.

2. Data Material

The data material consists of Hematoxylin Eosin Saffron (HES) stained WSI from patients diagnosed with primary papillary urothelial carcinoma, collected at the University Hospital of Stavanger, Norway. An expert pathologist has carefully annotated 239 selected regions from 50 WSI from 32 unique patients, where each region includes one of the five foreground classes. Regions belonging to the background class was annotated on seven randomly selected patients.

Tiles were extracted from these regions at 25x, 100x and 400x magnification in such a manner that the centre pixel is the same in all three tiles. All tiles have the same size of 128x128x3 pixels. Tiles belonging to the test set was extracted from patients not present in the training data. The remaining data was augmented to balance the dataset and was further randomly shuffled and split into 85% training and 15% validation data. A random seed was set to ensure that the shuffling was the same for each model. The final datasets consist of 128K training tiles, 23K validation tiles and 11K test tiles.

3. Method and Results

This paper compares three architectures referred to as the MONO-, DI- and TRI-CNN model. The three architectures have one (400x), two (100x, 400x) and three (25x, 100x, 400x) inputs, respectively. Each input is fed into a pre-trained VGG16 network (Simonyan and Zisserman, 2014) which acts as a feature extractor. The fully-connected (FC) layers of VGG16 are replaced with a classification network consisting of two FC-layers, each followed by a dropout layer, and a final softmax layer with one output node for each of the six classes. The DI-CNN and TRI-CNN models have two and three parallel VGG16 branches, respectively, which are concatenated before entering the classification network.

The FC-layers were tested with 512, 1024, 1536, 2048 and 4096 neurons, and dropout rates of 0, 0.3 and 0.5. This 15-model hyperparameter search was conducted on each of the three architectures, resulting in 45 models. These 45 models were run three consecutive times and averaged together for a more accurate result. Each model was trained using early stopping, stopping the model if validation loss did not decrease within 30 epochs. All model selections were based on the validation set performance. After training, the weight parameters from the epoch which performed best on the validation dataset were restored, and a final evaluation of the model was performed on the test dataset. The VGG16 networks had their weight parameters frozen during training. The model was written in Python 3.5 using the Keras machine learning library (Chollet et al., 2015).

Table 1 shows the hyperparameters for the best performing models and their average result from the three consecutive runs. The MONO-CNN model achieves a result similar to that of the autoencoder. The two multiscale models perform equally and significantly better than the mono-scale models. The multiscale models also have a lower standard deviation on all metrics. Since both multiscale models achieve the same result, one could argue that the simplest model of the two should be chosen. In that case, DI-CNN with its 36M parameters is a simpler model than TRI-CNN which has 47M parameters in total. DI-CNN also have a marginally lower standard deviation than TRI-CNN.

Table 1: Models evaluated on the test set. F1-Score is presented as the total average and standard deviation calculated across all six classes over three consecutive runs. Parameters are shown as no. of trainable parameters / no. of total parameters.

Model	Input Scale	Dropout	FC-Neurons	# Parameters	F1-Score
Autoencoder[1]	400x	0.1	256/512	89M/89M	0.934 ± 0.061
MONO	400x	0.3	2048	5.3M/20M	0.944 ± 0.007
DI	100x/400x	0.0	2048	6.3M/36M	0.986 ± 0.002
TRI	25x/100x/400x	0.5	1024	2.6M/47M	0.986 ± 0.003

4. Conclusion

In this paper, we present preliminary results from a multiscale tile-based classification model. Tiles from six classes were extracted at multiple scales from WSI of patients diagnosed with urothelial carcinoma. Three model architectures were compared: MONO-CNN (400x), DI-CNN (100x, 400x) and TRI-CNN (25x, 100x, 400x). Results for an autoencoder model from previous work was also included for reference. Both multiscale models outperform the two single-scale models and achieve a very good result indicating the advantage of utilising multiple scales. The model can be used as an ROI extraction method for relevant tissue areas in the large WSI, useful for both pathologist and computer-aided diagnostic systems. Some more experiments should be performed to clarify if the behaviour stems from the multiscale approach or the extended field-of-view.

1. Model trained and evaluated on the same dataset (Wetteland et al., 2019).

References

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries: Global cancer statistics 2018. *CA: A Cancer Journal for Clinicians*, 68, 09 2018. doi: 10.3322/caac.21492.

François Chollet et al. Keras, 2015.

Jiayun Li, Karthik V Sarma, King Chung Ho, Arkadiusz Gertych, Beatrice S Knudsen, and Corey W Arnold. A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1140. American Medical Informatics Association, 2017.

Ok Målfrid Mangrud. *Identification of patients with high and low risk of progresson of urothelial carcinoma of the urinary bladder stage Ta and T1.* PhD thesis, Ph. D. dissertation, University of Bergen, 2014.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Korsuk Sirinukunwattana, Nasullah Khalid Alham, Clare Verrill, and Jens Rittscher. Improving whole slide segmentation through visual context-a systematic study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–200. Springer, 2018.

Rune Wetteland, Kjersti Engan, Trygve Eftestøl, Vebjørn Kvikstad, and Emilius A. M. Janssen. Multiclass tissue classification of whole-slide histological images using convolutional neural networks. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM*, pages 320–327. INSTICC, SciTePress, 2019. ISBN 978-989-758-351-3. doi: 10.5220/0007253603200327.

Appendices

I: Multiclass tissue classification of whole-slide histological images using convolutional neural networks.

II: Multiscale deep neural networks for multiclass tissue classification of histological whole-slide images.

III: A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides.

Check for updates

Original Article

A Multiscale Approach for Whole-Slide Image Segmentation of five Tissue Classes in Urothelial Carcinoma Slides

Technology in Cancer Research & Treatment
Volume 19: 1-15
© The Author(s) 2020
DOI: 10.1177/1533033820946787
journals.sagepub.com/home/tct
SAGE

Rune Wetteland, MS[1], Kjersti Engan, PhD[1], Trygve Eftestøl, PhD[1], Vebjørn Kvikstad, MD[2,3], and Emiel A. M. Janssen, PhD[2,3]

Abstract

In pathology labs worldwide, we see an increasing number of tissue samples that need to be assessed without the same increase in the number of pathologists. Computational pathology, where digital scans of histological samples called whole-slide images (WSI) are processed by computational tools, can be of help for the pathologists and is gaining research interests. Most research effort has been given to classify slides as being cancerous or not, localization of cancerous regions, and to the "big-four" in cancer: breast, lung, prostate, and bowel. Urothelial carcinoma, the most common form of bladder cancer, is expensive to follow up due to a high risk of recurrence, and grading systems have a high degree of inter- and intra-observer variability. The tissue samples of urothelial carcinoma contain a mixture of damaged tissue, blood, stroma, muscle, and urothelium, where it is mainly muscle and urothelium that is diagnostically relevant. A coarse segmentation of these tissue types would be useful to i) guide pathologists to the diagnostic relevant areas of the WSI, and ii) use as input in a computer-aided diagnostic (CAD) system. However, little work has been done on segmenting tissue types in WSIs, and on computational pathology for urothelial carcinoma in particular. In this work, we are using convolutional neural networks (CNN) for multiscale tile-wise classification and coarse segmentation, including both context and detail, by using three magnification levels: 25x, 100x, and 400x. 28 models were trained on weakly labeled data from 32 WSIs, where the best model got an F1-score of 96.5% across six classes. The multiscale models were consistently better than the single-scale models, demonstrating the benefit of combining multiple scales. No tissue-class ground-truth for complete WSIs exist, but the best models were used to segment seven unseen WSIs where the results were manually inspected by a pathologist and are considered as very promising.

Keywords

bladder cancer, multiscale classification, tissue segmentation, deep learning, CNN, histological images

Abbreviations

AI, artificial intelligence; CAD, computer-aided diagnostic; CNN, convolutional neural network; CV, cross-validation; FC, fully-connected; FCN, fully convolutional networks; FN, false negative; FP, false positive; GAP, global average pooling; HES, Hematoxylin Eosin Saffron; IBD, inflammatory bowel disease; MDRAN, multiscale deep residual aggregation network; NSCLC, non-small cell lung cancer; ROI, region of interest; SGD, stochastic gradient descent; SVM, Support vector machine; TMB, tumor mutational burden; TP, true positive; WHO, World health organization; WSI, whole-slide image

Received: February 24, 2020; Revised: June 3, 2020; Accepted: June 19, 2020.

Introduction

Worldwide, 549 393 new cases of bladder cancer were diagnosed in 2018, in addition there were 199 922 deaths due to the disease. This makes bladder cancer the 10th most common type of cancer in the world.[1] Men are overrepresented, with approximately 75% of the cases.[2] The most common type of bladder cancer is urothelial carcinoma, with over 90% of the cases.[3] Of the patients diagnosed with bladder cancer, 50% to 70% will

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway
[2] Department of Pathology, Stavanger University Hospital, Norway
[3] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Norway

Corresponding Author:
Rune Wetteland, Department of Electrical Engineering and Computer Science, University of Stavanger, 4036 Stavanger, Norway.
Email: rune.wetteland@uis.no

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

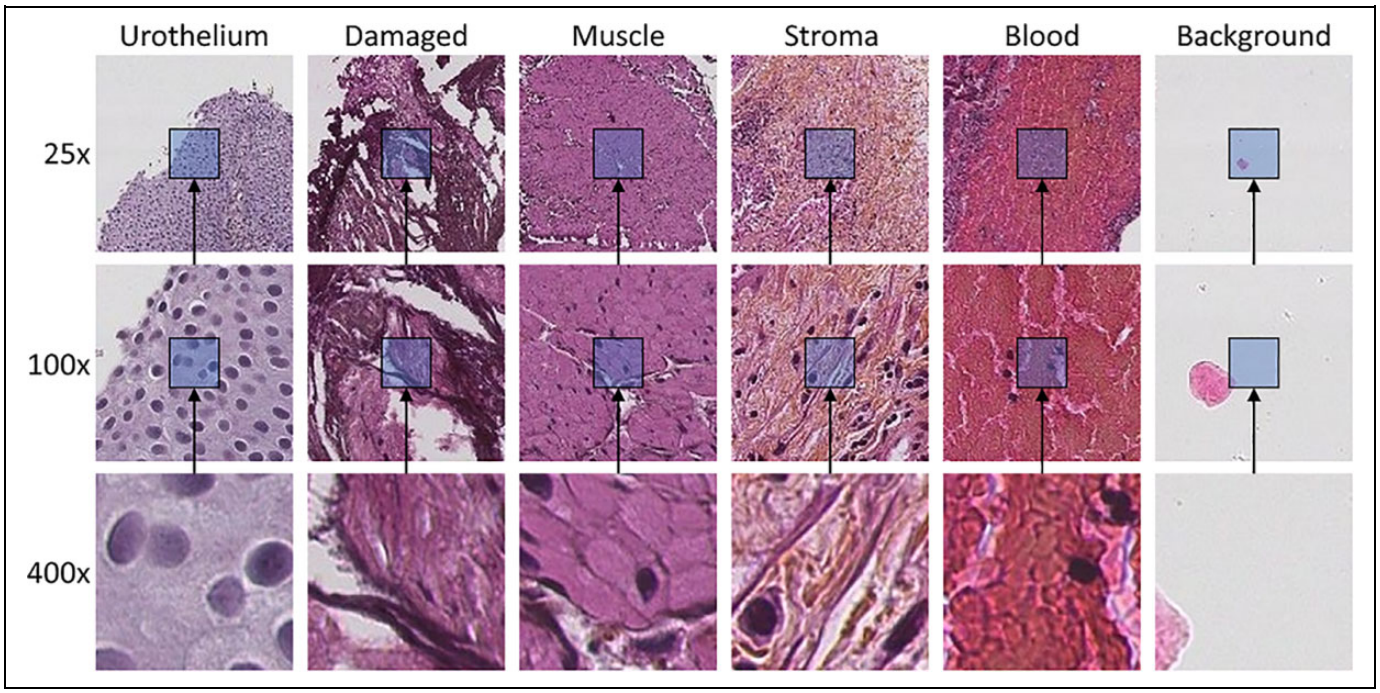

Figure 1. Example tiles of each class extracted at three magnification scales. Tiles at each scale are extracted from the same center pixel. The magnification scale is increased by a factor of 4 in each step, resulting in the tile covering 16 times as much area, even though they have the same size of 128 × 128 pixels.

experience recurrence, and 10% to 30% will advance to a higher disease stage.[4]

Treatment and follow up of urothelial carcinoma are primarily based upon histological grade and stage, evaluated manually by an expert pathologist studying the histological images of the tumor using the latest WHO16 classification system.[5] Correct grade and stage are essential to avoid over- or under-treatment, and thereby unnecessary suffering for the patient. For most pathology departments, evaluation of histological images is still performed through a microscope, a time-consuming process, not always reproducible.[6] Digital pathology has been introduced to improve diagnostic accuracy, and certain computer-aided diagnostic (CAD) tools are in use for other diseases. However, such tools are currently not in use for the assessment of urothelial carcinoma and could potentially be of great value to patients and clinicians.

Non-muscle invasive bladder cancer is usually treated with transurethral resection of the tumor. The removed tissue contains both atypical urothelium from the tumor as well as stroma, but can also contain smooth muscle from the bladder wall, normal urothelium from surrounding mucosa and blood. During the procedure, parts of the tissue can get damaged, for example in terms of heating damage induced by laser or electrically heated wire loop. Areas on the whole-slide images (WSI) with blood and damaged tissue will not be suitable for extracting diagnostic and prognostic information, and a pathologist will discard such regions on inspection. CAD systems processing WSI must be able to identify trustworthy interesting areas of resected tissue, but also identify damaged areas and regions that should be excluded from further analyses.

This paper proposes an automatic method for classifying WSI tiles from urothelial carcinoma cases into the following categories: urothelium, stroma, muscle, damaged tissue, blood, and background, utilizing different magnification scales. Examples from each class are shown in Figure 1. The output of such a system can be used as a guide for pathologists, providing a quick visualization of where the different tissue types can be found. To the best of the author's knowledge, a system for segmenting urothelial carcinoma WSIs into each tissue class does not exist. For determination of stage, pathologist wants to identify if muscle tissue is present or absent in the WSI and whether the tumor has infiltrated it. As muscle tissue is often sparse in the WSI, it can be time-consuming to get a full overview of its locations. However, with the help of segmented tissue images, it can be verified in a short amount of time. In the future, training data for a CAD system will be created by utilizing the best model developed through this paper by extracting diagnostic relevant features from the appropriate and relevant regions in the WSI. As this problem is not strictly dependent on classifying all six tissue classes, a binary approach is also experimented with in this paper classifying only urothelium vs. non-urothelium tissue to see if an increase in urothelium extraction can be achieved.

Tile-based classification of WSI has been done earlier.[7] However, by only classifying a single tile, it leaves out information from the surrounding area. Moreover, WSI viewed on different magnification scale identifies different information.

During an examination, a pathologist will integrate information across several magnification levels before reaching a final decision. Low magnification (25x) will show global context information such as papillary architecture, outline, and the border of the tissue, as well as color and texture. Nuclear polarity can be evaluated in the mid magnification (100x), while high magnification (400x) will reveal cytological features like cell size and shape, mitosis, as well as cell nucleus characteristics as contour, size and colorization intensity, and distribution.

The proposed method combines global context information found at lower magnifications (25x, 100x) with local information found at the highest magnification (400x) using deep neural networks to extract features from the different scales, thereafter concatenating the features feeding the last classifier layers of the network. Different neural network models were tested which utilized different combinations of the scales.

Related Work

It is not possible to feed an entire gigapixel WSI into a deep neural network, and a practical solution to this is to divide WSI into tiles and feed the tiles sequentially to the deep neural network. There are primarily two methods for semantic segmentation within medical applications. The first, which utilizes models capable of providing pixel-wise classifications, can output segmentations with high resolution. These networks are usually based on the fully convolutional networks (FCN) introduced by Long et al. in.[8] Popular models are the U-net model by Ronneberger,[9] and variants of this.[10,11] As these networks can detect small details, they are often used in cell and nuclei segmentation,[12,13] but also on tumor segmentation tasks.[14] The downside, however, is the need for pixel-wise ground-truth annotation for supervised learning, which is difficult and time-consuming to generate, especially in many medical applications. These networks are typically trained and tested on small example-patches from WSIs, since no dataset with a pixel-wise annotation of cells and tissue types on full WSI exist.

The second approach is based on tile-wise classification, where the models output a class label for each tile. This results in a coarser segmentation with the resolution of the tile size, and thus are more often seen for classification tasks rather than segmentation tasks. Nevertheless, it has been used in tumor segmentation methods.[15-19] As every pixel within the tile belongs to the same class, the tile-based ground-truth annotation process is significantly simplified for classification and localization of regions within histological images.

A combination of both tile-wise and pixel-wise classification has been seen for segmentation of WSI by Guo et al.[20] Firstly, a tile-based prediction using Inception-V3 gives a coarse segmentation of the WSI, followed by a pixel-wise classification of only the tumor tiles for refined segmentation of those areas. This approach can speed up the segmentation process relative to a pixel-wise segmentation of the entire slide; however, the need for pixel-wise ground-truth in all region of interests is still a significant challenge.

A pathologist studying a slide would typically zoom in and out, looking at both details and context. To similarly include these features in an artificial intelligence (AI) model, some multiscale approaches have been suggested. Models are trained with multiple input tiles, either taken from different magnification scales or taken from the same scale but with varying sizes to accommodate for a larger field of view. In the work of Sirinukunwattana et al.,[21] the author has performed a systematic comparison between five single-scale and five multiscale architectures, tested on four classes of prostate cancer and four classes of breast cancer. Both tiles extracted at different magnification levels, as well as tiles of various sizes, were tested; and the result supports the claim that incorporating a broader visual context improves the outcomes. Another multiscale approach was used by Vu et al.,[13] which created a network named multi-scale deep residual aggregation network (MDRAN). First, a tile is extracted from the WSI at 200x magnification, and then resized to x0.5 and x2 the original size. The three scales (0.5x, 1x, 2x) were then aggregated in the model and used to accurately segment nuclei of non-small cell lung cancer (NSCLC). Since the models uses multiple inputs, the architectures often become more complex, and the total number of parameters within the models also goes up. This affects both the training and inference time of the models.

Most previous work on WSI classification is targeted on segmenting cancerous vs. non-cancerous areas of the WSI, and often the non-cancerous class may include several tissue classes. E.g. the work just mentioned by Vu et al.[13] also performed WSI classification of NSCLC into three classes: NSCLC adeno (LUAD), NSCLC squamous cell (LUSC) and non-diagnostic (ND). The ND regions, in this case, consisted of fat, lymphocytes, blood vessels, red blood cells, normal stroma, cartilage, and necrosis without any attempt to separate these classes. Sometimes, however, there can be useful information in stroma, muscle, or other non-cancerous tissue types as well. There are some very few reported works on segmenting various tissue types. In,[22] Li et al. propose a model with dual inputs trained to segment WSI from the ICIAR2018 breast cancer dataset into normal, benign, situ, and invasive regions. Also, a transfer learning model with multiple inputs was explored by Wang et al.[23] to segment histological images of inflammatory bowel disease (IBD) into the four categories: muscle regions, messy regions, messy + muscle regions and background. Kather et al.[24] used a deep learning model to classify tiles from colorectal cancer into eight different classes of tissue: tumor epithelium, simple stroma, complex stroma, immune cell conglomerates, debris and mucus, mucosal glands, adipose tissue, and background.

Relatively little work is aimed at segmentation of bladder cancer WSIs. In the work of Xu et al.,[18] a method for predicting low or high tumor mutational burden (TMB) in bladder cancer patients was investigated. As a preprocessing step, a tile-wise tumor vs. non-tumor classifier was used to segment out the tumor regions from the surrounding tissue. An SVM classifier was then used to predict the patient's TMB state using extracted histological image features from the tumor regions.

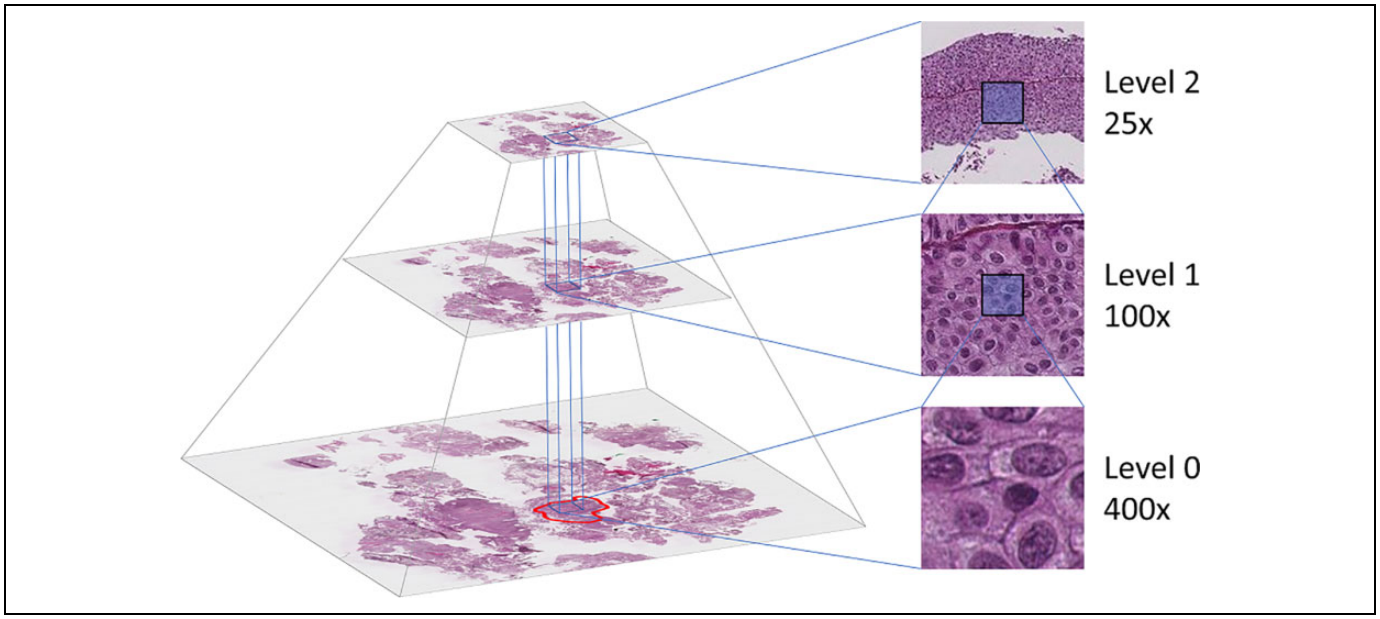

Figure 2. The WSI is stored in a pyramidal file format, including several down-sampled versions of the base image. The annotated region (marked with red at level 0) determines which tiles to extract. Tiles are then extracted at the desired location from all three levels.

A similar approach was used by Zhang et al.,[14] where a U-net like network was used to predict each pixel into tumor or non-tumor as a preprocessing step before using another neural network for predicting the slide level diagnosis. As urinary bladder tumors are removed using a laser, burnt and damaged tissue is often present at the WSI. Muscle, stroma, and blood will also be part of the removed tissue and visible in the WSIs. But no effort is aimed at identifying these regions, even though they may contain valuable information for a pathologist.

The recent research efforts show promising results utilizing deep neural networks in different configurations for classifying and localizing cancerous areas. However, most effort is made on the "big four" in cancer (i.e., breast, lung, prostate, and bowel), performed on some publicly available datasets. Still, there is relatively little work done on other cancer types, on multiclass classification, on tissue-type classification, and segmentation/heat maps of full WSI.

Aims and Contributions

In Wetteland et al.,[25] we presented a method based on convolutional neural networks (CNN) for classifying tiles of urothelial carcinoma WSI into the six classes shown in Figure 1. The model utilized the autoencoder architecture and was first pretrained on a large unlabeled dataset, and afterward fine-tuned on an annotated dataset. The models did not include any context, as both the unlabeled and labeled dataset was extracted at the full image resolution of 400x magnification.

The main contribution of the current paper is to combine histological images from different magnification scales into the model, giving the model access to a greater field of view and more context of the surrounding tissue. The resulting models are also used to generate segmented images of all the tissue classes within bladder cancer WSIs. An extensive number of experiments are conducted to find the best combination of inputs and magnification levels for the given task. The method utilizes the pyramidical image file format to extract tiles from existing down-sampled versions already present in the file, excluding any up- or down-sampling, limiting the number of necessary computational operations. Transfer learning is incorporated by building on the VGG16 network rather than the autoencoder model. To summarize, this paper proposes an automatic multiscale system, merging inputs of 25x, 100x, and 400x magnification, based on a CNN for classification of whole-slide histological images into six classes.

A preliminary study of this work was published by Wetteland et al. as an abstract.[26] Here we present much more comprehensive experimental work and a description of the method.

Materials and Methods

First, the data material will be introduced and explain how the datasets are prepared. Afterward, the proposed system for tissue segmentation is presented. Then the structure of the model is described, and finally, the training procedure and model selection is explained.

Data Material

The data material consists of digital whole-slide images from patients diagnosed with primary papillary urothelial carcinoma, collected at the University Hospital of Stavanger, Norway, in the period 2002-2011. The biopsies are formalin-fixed and paraffin-embedded, from which 4 μm slices are cut and stained with Hematoxylin Eosin Saffron (HES).

The prepared tissue samples are scanned at 400x magnification using the Leica SCN400 slide scanner, producing image files in Leica's SCN file format. The images are stored as a pyramidal tiled image with several down-sampled versions of the base image in the same file to accommodate for rapid zooming. Each level in the file is down-sampled by a factor of 4 from the previous level. Figure 2 shows an example of a pyramidal histological image with three levels. The Vips library[27] is capable of extracting the base image as well as the down-sampled versions, making it easy to extract the dataset at each resolution.

Two datasets were collected from the described data material, referred to as the CV dataset and the inference dataset, both are described below.

CV dataset. An expert pathologist carefully annotated selected regions in the WSI, where each region includes one of the six classes. A total of 239 regions belonging to the five foreground classes was annotated in WSI from 32 unique patients. The background regions were extracted from seven randomly selected patients.

The annotated regions contain tight corners and narrow passages to accommodate the shape of the tissue regions in the WSI. When extracting tiles from the WSI, a grid of non-overlapping tiles was superimposed upon the annotated region at 400x magnification level. The tiles in the grid which lie outside of the region are regarded as invalid and will not be used, whereas tiles within the region are valid. By shifting the grid in the X- and Y- direction, more or fewer tiles become valid. To maximize the number of valid tiles, an automatic search algorithm was developed. The algorithm checks the number of valid tiles for all possible positions of the grid. The grid location with the highest number of valid tiles was used to extract the dataset from that region. This search was performed individually for each region.

Tile sizes of 64×64, 128×128, and 256×256 pixels were tested when extracting tiles with the automatic program. Using a tile size of 64×64 extracted the most extensive dataset, but the size may be too small as each tile contain little context information. With a tile size of 256×256, the extracted dataset became very small, especially for the stroma and muscle class. A tile size of 128×128 was thus chosen as a trade-off between the other two sizes. When a tile is saved from the region, the corresponding tiles from 25x and 100x magnification were also extracted in such a manner that the center pixel is the same in all three magnification levels, as can be seen in the right-half of Figure 2.

The extracted 400x magnification tiles are ensured to stay within the region border. However, by keeping the tile size the same, the lower magnification (25x, 100x) tiles will have a wider field of view, allowing for more context of the surrounding tissue to be included. Consequently, these tiles will, in some cases, include several classes. Because the annotation process requires specific expertise input, the dataset contains a limited number of samples. Furthermore, the labels are imprecise as they do not include samples of the labeled border between tissue regions. This would require multi-label samples, an even more expensive annotation process. As a result of this, the dataset is weakly labeled in both quantity and quality.

Table 1. The Resulting CV Dataset Is Listed in the Table With the Total Number of Tiles Extracted for Each Class. The Number of Tiles Refers Only to Tiles Extracted at 400x Magnification. For the DI- and TRI-CNN Models, the Numbers Need to be Multiplied by 2 and 3, Respectively. Classes Marked With an Asterisk Shows the Number of Tiles After Augmentation.

Class	Tiles	Patients
Urothelium	29 728	28
Damaged	33 607	9
Stroma*	9 750	5
Blood	19 832	5
Muscle*	19 932	4
Background	27 012	7

No normalization of the stain color is performed on the data, and the raw pixel intensity is used to train the models.

Stroma- and muscle-tissue are more sparsely distributed in the WSI, resulting in a smaller amount of data for these classes. Data augmentation techniques have been utilized to balance the dataset. Tiles from these two classes are extracted with 50% overlap, and further rotated and flipped during training to achieve a more balanced dataset. The size of each class is listed in Table 1.

Due to the low number of patients in the dataset, a traditional train/validation/test split could potentially hurt both the training and evaluation of the models. Instead, stratified 5-fold cross-validation is used. This enables the usage of all WSIs in both training and testing of the models. Stratification is performed on the patient-level to ensure that tiles from the same patient are not present in both the training and test set. A random seed is set to ensure that the folds are the same for each model, making the included samples in the training and test sets identical for all models.

Inference dataset. In addition to the CV dataset, seven WSIs were selected to be used as inference on the retrained models. The WSIs included in the inference dataset is not part of the CV dataset, and thus unseen by the models. As with the CV dataset, no normalization is performed on the WSIs in the inference dataset.

Due to the large size of the histological images, the WSIs included in the inference dataset do not have any annotations, and therefore any quantitative measurements are lacking. However, the resulting segmented images have been examined by a pathologist to be promising and confirm that the models can go from predicting smaller regions of the WSI to segment the full WSI.

Proposed System

An overview of the proposed system for tissue segmentation of whole slide images is presented in Figure 3. The system accepts

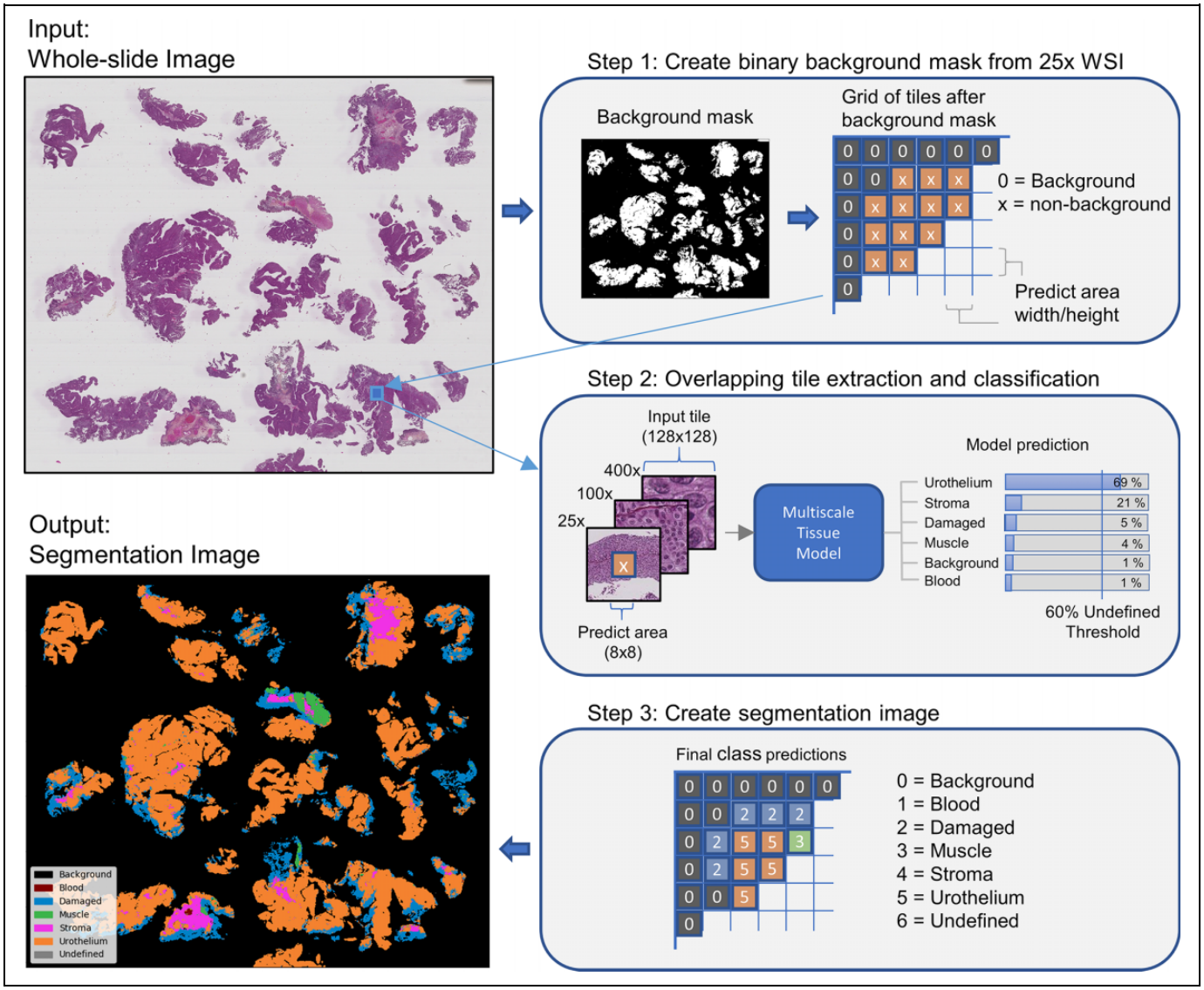

Figure 3. Overview of the proposed system. A background mask is created from the 25x WSI to exclude the background from further processing. Areas in the WSI selected as non-background is then extracted and fed through the multiscale model from Figure 4, which outputs tissue predictions. The prediction needs to exceed a set threshold to be valid. Finally, the segmentation image is generated by giving each class a separate color. The values shown in the figure are for illustration purposes only.

input WSI of any size and outputs a corresponding segmentation image from the input. The system is tested on the seven WSIs in the inference dataset. The system consists of three main steps which will be described here. The multiscale model in step 2 is described in more detail in the next section. Note that the blue box in step 2 in Figure 3 marked with "Multiscale Tissue Model" can be exchanged with any of the models described in the model structure section below.

First, a binary background mask is produced from the 25x level of the WSI, generated by checking the pixel intensity value and splitting them into background or non-background tiles. About 60 to 80% of the WSI is covered by background, so this step reduces the number of tiles that needs to be processed by the inference model. Tiles selected as non-background are then extracted and fed to the multiscale model for further classification.

Depending on which model architecture is used (MONO, DI, or TRI), one, two, or three tiles are extracted from the same location but with different magnification. The extracted tile will always be 128 × 128 pixels, as this is the required input size of the inference model. However, the prediction only holds for a smaller area within the tile, typically 8 × 8 pixels, but can be set to any size. The input tiles are then overlapped, such that the inner area is located next to each other with no overlap.

Tiles are classified according to the highest prediction score. The outcome of a prediction may be equally split between multiple classes (e.g., two classes getting a score of 0.5 each, or four classes getting 0.25 each). To avoid such cases, a threshold value is set to determine if a prediction is valid. To ensure that the majority of the predicted score falls to a single class, the threshold needs to be above 0.51. Also, by setting the

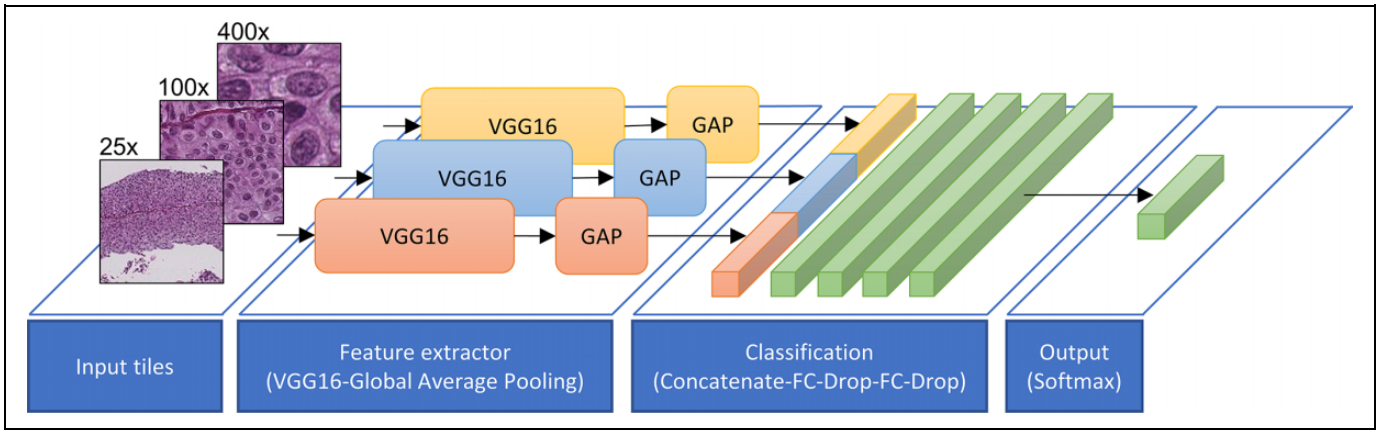

Figure 4. A block diagram of the TRI-CNN model proposed in the current paper. The input tiles are fed through individual pre-trained VGG16 network and global average pooling (GAP) layer to create feature vectors. The feature vectors are concatenated and fed through the classification network before entering the final output layer consisting of a softmax function. The softmax function outputs a prediction score for each of the six classes.

threshold too high may result in removing too many tiles. A threshold value of 0.6 is therefore determined as a trade-off between removing the unwanted conflicting predictions and not removing too much. Tiles with all prediction scores below the threshold are labeled as *undefined*.

Finally, each class is given a separate color, and the final segmentation image is saved. The segmentation images are ensured to only show classes with prediction scores higher than 0.6 but do not show the exact score. A method for creating heat maps has also been implemented, where no thresholding is performed, and the score for each class is visualized. A disadvantage of this is that one image must be created for each class. We earlier showed this approach in Wetteland et al.,[25] but have omitted it from this paper.

Multiscale model structure. This paper compares three architectures referred to as the MONO-, DI-, and TRI-CNN models. The three architectures have one, two, and three inputs, respectively. To differentiate the models from each other, they are named according to their main architecture, and the input scale, e.g. MONO-400x is a MONO-CNN model trained on tiles extracted at 400x magnification. Tiles in the dataset are extracted at three magnification levels, yielding three MONO models: MONO-25x, MONO-100x, and MONO-400x. These three magnification scales can further be combined in three configurations for the DI-CNN model: DI-25x-100x, DI-25x-400x, and DI-100x-400x. The TRI-CNN model has only one configuration: TRI-25x-100x-400x, and is depicted in Figure 4. The different MONO- and DI-CNN models can easily be derived from the same figure. E.g. to create the DI-25x-400x model, remove the 100x input and blue blocks, and to create the MONO-100x model, remove the 25x input, 400x input, red and yellow blocks.

The overall structure of each model is the same. Each input is fixed at $128 \times 128 \times 3$ pixels, which is the size of each tile. The input is fed into a pre-trained VGG16 network[28] which acts as a feature extractor, followed by a global average pooling (GAP) layer providing a feature vector representation of the input. This feature vector is then fed into a classification network consisting of two fully-connected (FC) layers, each followed by a dropout layer, and a final softmax layer with one output node for each class. The DI- and TRI-CNN models have two and three parallel VGG16 branches, respectively, resulting in multiple feature vectors. These feature vectors are concatenated before entering the classification network. The FC-layers has the same size of 4096 neurons as the original layers in the VGG16 network. Dropout layers are added after each FC-layers to add regularization to the network due to the small dataset.

Training procedure and model selection. All models were trained using the SGD optimizer with a learning rate of 1.5e-4, batch size of 128, a dropout rate of 0.3, and a cross-entropy loss function. Early stopping was enabled, stopping the model when no increase in performance during the past 10 epochs was seen. Due to the cross-validation training scheme, no validation set was used, and the early stopping process was thus monitoring the training loss. The model is written in Python 3.5 using the Keras machine learning library,[29] and Scikit-learn module[30] for evaluation.

The models were trained in a stratified 5-fold cross-validation fashion. To produce an unbiased evaluation score, the output from each fold was summarized in a micro-average manner, as suggested by Forman and Scholz.[31] All the true positive (TP), false positive (FP), and false negative (FN) values were summarized for each class over all the folds to produce a final micro-averaged F1-score.

The VGG16 network, which is used as a base model in our architectures, is pre-trained on the ImageNet dataset.[32] It is possible to have the base model fixed during training by freezing the parameters, preventing the base model from being updated. Freezing the parameters will allow for faster training as fewer parameters need to be learned, however, as the nature of the histological images is not part of the ImageNet domain, it could affect the model's ability to fully grasp the new images.

Table 2. Results for all 28 Models, Trained Using Stratified 5-Fold Cross-Validation. Each Score Is Shown as Micro-Averaged F1-Score Aggregated Across all Classes, Marked as "All" in the Table. F1-Score Only for the Urothelium Class Is Shown in the Columns Marked "Uro." Numbers in Bold Refer to the Highest Score in Their Respective Column.

		Multiclass				Binary-class			
		Frozen		Unfrozen		Frozen		Unfrozen	
	Model	All	Uro.	All	Uro.	All	Uro.	All	Uro.
Single-scale	MONO-25x	93.4	92.9	**96.4**	96.8	96.3	92.5	98.1	96.1
	MONO-100x	94.4	96.6	94.8	97.8	98.3	96.5	99.1	98.1
	MONO-400x	87.2	89.7	86.4	86.3	94.2	88.1	93.7	87.2
Multiscale	DI-25x-100x	**96.5**	97.4	96.2	98.1	98.1	96.2	**99.3**	**98.5**
	DI-25x-400x	95.6	96.3	96.0	97.6	97.8	95.4	98.3	96.5
	DI-100x-400x	95.0	96.8	95.3	97.6	98.4	96.6	98.9	97.7
	TRI-25x-100x-400x	**96.5**	**97.6**	**96.4**	**98.3**	**98.5**	**97.0**	99.2	98.3

By unfreezing the weights, it may allow to better adapt to the histological domain, at the cost of longer training time. Both freezing and unfreezing the weights were tested in the experiments.

As one of the objectives is to be able to automatically extract *urothelium tissue* from the histological images, to be used in diagnostic systems in the future, it is therefore not strictly necessary to classify all six tissue classes. A possible easier problem would be to define a binary problem, classifying urothelium vs. non-urothelium tissue. Each model was therefore also tested with this binary-class approach to see if it improved classification results for urothelium tissue. By simply combining the remaining five classes into one non-urothelium class, the dataset becomes heavily unbalanced toward the non-urothelium class. To counteract against this, augmentation using rotation and flipping was applied to balance out the dataset. By augmenting all the tiles from the muscle, stroma, and urothelium class 4x during training, the dataset became evenly distributed between the two classes urothelium and non-urothelium.

After evaluating the model using stratified cross-validation, a new and final inference model was trained by utilizing all available data as training data. The average number of epochs used during cross-validation was used when training the inference model. This inference model was then used to predict new WSIs from the inference dataset.

Results

This section will present the results for the different models. A total of 28 models were trained using stratified 5-fold cross-validation, including *single*- and *multiscale*, and *binary*- and *multiclass* models. Each model was trained using weakly labeled data, with both frozen and unfrozen weights in the VGG16 network.

Table 2 shows the cross-validation results for all the models. Aggregated micro-average F1-score across all classes are included, as well as the F1-score for only the urothelium class to better compare multiclass vs. binary-class models. Figure 5 displays the confusion matrices for the best multiclass models. The matrices are normalized to allow for more easy comparison. For the number of samples in each class, refer to Table 1.

Some of the best models have been retrained on the entire CV dataset and used to segment the seven WSIs included in the inference dataset. The resulting segmented images have then been inspected by an expert pathologist and are considered to be very promising. Figure 6 shows four WSIs and their corresponding tissue segmented images generated by the best multiclass model. Figure 7 shows a comparison between segmentation images generated by the best binary-class model and the best multiclass model. A DICE-score is calculated to measure the similarity between the predicted urothelium tissue between these two models, with an average DICE-score of 0.87 for the three WSIs. Figure 8 shows a close-up region taken from the top-right corner of the first WSI in Figure 6. This region is then segmented with all the best MONO-, DI-, and TRI-models for comparison.

Discussion

The results in Table 2 are shown as micro-averaged F1-score across all classes, as well as for the urothelium class. The results are overall good for all models, and a discussion of each case follows below. Afterward, the confusion matrices and the segmented images will be discussed, and finally, different usage scenarios of the system will be considered as well as some limitations of the study.

Binary-class vs. multiclass. As expected, the binary-class models achieve a higher average F1-score than the multiclass models, with all 14 of the binary models getting a higher score than their multiclass counterparts. This is expected because five of the classes are now grouped, and misclassification within these classes is canceled out. The best multiclass model is the frozen TRI-25x-100x-400x with an F1-score of 96.5% across six classes, whereas the best binary model is the DI-25x-100x with unfrozen weights, which got an F1-score of 99.3% across its two classes.

By looking at the F1-score for the urothelium class alone, the multiclass models are now superior, with 9 of the 14 models

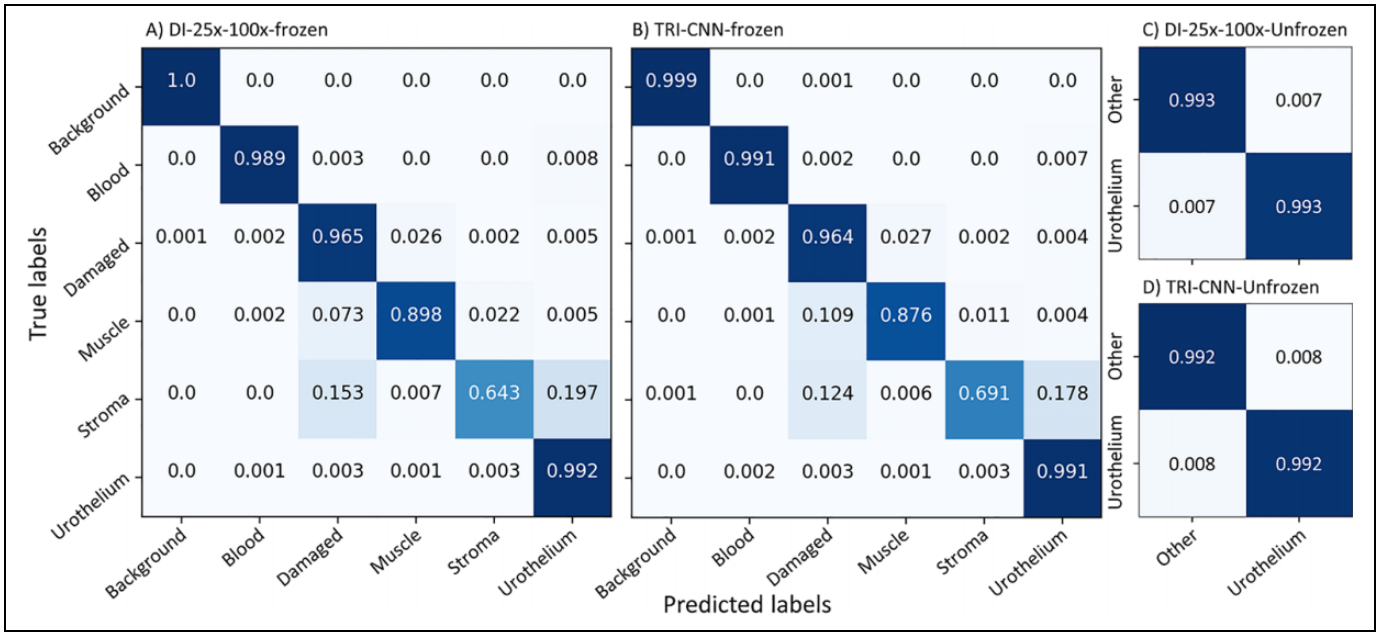

Figure 5. Normalized confusion matrices for the best multiscale models. Aggregated results across all 5 folds in the cross-validation test. A) Best multiclass DI-CNN, B) Best multiclass TRI-CNN, C) Best binary-class DI-CNN, and D) Best binary-class TRI-CNN.

being ahead of their binary-class counterparts. The few binary-models which have a higher score, are only marginally so, with the largest difference being the unfrozen MONO-400x, where the binary version is 0.9% better than the multiclass version. It is clear that by simplifying the problem into a 2-class problem, did not help with getting better urothelium extraction. The highest urothelium score is achieved by the TRI model, where both the unfrozen multiclass and unfrozen binary-class version each got an equal F1-score of 98.3% for the urothelium class.

Frozen vs. unfrozen. The three architectures MONO, DI and TRI, have 19 M, 21 M, and 23 M trainable parameters, respectively, when the VGG16 weights are frozen. By unfreezing the weights, the same models get 34 M, 50 M, and 67 M trainable parameters. When comparing results for these models, there is on average an increase of +0.6% by unfreezing the weights. Of the 14 unfrozen models, 10 get a higher score than the corresponding frozen models. The largest increase is seen in the binary MONO-25x model, which goes from an F1-score of 96.3% to 98.1% by unfreezing the weights.

The increase in the number of trainable parameters also affects the training time of the models. The average time per epoch for all the frozen models was 9 minutes, while the unfrozen models needed on average 10 minutes to compute one epoch. This is an increase of 11% processing time per epoch. However, the frozen models needed on average 162 epochs to reach the early stopping criteria, whereas the unfrozen models only needed 58 epochs. Thus, the models with unfrozen weights needed about 60% less processing time during training.

Single-scale vs. multiscale. When comparing the single-scale MONO-models with the multiscale DI- and TRI-models, the multiscale models achieve better results across all columns in Table 2, with the exception for the unfrozen MONO-25x model which matches the performance of the TRI-scale model. If we limit ourselves to the multiclass models, the best models for the three architectures are the unfrozen MONO-25x with 96.4%, frozen DI-25x-100x with 96.5%, and frozen TRI-25x-100x-400x which got an F1-score of 96.5%. The story is similar for the binary models, with unfrozen MONO-100x being the best with 99.1%, unfrozen DI-25x-100x with 99.3%, and unfrozen TRI-25x-100x-400x with 99.2%.

By looking at the single-scale models alone, it is clear that the two lower scales (25x, 100x) are performing better than the 400x scale, and that having a greater field of view is preferable. The multiscale models, consisting of two and three VGG16 networks, have a more complex structure involving more parameters than the MONO models. In addition, they have access to a greater field of view in all its models. These two features seem to help the performance of these models.

Naturally, the MONO models take the least amount of training time, with an average of 4:40 minutes per epoch. The DI-models take 136% longer with an average of 11:01 minutes, and finally, the TRI-models take the most time with 19:38 minutes on average per epoch. That is 321% and 78% longer than MONO and DI, respectively. The average number of epochs before reaching the early stopping criterion for the three architectures was 147, 88, and 64 epochs for the MONO-, DI-, and TRI-models, respectively.

Confusion matrices. Figure 5 shows the resulting normalized confusion matrices for the best multiscale models for both multiclass and binary-class models.

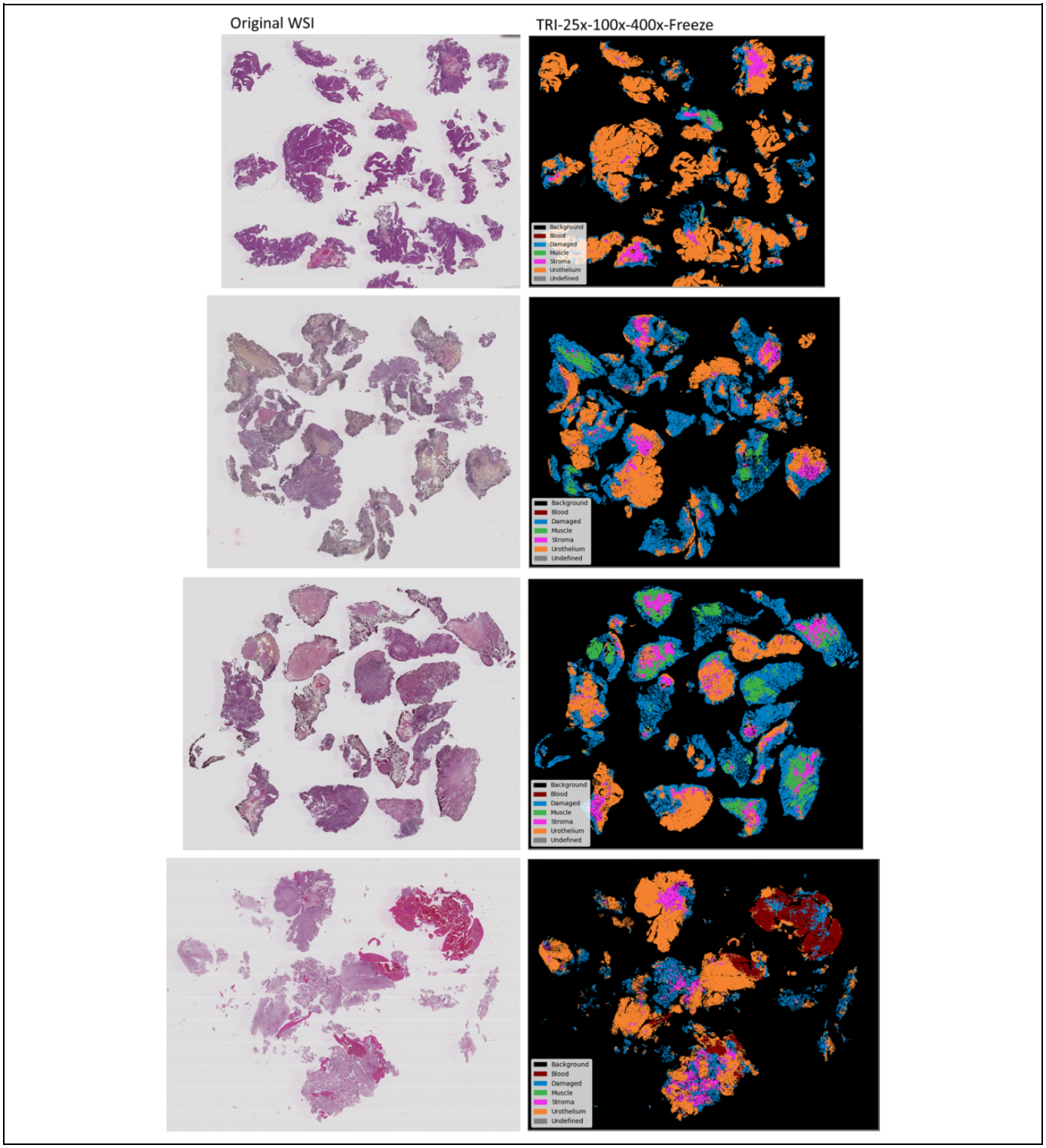

Figure 6. The best multiclass model was retrained and used to generate segmentation images from four WSI not present in the training data.

In the two multiclass matrices (A) and (B), the models did an excellent job at classifying background, blood, and urothelium correctly, and a great job with the damaged class as well. Both models struggled mostly with the muscle and stroma classes. These are the classes with the fewest number of labeled samples in the dataset. As a result of this, the models may have achieved a weaker generalization for these classes, and thus misclassified them more often. Most notable misclassifications are related to muscle and stroma being misclassified as damaged tissue, and also stroma being misclassified as urothelium.

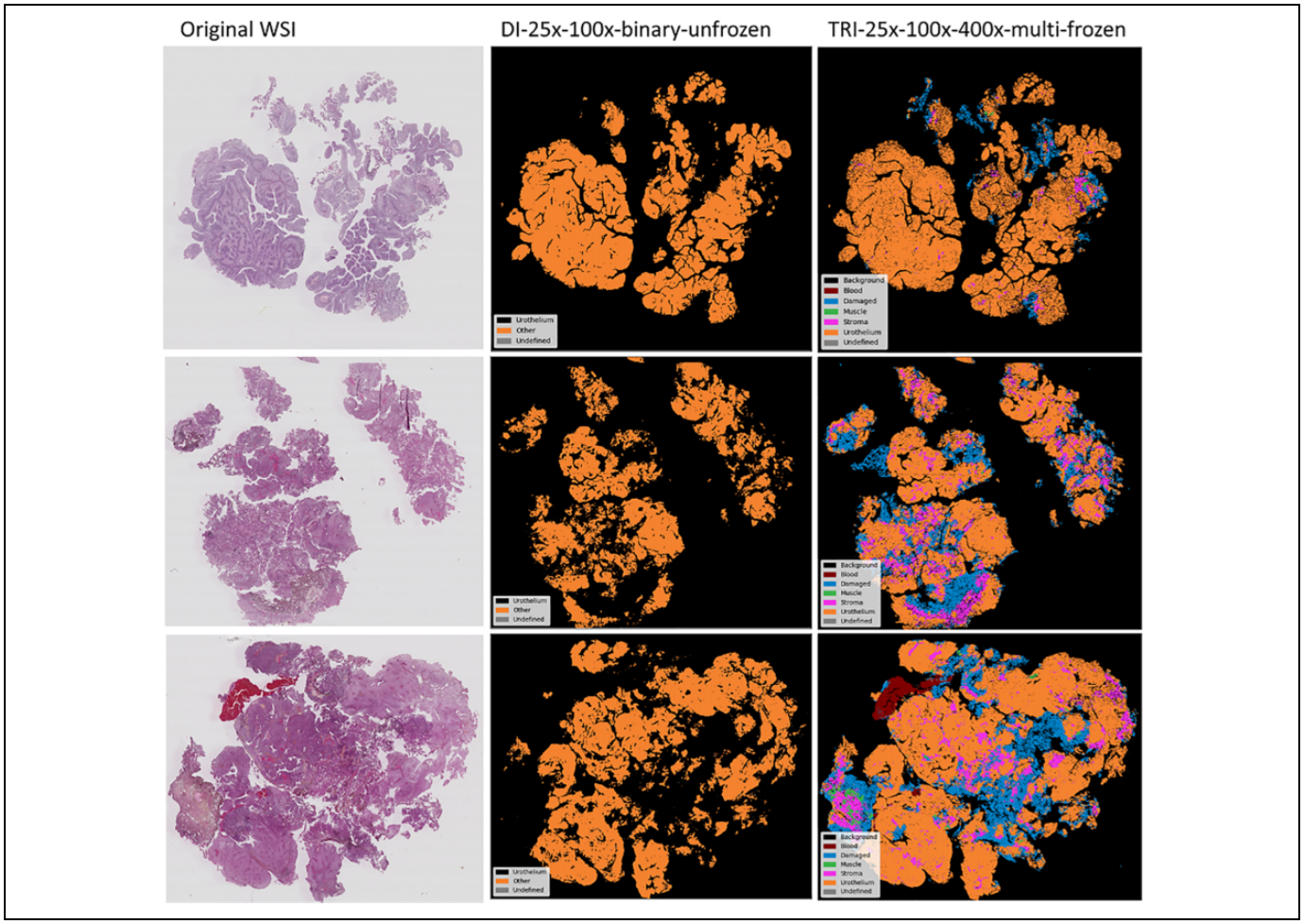

Figure 7. The best binary-class model vs. the best multiclass model. A DICE-score is calculated to measure the similarity between the predicted urothelium tissue between the two models. DICE-score from top to bottom are 0.92, 0.85 and 0.85 .

The two binary-class models in Figure 5 (C) and (D) got an equally good performance. Five of the classes are now combined into one class named *other* in the figure and thereby removing most of the misclassifications from the multiclass cases. However, this did not significantly increase the performance of model (C) and (D). Model (D) got the same normalized score as (A), and model (C) is only marginally better.

Inference dataset results. The seven WSIs included in the inference dataset were processed with overlapping tiles according to Figure 3, where only the inner 16 $\times$ 16 pixel of the tile was classified. The average processing time was 7 hours 18 minutes, including all three steps in Figure 3. On average, only 0.9% of the WSIs were categorized as undefined. Four of the WSIs are presented in Figure 6, and three in Figure 7.

Segmentation image results. The best multiclass model, according to Table 2, is split between two models. The frozen DI-25x-100x and frozen TRI-25x-100x-400x both have a similar F1-score of 96.5%, but the latter model has a higher urothelium F1-score and is thus regarded as the best multiclass model. The model was retrained and used to process four new WSIs, not present in the training data, to demonstrate its usage. Figure 6 shows the original WSI with the corresponding segmentation images. The segmented images are intuitive, easy to understand, and allow even untrained personnel to both identify and locate the difficult to find regions, e.g. like muscle tissue.

Fully multiclass-annotated WSI in our dataset is not available. The resulting segmentation images for the WSI have, however, been manually inspected by an expert uropathologist and are considered to be very promising, especially considering that the WSIs were only weakly annotated. Large homogeneous areas with a certain tissue type are clearly recognized. Most models are really challenged by smaller, more heterogeneous areas.

Binary-class vs. multiclass segmentation images. The best multiclass and binary-class models were retrained and used to create the segmentation images seen in Figure 7. The multiclass segmentation image may be of more interest to a pathologist, as it outlines regions of all six classes, whereas the binary-class segmentation image only outlines the urothelium class. However, both the multiclass and binary-class models have about the same F1-score for the urothelium class, and the additional

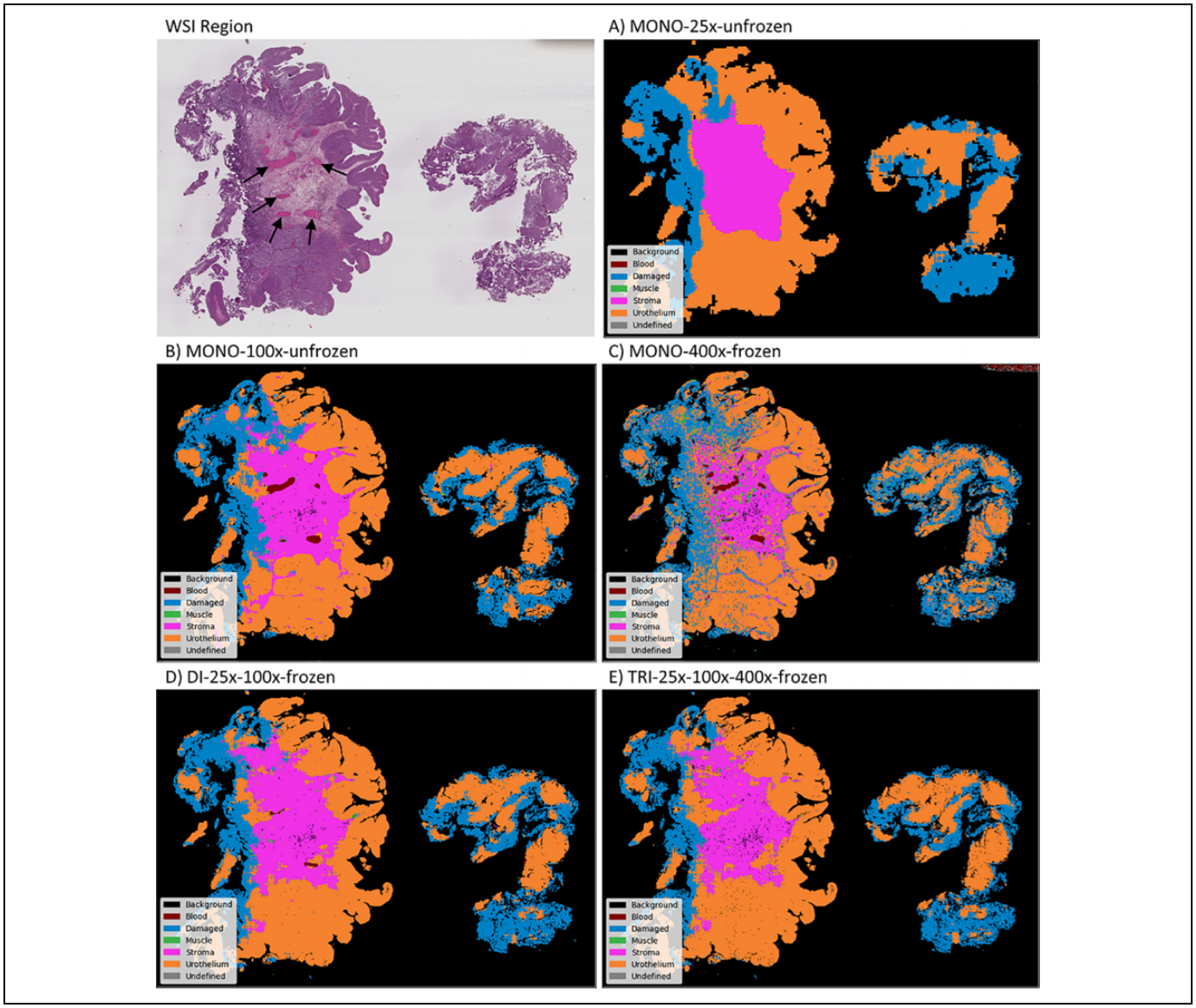

Figure 8. Segmentation of close-up region taken from the top-right corner from the first WSI in Figure 6. A) Best MONO-25x, B) Best MONO-100x, C) Best MONO-400x, D) Best DI-CNN model, E) Best TRI-CNN model. Arrows in the WSI region points to small areas of blood that the models struggle to identify.

information in the multiclass segmentation images favor the former model in a final system.

After comparing the urothelium regions in the two segmented images for each WSI, they are very similar. The DICE-score is calculated to measure the similarity between the regions, and the three cases have an average DICE-score of 0.87, which confirms that the two model's prediction for urothelium is quite similar. However, there is no truth annotation, so the DICE-score does not reveal if one of the models is better than the other.

Close-up segmentation regions. Even though the system is trained on weakly labeled data, consisting of single-class samples, using tile-based classification and not a per-pixel classification, it is still interesting to see how the system performs on a detailed level. This also allows us to compare the different models. Figure 8 shows a close-up region taken from the top-right corner from the first WSI in Figure 6, processed using an 8 $\times$ 8 pixel predict area.

All models do a decent job of outlining the major regions in the image. The different models process the image on different scales, and so the prediction tile covers a larger area for the smaller scales. The effect of this is visible at the three MONO models, where the level of detail goes up with each scale. The MONO-100x and MONO-400x models, with its smaller field of view, are able to detect some of the small regions containing blood in the middle of the image. The MONO-25x, however, is not able to identify this. The DI-25x-100x model, which has access to both the mid and broad field of view, barely identifies a small part of

the blood, whereas the TRI scale model does not identify it at all.

Usage Scenarios

As seen from both Table 2 and the segmented images in Figure 6, the model is fully capable of distinguishing between the different tissue types. The presented system has several possible usage scenarios, which will be discussed here.

The segmented images in Figure 6 can be used as a digital tool for pathologists to help them become more efficient in their work. It can be used to guide them to the diagnostic relevant areas of the WSI, such as urothelium, muscle, and stroma tissue. It can also be used to find edges of the urothelium tissue without damage more easily. During an examination, a pathologist needs to verify if muscle tissue is present or not in the current WSI. With the segmented images, this can be verified within a short amount of time.

Another use case for the system is as a preprocessing step for an automatic diagnostic system. For instance, each patient has follow up records about whether the patient experienced recurrence and progression. By training a diagnostic model on the entire WSI, the dataset quickly becomes too large if many patients are included. Also, by randomly selecting a subset of tiles within each WSI, the dataset will include a large portion of damaged tissue and blood, which will add noise to the diagnostic model. By using the multiscale tissue model presented in this paper as a preprocessing step, areas of clean, undamaged urothelium and other diagnostic relevant types can easily be extracted and used as training data.

Limitations

One limitation of the current study is that the dataset is relatively limited in size. A small training dataset may lead to overfitting of the model, resulting in poor performance, and a small test set may cause an optimistic estimate of the performance. Several measures have been taken to reduce these negative effects. Pre-trained models, dropout, and early stopping was used to reduce overfitting, and cross-validation was used to get a realistic estimate of each model's performance.

As mentioned in the data material section, the labels are accurate in the highest resolution (400x) but are imprecise on the lower scales (25x, 100x), meaning the ground-truth is based on weak annotations of the dataset, which may impact the accuracy. The experimental results show that having access to a greater field of view outweighs the potential negative effects of imprecise labels.

It is difficult to compare the presented models against other approaches or to perform a test on an independent dataset. To the best of the authors' knowledge, no other open dataset exists with annotations of the same six classes. As mentioned in the related work section, some research and models exist for segmentation of histological images. However, these are based on other cancer types or trained on other classes than the six classes used in this paper.

Conclusion

This paper investigates the effect of using multiple scales during tissue classification from WSI of urothelial carcinoma into six classes. The classification is performed on smaller tiles and can be useful for a coarse segmentation, or ROI-extraction, of WSI. Three main architectures are presented: MONO-, DI-, and TRI-CNN model, and a total of 28 different models were trained using weakly labeled data and evaluated in a stratified 5-fold cross-validation scheme.

The multiscale models achieved a better result than the MONO-CNN models. There was not a substantial increase in urothelium classification by using the binary-class models, neither by cross-validation or by inspection of the segmented images. The best multiclass model was used to generate intuitive and easy to understand segmented images from unseen WSIs, and after inspection by a pathologist is considered to be very promising.

The segmented regions shown in Figure 8 demonstrates the importance of including the highest magnification scale (400x) during tile-wise classification. The models which do not include this scale are not able to identify the smaller details within the WSI.

As the three MONO models pick up different levels of details, we will in the future experiment on employing them in a multiscale ensemble model by combining their outputs, instead of combining the different scales within the models, as the DI- and TRI-CNN models do. We also plan to use the model for automatic ROI-extraction of relevant tissue in the WSI to create training datasets for a diagnostic and prognostic classification model. By only extracting the diagnostic relevant areas of the WSIs, a dataset of much higher quality can be collected.

Authors' Note

Ethical approval from Regional Committees for Medical and Health Research Ethics (REC), Norway, ref.no.: 2011/1539, regulated in accordance to the Norwegian Health Research Act. As this is a retrospective study, Ethical approval was given without written consent from the patients. All insights in a patient's journal are monitored electronically, and all except the treating physician were required to state the reason why they needed to read that patient's journal. This log is always open for the patient to view. All patients were checked if any had registered themselves in the register for research reservation from the National Institute of Health (Registry of Withdrawal from Biological Research Consent, Norway).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partly funded by the University of Stavanger, Faculty of Science and Technology, Strategic PhD scholarship in health technology.

ORCID iD

Rune Wetteland https://orcid.org/0000-0002-9995-4204

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A.Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/caac.21492.
2. Antoni S, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F. Bladder cancer incidence and mortality: a global overview and recent trends. *Euro Urol*. 2017;71(1):96-108. doi:10.1016/j.eururo.2016.06.010
3. Eble JN, Sauter G, Epstein JI, Sesterhenn IA. *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs*. IARC Press; 2004;1-359.
4. Mangrud OM. *Identification of Patients with High and Low Risk of Progression of Urothelial Carcinoma of the Urinary Bladder Stage Ta and T1*. [PhD Thesis, Ph. D. Dissertation]. University of Bergen; 2014.
5. Babjuk M, Böhle A, Burger M, et al. EAU guidelines on non–muscle-invasive urothelial carcinoma of the bladder: update 2016. *Euro Urol*. 2017;71(3):447-461.
6. Mangrud OM, Waalen R, Gudlaugsson E, et al. Reproducibility and prognostic value of WHO1973 and WHO2004 grading systems in TaT1 urothelial carcinoma of the urinary bladder. *PLoS One* 2014;9(1):e83192. doi:10.1371/journal.pone.0083192
7. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using convolutional neural networks. *PLoS One* 2017;12(6):e0177544. doi:10.1371/journal.pone.0177544.
8. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on CVPR*; 2015:3431-3440.
9. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *MICCAI*. 2015;9351:234-241. doi:10.1007/978-3-319-24574-4_28
10. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: hybrid densely connected U Net for liver and tumor segmentation from CT volumes. *IEEE Tran Med Image*. 2018;37(12):2663-2674. doi:10.1109/TMI.2018.2845918
11. Li J, Sarma KV, Chung Ho K, Gertych A, Knudsen BS, Arnold CW. A multi-scale U-net for semantic segmentation of histological images from radical prostatectomies. *AMIA Annu Symp Proc*. 2018;2017:1140-1148.
12. Graham S, Vu QD, Raza SE, et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Ana*. 2019;58:101563. doi:10.1016/j.media.2019.101563
13. Vu QD, Graham S, Kurc T, et al. Methods for segmentation and classification of digital microscopy tissue images. *Front Bioengine Biotech*. 2019;7. doi:10.3389/fbioe.2019.00053
14. Zhang Z, Chen P, McGough M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat Mac Intel*. 2019;1(5):236-45. doi:10.1038/s42256-019-0052-1
15. Huang X, He H, Wei P, Zhang C, Zhang J, Chen J. Tumor tissue segmentation for histopathological images. In: *Proceedings of the ACM Multimedia Asia on ZZZ*; 2019:1-4. doi:10.1145/3338533.3372210
16. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Med*. 2019;25(8):1301-1309. doi:10.1038/s41591-019-0508-1
17. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the IEEE conference on CVPR*; 2016:2424-2433.
18. Xu H, Park S, Lee SH, Hwang TH. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. *Bio Rxiv*. 2019:554527. doi:10.1101/554527
19. Halicek M, Shahedi M, Little JV, et al. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. *Sci Rep*. 2019;9(1):1-1. doi:10.1038/s41598-019-50313-x.
20. Guo Z, Liu H, Ni H, et al. A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. *Sci Rep*. 2019;9(1):1-0. doi:10.1038/s41598-018-37492-9.
21. Sirinukunwattana K, Alham NK, Verrill C, Rittscher J. Improving whole slide segmentation through visual context-a systematic study. *MICCAI*. 2018;11071;192-200. doi:10.1007/978-3-030-00934-2_22
22. Li Z, Tao R, Wu Q, Li B. CA-Refine net: a dual input WSI image segmentation algorithm based on attention. *Ar Xiv*. 2019:arXiv:1907.06358.
23. Wang J, MacKenzie JD, Ramachandran R, Chen DZ. A deep learning approach for semantic segmentation in histology tissue images. *MICCAI*. 2016;9901:176-184. doi:10.1007/978-3-319-46723-8_21
24. Kather JN, Weis CA, Bianconi F, et al. Multi-class texture analysis in colorectal cancer histology. *Sci Rep*. 2016;6:27988. doi:10.1038/srep27988.
25. Wetteland R, Engan K, Eftestøl T, Kvikstad V, Janssen EAM. Multiclass tissue classification of whole-slide histological images using convolutional neural networks. *ICPRAM*. 2019;1:320-327. doi:10.5220/0007253603200327
26. Wetteland R, Engan K, Eftestøl T, Kvikstad V, Janssen EAM. Multiscale deep neural networks for multiclass tissue classification of histological whole-slide images. *MIDL Extended Abstract Track*. 2019;arXiv:1909.01178.

27. Martinez K, Cupitt J. VIPS-a highly tuned image processing software architecture. *IEEE ICIP*. 2005;2:574. doi:10.1109/ICIP.2005.1530120
28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *Ar Xiv preprint arXiv:14091556*; 2014.
29. Chollet F. The Keras Blog. Published 2015. Accessed September 24, 2020. https://keras.io
30. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *JMLR*. 2011;12(10):2825-2830.
31. Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Exp Newsletter*. 2010;12(1):49-57. doi:10.1145/1882471.1882479
32. Russakovsky O, Deng J, Su H, et al. Image net large scale visual recognition challenge. *IJCV*. 2015;115(3):211-252. doi:10.1007/s11263-015-0816-y

Appendices

I: Multiclass tissue classification of whole-slide histological images using convolutional neural networks.

II: Multiscale deep neural networks for multiclass tissue classification of histological whole-slide images.

III: A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides.

IV: Automatic diagnostic tool for predicting cancer grade in bladder cancer patients using deep learning.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 19, 2021, accepted August 3, 2021, date of publication August 13, 2021, date of current version August 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104724

Automatic Diagnostic Tool for Predicting Cancer Grade in Bladder Cancer Patients Using Deep Learning

RUNE WETTELAND[1], VEBJØRN KVIKSTAD[2,3], TRYGVE EFTESTØL[1], (Senior Member, IEEE), ERLEND TØSSEBRO[1], MELINDA LILLESAND[2], EMIEL A. M. JANSSEN[2,3], AND KJERSTI ENGAN[1], (Senior Member, IEEE)

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway
[2]Department of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway
[3]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway

Corresponding author: Rune Wetteland (rune.wetteland@uis.no)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Regional Committees for Medical and Health Research Ethics (REC) under Application No. 2011/1539, and performed in line with the Norwegian Health Research Act.

ABSTRACT The most common type of bladder cancer is urothelial carcinoma, which is among the cancer types with the highest recurrence rate and lifetime treatment cost per patient. Diagnosed patients are stratified into risk groups, mainly based on grade and stage. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists, potentially leading to under- or overtreatment of the patients. The economic burden, unnecessary patient suffering, and additional load on the health care system illustrate the importance of developing new tools to aid pathologists. We propose a pipeline, called $\text{TRI}_{\text{grade}}$, that will identify diagnostic relevant regions in the whole-slide image (WSI) and collectively predict the grade of the current WSI. The system consists of two main models, trained on weak slide-level grade labels. First, a WSI is segmented into the different tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background). Next, tiles are extracted from the diagnostic relevant urothelium tissue from three magnification levels (25x, 100x, and 400x) and processed sequentially by a convolutional neural network (CNN) based model. Ten models were trained for the slide-level grading experiment, where the best model achieved an F1-score of 0.90 on a test set consisting of 50 WSIs. The best model was further evaluated on a smaller segmentation test set, consisting of 14 WSIs where low- and high-grade regions were annotated by a pathologist. The $\text{TRI}_{\text{grade}}$ pipeline achieved an average F1-score of 0.91 for both the low-grade and high-grade classes.

INDEX TERMS Automated cancer grading, bladder cancer, convolutional neural networks, multiscale classification, urothelial carcinoma, weakly labeled data, whole-slide image.

I. INTRODUCTION

Bladder cancer is the 10$^{\text{th}}$ most commonly diagnosed cancer disease worldwide, with 573 278 new cases in 2020 [1]. The most common type of bladder cancer is urothelial carcinoma, in which men are overrepresented. It is among the cancer types with the highest recurrence rate, approximatly 50 to 70%, which makes it especially challenging [2]. It requires an intensive treatment and follow-up plan, which results in it being one of the cancer types with the highest lifetime treatment cost per patient [3], [4]. In the case of muscle-invasive bladder cancer (MIBC), where the cancer has invaded the muscle wall of the bladder, a cystectomy is often required. However, cancers that stay confined in the bladder mucosa are referred to as non-muscle-invasive bladder cancer (NMIBC) and are easier to treat.

In histopathological diagnostics, pathologists use grading and staging to describe the tumor. These parameters are used to stratify patients into risk groups and form a suitable treatment and follow-up plan. The grade of a tumor describes the differentiation state of the tumor cells under a microscope. Different cancers have different grading scales, but in general, if the cancer cells are similar to that of healthy non-cancerous cells, the grade will be low, and the cancer will have a lower likelihood of spreading. On the other hand, if the cells have a more abnormal appearance and are disorganized, the grade will be higher. In addition to the grade, tumor stage is also important and is determined by the size of the primary

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

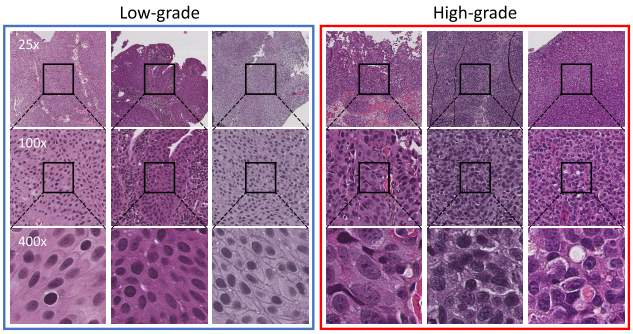

FIGURE 1. Examples of low-grade and high-grade tiles extracted from a WSI. The tiles are extracted from three magnification levels (25x, 100x, and 400x) and all have the same size of 256 × 256 pixels.

tumor, how far it has spread into the surrounding tissue, and the number of primary tumors present. In this paper, we focus on grading of NMIBC. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists [5], [6], which can lead to both under- or overtreatment of the patients. New tools to aid pathologists are therefore desired.

The World Health Organization (WHO) has proposed three grading systems for bladder cancer. The first grading system was introduced in 1973, referred to as WHO73, which is still somewhat used today. It consists of three categories, grade 1, grade 2, and grade 3, where grade 3 is the most severe state. A revised edition of the grading system was introduced in 2004 called WHO04, and further updated in 2016 as WHO16. In these versions, cases are split into low- and high-grade carcinoma. Some examples of low- and high-grade areas are shown in Fig. 1. Grade 1 patients are referred to as low-grade patients, and grade 3 patients are high-grade patients. Patients diagnosed as grade 2, however, are now split into either the low- or high-grade case. This might seem like a minor change, but for a patient to be diagnosed as low- or high-grade may result in very different follow-up regimes and local treatment with potential adverse events. A patient falsely diagnosed as a high-risk patient is an example of unnecessary patient suffering by overtreatment, additional load on the health care system, and extra cost. The data material used in this paper was collected and diagnosed prior to 2016 and will therefore focus on the WHO04 grading system.

After the tumor is removed, it is placed on an object glass and stained before a pathologist examines it. This is usually done through a microscope; however, with the introduction of digital pathology, digital versions of the stained specimen are also available in the form of whole-slide images (WSI). This has multiple advantages, such as remote access, storage and sharing cases between institutes, cloud computing, improved workflow, as well as computational pathology, which enables the use of new tools to process and interpret the tissue samples. All of which can improve the diagnostic accuracy and the clinical outcome of the patients [7]–[11].

Recent years have seen a rapid increase in both interest and usage of machine learning applications. Such tools could potentially be used to assist pathologists and help reduce the increasing workload. Also, because the errors made by a machine learning system may be different from that of a pathologist, the two may be combined for improved accuracy by the pathologist, as shown by Wang *et al.* [12]. Low reproducibility and variability in interpretations may also be reduced if a trustworthy computer-aided diagnosis (CAD) system could be implemented in a clinical setting.

With a CAD system, we want to map a WSI input to one of the disease output categories. The traditional machine-learning method to achieve this is by supervised learning. A set of known image and label pairs are shown to the model, which uses a gradient descent algorithm to optimize its parameters. For these algorithms to work efficiently and create robust models, a large set of image-label pairs are needed. Within digital pathology, we have access to a large amount of image data in the form of WSIs. However, annotated data is limited, challenging the practicability of supervised learning approaches. The nature of the images also calls for expert input to be able to annotate them. This is a time-consuming and, in some cases, challenging task. To create enough of the image-label pairs necessary to train these models and avoid the expensive annotation process, one possibility is to utilize data already available in the form of the slide-level diagnosis information. The WSIs are split into smaller images in the form of tiles, and the slide-level diagnosis will be assigned to each of the tiles.

For patients diagnosed with NMIBC, the tumor is usually removed through transurethral resection of bladder tumor (TURBT). During this process, parts of the tissue get damaged, either heating damage from the cauterization process or physical damage from tearing. Other tissue types, like stroma or muscle, as well as blood, are also often present in the slides of urothelial carcinoma. For the purpose of grading NMIBC, urothelium is the most diagnostic relevant tissue. For staging, both urothelium and stroma, and particularly the border between them, is essential. The presence of muscle tissue also has importance for correct staging. However, cauterized tissue from the TURBT process, as well as areas containing blood, have no diagnostic relevance. Feeding a deep learning model with these irrelevant tissue classes, e.g., blood or damaged tissue, may harm the diagnostic model's accuracy. To avoid this, we have previously proposed a method based on convolutional neural networks (CNN), which automatically segments NMIBC slides into background and five foreground classes (urothelium, stroma, muscle, blood, and damaged tissue). This tissue classification model is referred to as the $\text{TRI}_{\text{tissue}}$-model in the following and is explained in detail in Wetteland *et al.* [13].

In the current paper, we propose a system called $\text{TRI}_{\text{grade}}$ for automatically grading WSI according to the WHO04 grading system. The proposed system uses the $\text{TRI}_{\text{tissue}}$-model as a first-stage network for preprocessing the WSI to find regions of urothelium tissue. The extracted

urothelium tissue is then fed through a second-stage network called the $\text{TRI}_{\text{WHO04}}$-model for automatic grading.

The large size of the gigapixel images causes some challenges. It is not possible to feed the entire image into a deep learning algorithm; instead, tiles of a suitable size are extracted from the WSI and fed to the algorithm sequentially. The CNN-based model assigns a prediction score to every tile. These predictions are used to create a heatmap showing which regions were predicted with low- or high-grade carcinoma. The final decision can further be aggregated from the micro predictions into a slide-level prediction.

A WSI is stored in a pyramid format with multiple magnification levels, where the different levels will give different information. An example of such a pyramidal WSI is shown in Fig. 2. A pathologist will typically zoom in and out of a WSI to gather information at several scales before reaching a final decision. Our proposed method mimics this behavior by combining global context information and local details by utilizing a multiscale model architecture.

A. PREVIOUS WORK

With the introduction of digital pathology, there has been an increase in medical application research utilizing machine learning and deep learning approaches. Most research is related to cancer diseases such as breast, lung, prostate, brain, and skin cancer [14]. By looking at the list of US Food & Drugs Administration (FDA) approved artificial intelligence (AI) based medical technologies, most are in the fields of radiology, cardiology, and Internal Medicine/General Practice [15]. Still, a lot of effort is also aimed towards histological images [16]–[20].

The majority of CAD research conducted on histological images utilize two or more seperate models in their methods [16], [21]–[24]. First, a segmentation algorithm or region of interest (ROI) selection step is performed to narrow down the area which needs additional processing. This is an important step that helps in several ways. Compared to standard images, the WSIs are very large in size, and it is computationally expensive to process the entire WSI. By limiting the number of extracted tiles, the classification runtime is reduced, speeding up the classification step. Also, by removing the unwanted and diagnostically irrelevant areas, the extracted datasets will consist of higher quality tiles, which aids the classification algorithm in the following steps. After segmentation, tiles from the ROI are processed, usually by a classification model, which will predict the class of the tiles. Examples of tile classes can be cancer vs. non-cancer, recurrence vs. no recurrence, cancer grading or staging, or other classes related to cancer diagnosis. After all the selected tiles have been classified, the predictions are aggregated into a final slide-level prediction, usually using statistical or machine-learning methods.

Some research has been aimed towards urothelial carcinoma, otherwise known as bladder cancer. In Jansen *et al.* [22], they utilized two individual single-scale neural networks to detect and grade 328 cases of bladder cancer collected from 232 patients. A U-net-based segmentation network was trained to detect and segment the urothelium tissue, used as input to a second network trained to grade the urothelium tissue according to the WHO04 grading system. The classification network assessed the WHO04 grading on slide-level, using the majority vote of all classified tiles. The predictions were compared with the grading of three experienced pathologists. According to the consensus reading, the classification model achieved an accuracy score of 74%. The included whole-slide images were all exported at 20x magnification (0.5 μm per pixel).

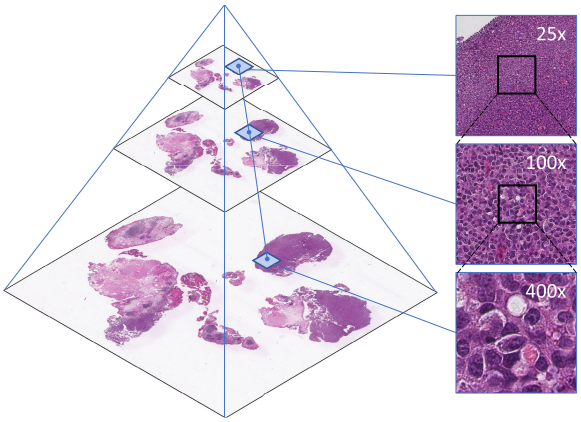

FIGURE 2. WSI images are stored in a pyramidal format, where the base image corresponds to the highest magnification level. The right-hand side shows a set of three tiles extracted so that the center of the tile corresponds to the same physical area in the WSI, forming a triplet.

From the same research group, the work of Lucas *et al.* [24] utilized the same urothelium segmentation model as presented in [22]. Regions of urothelium were then fed into a selection network which classified tiles into recurrence vs. no recurrence. A strategy was applied to select features from 200 tiles fed into a final bidirectional gated recurrent unit (GRU) classification network that predicts 1-year and 5-year recurrence-free survival (RFS) in bladder cancer patients.

The work of Zhang *et al.* [23] was also performed on bladder cancer. They used three different neural networks referred to as s-net, d-net, and a-net. The s-net model is a U-net-like architecture that classifies each pixel as tumor vs. non-tumor. The d-net then characterizes the tumor ROIs and generates an interpretable diagnosis and low-dimensional encodings. Finally, the a-net uses the ROI encodings and predicts a slide-level WHO04 grading.

Multiscale cancer subtype classification, where two or more different magnification scales are fed to the classification model, has been shown to improve the accuracy compared to single-scale models [13], [25]. This mimics the pathologist's process, which will zoom in and out to investigate the tissue area at several scales.

In Skrede *et al.* [21] the WSI is first segmented, before tiles are extracted at 10x and 40x resolution. The tiles from each scale are fed to an ensemble of 5 models, using a total of ten CNN-based models. The average score from the ensembles is used to predict the prognosis of colorectal patients.

TABLE 1. Overview of how the data material in this study is distributed into training, validation, and test sets. For triplets in the training dataset, see Table 2.

	Low-grade WSIs	High-grade WSIs	Total WSIs	Total triplets
Training	124	96	220	Table 2
Validation	17	13	30	301 775
Test	28	22	50	473 678

TABLE 2. Extracted triplets for the training dataset.

N	Total triplets before aug.	Total triplets after aug.	Percentage increase
250	54 564	55 000	0.8%
500	106 577	110 000	3.1%
1 000	202 904	219 560	7.6%
3 000	534 734	647 368	17.4%
5 000	812 588	1 051 752	22.7%

In the work of Hashimoto *et al.* [26] WSIs from malignant lymphoma were fed to a multiscale CNN-based model. They compared the results of models using tiles extracted at 10x or 20x resolution. However, the best result was achieved by combining the two scales into a multiscale model. The authors of this study also confirm that class-specific features exist at different magnification scales.

Previous work from our group, on bladder cancer, included tissue segmentation [13], [27], [28], and prediction of recurrence in NMIBC patients [29]. In Wetteland *et al.* [13], we experimented with three magnification scales and any combination of these. We proposed three MONO-models (Mono-25x, Mono-100x, and Mono-400x), three DI-models (DI-25x-100x, DI-25x-400x, and DI-100x-400x), and finally a model utilizing all three magnification scales, TRI-25x-100x-400x. All models used the VGG16 network as a feature extractor and were trained and evaluated on six tissue classes. The MONO-models performed worst, and the best result was achieved with the TRI-model utilizing all scales, supporting the claim that multiscale models achieve better results. Both frozen and unfrozen weights were experimented with, but the TRI-model trained with frozen weights in the VGG16 models performed best and achieved an average F1-score of 96.5% when evaluated on all six classes, and an average F1-score of 97.6% for the urothelium class alone.

Based on this result, we continued with the TRI-model and VGG16 as feature extractors in the current paper. We have not evaluated the MONO- or DI-models on the diagnostic data. The model referred to as TRI-25x-100x-400x in [13] is in the current paper referred to as the $\text{TRI}_{\text{tissue}}$-model. It is used for tissue extraction as shown in Fig. 4. The name, architecture, and base model have also been carried over to this paper and are the basis for the $\text{TRI}_{\text{WHO04}}$-model we propose here.

B. OUR CONTRIBUTIONS

The current study's main contributions is listed below.

- A novel, fully automated pipeline called $\text{TRI}_{\text{grade}}$ is proposed. The system consists of a tissue segmentation model and a diagnostic WHO04 grade model. The system's output consists of a tissue segmentation map, a WHO04 heatmap, and a predicted slide-level WHO04 grade. The proposed $\text{TRI}_{\text{grade}}$ system correctly predicted 45 of the 50 WSIs in the test set, achieving an accuracy of 90%.
- The $\text{TRI}_{\text{grade}}$ system-generated heatmaps are both visualized and evaluated against a segmentation test set. This helps to demonstrate the usage of such a system for a pathologist in a clinical setting.
- An algorithm for finding the optimal value of a decision threshold for classifying WSIs at slide-level is proposed.
- We trained models on differently sized training sets. The results for this provide insight on how dataset sizes affect the performance of the models, training time per epoch, and trained epochs before reaching stopping criteria during early stopping.
- Source code for this paper is accessible at the following URL address https://git.io/J3OdW.

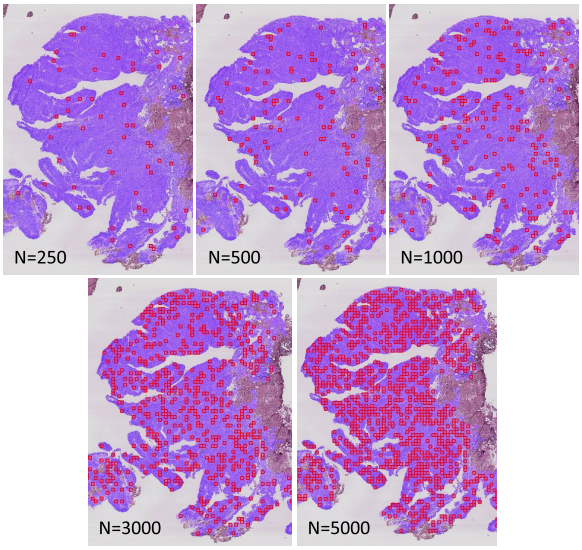

FIGURE 3. A close-up image from a WSI with a superimposed urothelium ROI mask (semi-purple). As *N* increases, the density of the tiles (red squares) also increases. The illustrated tiles are shown on 400x magnification level, but tiles from 25x and 100x are also extracted.

II. METHODS

The proposed $\text{TRI}_{\text{grade}}$ system presented in this paper utilizes multiscale models, which use tiles extracted at multiple magnification levels as input. For improved readability, we define these tiles as a *triplet*. A triplet is denoted T_i and is defined as a set of three tiles extracted from a WSI at three different magnification levels (25x, 100x, and 400x). Let $\mathcal{T}$ denote a set of triplets in a WSI, where $\mathcal{T} = \{T_1, T_2 \dots T_i \dots T_{max}\}$, and the number of elements in the set is given by the cardinality $|\mathcal{T}|$. An example of a triplet is shown in Fig. 2.

A. DATA MATERIAL

The data material consists of 300 digital whole-slide images from patients diagnosed with NMIBC, where the tissue is

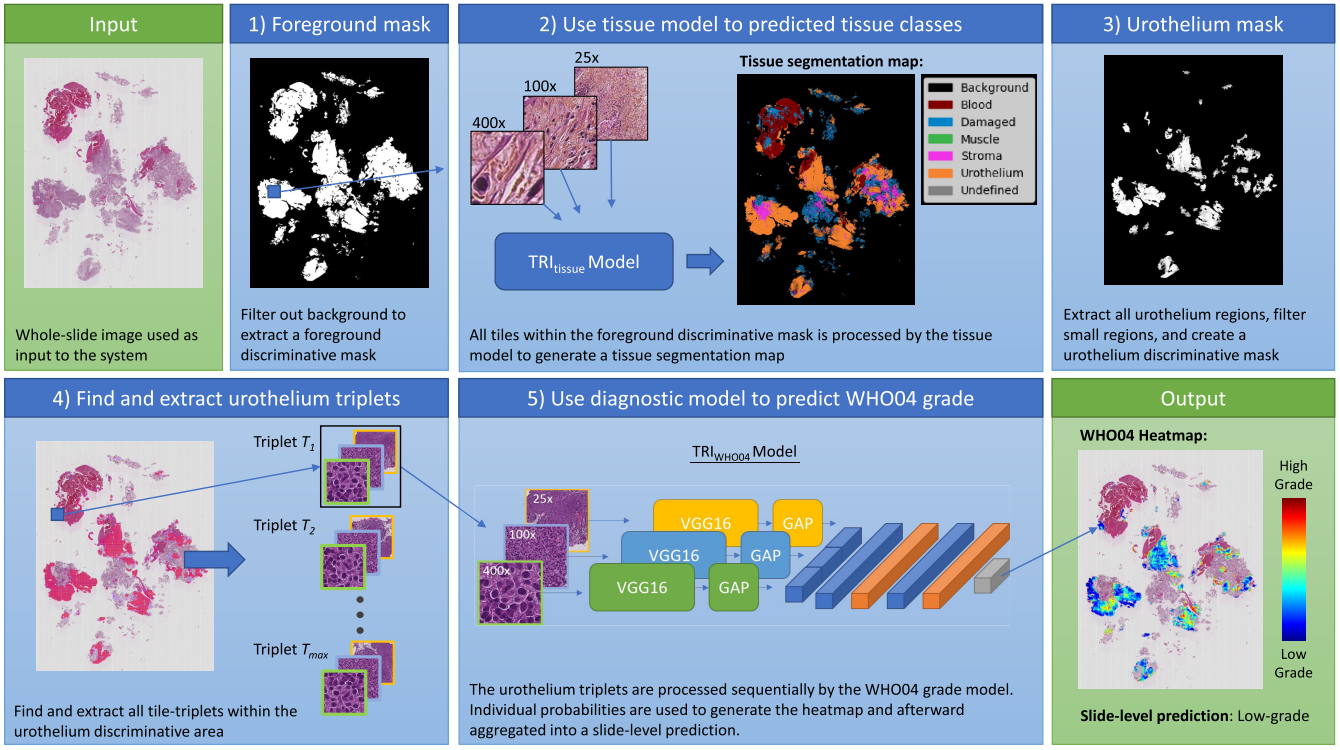

FIGURE 4. **This figure presents the pipeline for our proposed system, TRI_{grade}. Input) A WSI of urothelial carcinoma is used as input. 1) A foreground discriminative mask is found by evaluating the pixel intensity values and used as a reference to extract tiles from the WSI. 2) The TRI_{tissue}-model is used to generate a tissue segmentation map. 3) The urothelium regions are used to create a urothelium discriminative mask. 4) Using the urothelium mask, triplets consisting of tiles from three magnification levels are extracted from the input WSI. 5) The urothelium triplets are fed sequentially to the TRI_{WHO04}-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. Output) The system will output a WHO04 grade heatmap and a slide-level WHO04 prediction.**

removed from the patient through transurethral resection of bladder tumor. The data was collected at the University Hospital of Stavanger, Norway, in the period 2002-2011. All non-muscle invasive bladder cancers are included in the dataset, making it a true population based dataset. The biopsies were formalin-fixed and paraffin-embedded, from which 4 μm thick sections were cut and stained with Hematoxylin, Eosin, and Saffron (HES).

The slides were diagnosed and graded according to WHO73 and WHO04 [30]. All slides have the label low-grade or high-grade in the WHO04 system. In addition, cancer stage and follow-up data on recurrence and disease progression are recorded, and all patients have stage Ta or T1, i.e., non-muscle invasive. We have, however, no annotated regions with healthy non-cancerous urothelium available. All WSI have gone through a manual quality check at the department of pathology, Stavanger University Hospital, and only high-quality slides, with little or no blur, have been included in the dataset. However, as mentioned, NMIBC is removed by cauterization, which will leave burned and damaged tissue areas. All WSI are from the same laboratory, and the variation in staining color is relatively low. Ethical approval from Regional Committees for Medical and Health Research Ethics (REC), Norway, ref.no.: 2011/1539, regulated according to the Norwegian Health Research Act.

The glass slides were digitized using a Leica SCN400 slide scanner, producing WSI images in the vendor-specific scn file format. These WSI images are gigapixel images with a typical resolution of 100 000 × 100 000 pixels, stored as a pyramidal tiled image with several down-sampled versions of the base image in the same file to accommodate for rapid panning and zooming. The pyramidal structure of the WSI is depicted in Fig. 2. The Vips library [31] can extract the base image and the down-sampled versions, making it easy to extract the dataset at each resolution.

Tiles are extracted from the image pyramid at levels corresponding to 25x, 100x and 400x magnification, which is equivalent to a spatial resolution of 4 μm/pixel, 1 μm/pixel and 0.25 μm/pixel, respectively. For the TRI_{tissue}-model, we used a tile size of 128 × 128 pixels, which for the three magnification levels correspond to (512 μm × 512 μm), (128 μm × 128 μm), and (32 μm × 32 μm). For the TRI_{WHO04}-model, we had access to a much larger library of WSIs, and thus a larger tile size of 256 × 256 pixels was chosen. For the three magnification levels, this corresponds to (1 024 μm × 1 024 μm), (256 μm × 256 μm), and (64 μm × 64 μm).

The 300 WSIs included in this study were split into 220/30/50 WSIs for training, validation, and testing, respectively. Demographic characteristics of the data material were not used when splitting the data material into the different

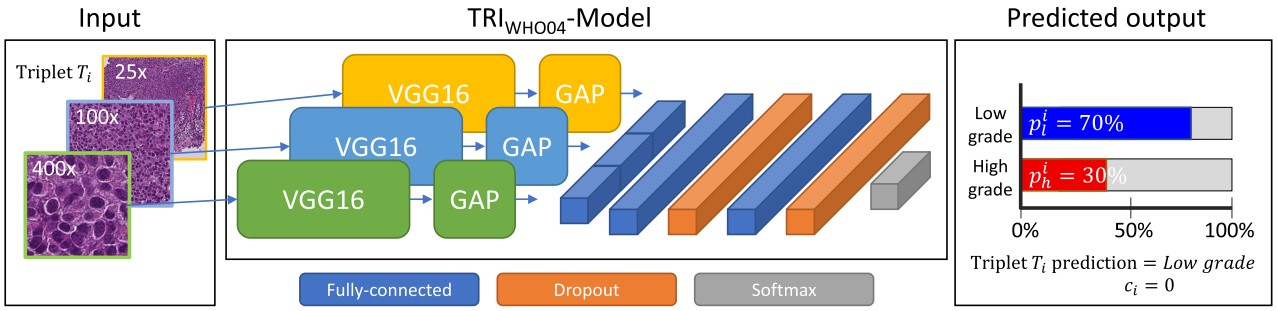

FIGURE 5. Architecture of the TRI$_{WHO04}$-model. Three separate VGG16 networks are used to extract features from each magnification scale. The global average pooling layer (GAP) is used to flatten the features into feature vectors, which are concatenated. The classification network consists of fully-connected layers and dropout layers. The output uses a softmax activation function to predict the input tiles to the two classes, low-grade and high-grade carcinoma.

datasets. Instead, the WSIs were randomly selected and stratified to include the same ratio of all diagnostic outcomes based on the WHO73 and WHO04 grading, stage, recurrence, and disease progression, to represent the data material best. The distribution of low- and high-grade WSIs in each dataset, as well as the number of triplets in the validation and test set, can be seen in Table 1.

The 50 WSIs in the test set will use the slide-level label as ground truth to evaluate the TRI$_{WHO04}$-model. In addition, a pathologist has carefully annotated low- and high-grade regions in 14 of the 50 WSIs. The 14 WSIs are a sub-set of the test set and are referred to as the *segmentation test set* and will be used to evaluate the low- and high-grade segmentation performance of the best TRI$_{WHO04}$-model.

From the 220 WSIs used for training, five datasets were extracted with a different number of triplets extracted from each WSI. A set of N triplets was selected randomly from the predicted urothelium regions in each WSI, where N was set to 250, 500, 1 000, 3 000, and 5 000.

Some of the WSIs in the data material contain only small amounts of urothelium, either because the tissue sample itself is small or because most of the tissue sample consists of damaged tissue or other tissue classes. For these WSIs, an augmentation strategy was employed, where a randomly selected set of triplets were augmented. The aim of this process is for each WSI to contribute equally, or as close as possible, to the number of triplets specified by N. Augmentation was performed by rotation and vertical/horizontal mirroring of the individual tiles in the triplet. All tiles in the triplet were augmented in the same manner. By combining rotation and mirroring, a tile can be oriented in eight uniquely defined ways, making this the maximum number a particular tile can be augmented. For $N \geq 1\,000$, some WSIs did not reach the desired number of triplets, even with 8x augmentation. No augmentation was performed on the validation or test datasets. Table 2 shows a list of total triplets extracted, before and after augmentation, for each value of N.

Fig. 3 shows a region from one WSI with the extracted tiles superimposed. The semi-transparent purple color indicates the predicted urothelium region. From this region, N randomly selected tiles are extracted as indicated by the red tiles on the image. As N increase, the density of extracted tiles also increases. Also, note that only the tile extracted at magnification level 400x is visualized in the figure. At each tile position, tiles from all three magnification levels (25x, 100x, and 400x) are extracted in such a manner that the center position of each tile corresponds to the same physical location, as illustrated in Fig. 2.

For preprocessing, all pixel intensity values were normalized from 0-255 values into 0-1 values, and the order of the color channels was altered from RGB to BGR. These steps ensure that the input data is presented to the VGG16 network in the same fashion as when it was pre-trained on the ImageNet data. No stain normalization was performed on the extracted tiles.

Our data material contains slide-level diagnostic information; however, no location annotations exist, showing where in the WSI the low- or high-grade regions are found, except on our segmentation test set, as explained. As manual annotation is time-consuming, expensive, and requires expert input, it is not feasible to get this type of detailed annotations on large datasets as needed for training such models, particularly considering both the size of each WSI and the total number of WSIs in the data material. Instead, each extracted tile inherits the slide-level WHO04 grade as its label. This is not ideal, as high-grade slides may contain regions with low-grade tissue. Consequently, all the extracted datasets are thus regarded as weakly labeled due to the inaccurate labels, which is consistent with what is called a weak label in many tasks [32]. The segmentation test set is considered strongly labeled.

B. PROPOSED SYSTEM

We propose a pipeline, called TRI$_{grade}$, that takes a WSI as input and outputs a tissue segmentation map, a WHO04 grading heatmap, and a slide-level WHO04 grade prediction. The pipeline consists of two main models, denoted as TRI$_{tissue}$-model and TRI$_{WHO04}$-model. The task of the TRI$_{tissue}$-model is to classify an input triplet as a tissue type which then can be used to make a tissue segmentation map. The task of the TRI$_{WHO04}$-model is predicting the cancer grade, i.e., low- or high-grade, based on the urothelium tissue. The TRI$_{grade}$ pipeline is depicted in Fig. 4 and explained in detail below.

Algorithm 1 Find Optimal Threshold Value D_t

Initialize: $\mathcal{Y}$, $\hat{\mathcal{Y}}$, $\mathcal{R}$, $\mathcal{D}_{c_{best}}$ are empty lists
Initialize: $Acc_{max} = 0$
for *WSI* $\leftarrow$ training set **do**
  Feed *WSI* through pipeline in Fig. 4
  $R_{high} = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|} c_i$
  Append R_{high} to the list $\mathcal{R}$
  Append the true grade Y of *WSI* to the list $\mathcal{Y}$
end for
for $D_c \leftarrow 0$ to 50 **do**
  for $R_{high} \leftarrow \mathcal{R}$ **do**
    $\hat{Y} = \begin{cases} \text{High-grade}, & \text{if } R_{high} \geq D_c \\ \text{Low-grade}, & \text{otherwise} \end{cases}$
    Append the slide-level prediction $\hat{Y}$ to the list $\hat{\mathcal{Y}}$
  end for
  $Acc_{D_c} = \text{sklearn.metrics.accuracy\_score}(\mathcal{Y}, \hat{\mathcal{Y}})$
  if $Acc_{D_c} > Acc_{max}$ **then**
    $Acc_{max} \leftarrow Acc_{D_c}$
    Clear list $\mathcal{D}_{c_{best}}$
  end if
  if $Acc_{D_c} \geq Acc_{max}$ **then**
    Append D_c to list $\mathcal{D}_{c_{best}}$
  end if
end for
$D_t = \lceil \frac{1}{|\mathcal{D}_{c_{best}}|} \sum \mathcal{D}_{c_{best}} \rceil$

1) $\text{TRI}_{\text{grade}}$ PIPELINE

The $\text{TRI}_{\text{grade}}$ pipeline depicted in Fig. 4 contains five steps explained here. The input to the pipeline consists of a WSI file in the vendor-specific.scn file format. First, in step 1, a foreground discriminative mask is found on the 400x level by evaluating the pixel intensity values as grey background or not. Using the foreground mask as reference, tiles with dimension 128 $\times$ 128 pixels were extracted from the WSI with 87.5% overlap, resulting in the inner 16 $\times$ 16 pixels being classified for each tile. Three tiles were extracted in the WSI (25x, 100x, and 400x) for each location, forming a triplet. All tiles in each triplet have the same dimension of 128 $\times$ 128 pixels and are extracted such as the center point corresponds to the same physical location in the WSI for all three tiles, as shown in Fig. 2.

In step 2, triplets are sequentially fed into the $\text{TRI}_{\text{tissue}}$-model we proposed in Wetteland *et al.* [13]. This model will evaluate the triplets and predict which of the six tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background) the current triplet belongs. In our case, the class of damaged tissue is a collection of all tissue that is not one of the other classes, and in our dataset, this is mainly cauterized or torn tissue areas. If blurred regions are a problem in the dataset, this can be made as a separate class or included in the damaged tissue class. After predicting all triplets, a segmented tissue map is created, visualizing all tissue regions in the WSI. This tissue map can also be presented to the clinician to help guide them more efficiently to the specific tissue types in the WSI.

From the generated tissue map, all urothelium regions are extracted in step 3. Small regions are filtered to suppress noise, and a urothelium discriminative mask is created on the 400x level. In step 4, a grid of non-overlapping tiles is overlayed on the WSI at the 400x level, this time using tiles of dimension 256 $\times$ 256 pixels. The individual tiles in the grid are checked against the discrimination mask. If 80% or more of a tile lay within the discriminate mask, the position is saved, while the remaining tiles are discarded. For the validation and test sets, triplets from all the saved positions are extracted. Whereas for the training set, N randomly selected triplets are extracted from the saved positions, where training sets are formed with N set to 250, 500, 1 000, 3 000, and 5 000. If fewer than N positions are saved, the augmentation strategy explained in the data material section is employed. The total number of extracted triplets for each dataset is shown in Tables 1 and 2.

A comprehensive description of how triplets are extracted from the WSI is given in Wetteland *et al.* [33], where a parameterized method for extracting tiles in multilevel images is given. The parameters used in this paper are the tile size parameter $L_T = 256$. The overlap-ratio between a tile and the discriminative mask is set to 80%, which corresponds to a value of $\phi = 0.8$. Tiles are checked at the 400x level by setting $\alpha = 0$, and the corresponding tiles in the triplets are found at level 25x and 100x, i.e., $\mathcal{S}_\beta = \{1, 2\}$. The binary mask B^k is set as the urothelium discriminative mask, and the starting coordinate of the grid is at position (0, 0). With these parameters and the methods described in [33], extraction of the triplets in the WSIs is repeatable and reproducible.

In step 5, the extracted urothelium triplets are fed to the $\text{TRI}_{\text{WHO04}}$-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. Finally, all scores are used to generate a heatmap which is overlayed on the WSI, and the aggregated micro-predictions are measured against the decision threshold D_t to get the final slide-level prediction.

2) MODEL ARCHITECTURE

The proposed pipeline in Fig. 4 contains two CNN-based models used for different tasks; the $\text{TRI}_{\text{tissue}}$-model is used for tissue classification and the $\text{TRI}_{\text{WHO04}}$-model for grading of urothelium tissue. The models are built upon the same architecture but have different inputs and outputs. The architecture consists of three separate VGG16 networks, one for each input scale. Both the model architecture and the TRI-terminology comes from our previous work on the tissue model in Wetteland *et al.* [13].

The input to the $\text{TRI}_{\text{tissue}}$-model is a triplet consisting of three 128 $\times$ 128 pixel tiles (25x, 100x, and 400x). The model can predict triplets extracted from anywhere in the WSI, but a foreground discriminative mask is usually used to save processing time by removing the background. The output of the $\text{TRI}_{\text{tissue}}$-model is a probability distribution over the six predicted classes (urothelium, stroma, muscle, blood, damaged tissue, and background). The input to the $\text{TRI}_{\text{WHO04}}$-model

is a triplet consisting of three 256 × 256 pixel tiles (25x, 100x, and 400x) extracted from urothelium tissue regions. The model outputs a probability distribution over the two predicted classes, low- and high-grade carcinoma. A block diagram of the $\text{TRI}_{\text{WHO04}}$-model architecture is depicted in Fig. 5. The $\text{TRI}_{\text{tissue}}$-model has almost the same architecture but has six output classes instead of two.

The individual tiles in the input triplet are fed to separate VGG16 networks. The VGG16 networks are used as base models with weights pre-trained on the ImageNet dataset, a large dataset containing annotated photographs used for computer vision research. Each VGG16 network acts as a feature extractor and takes a high dimensional tile as input (128 × 128 × 3 or 256 × 256 × 3 pixels) and compresses it down to a feature volume (8 × 8 × 512). A global average pooling (GAP) layer is used as the output layer for each VGG16 network, transforming the feature volume into a feature vector of length 512. The three feature vectors, one for each scale, are concatenated into one final feature vector of length 1 536 and fed to the classification network.

The classification network consists of two fully-connected (FC) layers using a rectified linear unit (ReLU) activation function, each followed by a dropout layer for regularisation. Lastly, an output layer with a softmax activation function is used to provide the prediction of the model. The two FC-layers and the two dropout layers each have a dimension of 4 096 neurons, and the output layer has one output neuron for each class. The $\text{TRI}_{\text{WHO04}}$-model consists of 67M parameters, where 23M of the parameters are trainable parameters belonging to the classification network.

3) TILE-LEVEL PREDICTION

When a triplet T_i is fed to the $\text{TRI}_{\text{WHO04}}$-model, the model outputs a list of probabilities for the two classes, low- and high-grade. These probabilities are denoted as $[p_l^i, p_h^i]$. To find the class with the largest predicted probability, the argmax function is used.

$$c_i = argmax([p_l^i, p_h^i]) \tag{1}$$

where c_i is the index to the predicted class for the triplet at position T_i. The low-grade class has an index of 0, and the high-grade class has an index of 1.

The proposed system can also produce a heatmap from the individual triplet probabilities, which indicates the location of low- and high-grade regions. This is useful for pathologists who can focus their limited per-patient investigation time on the diagnostic relevant areas in the WSI. A color mapping function converts the high-grade probability p_h^i into a color based on its value. This color is then superimposed on the WSI at the current triplet's position, covering the same area as the 400x magnification tile in the triplet. This results in the heatmap, as seen in the bottom-right of Fig. 4. Only the model's probabilistic score for the high-grade class is used to generate the heatmaps. However, because there are only two classes, a low probabilistic score of the high-grade class implicitly means a high score for the low-grade class. I.e., red highlighted regions in the heatmaps are associated with the high-grade class, and blue highlights indicate the low-grade class.

4) SLIDE-LEVEL PREDICTION

In addition to predicting the individual triplets, we also output a WHO04 slide-level prediction. A pathologist will often assign the worst case to a slide during a clinical examination, meaning that if a high-grade region exists in the WSI, the WHO04 grading should be high-grade. However, we must assume some misclassification in the WSI from both the $\text{TRI}_{\text{tissue}}$-model and $\text{TRI}_{\text{WHO04}}$-model, so there must be a minimum amount of high-grade triplets before the slide-level prediction becomes high-grade, and we would like to find a decision threshold, D_t, which maximizes correct prediction of the WSIs.

By summing over c_i, the number of triplets predicted as high-grade is counted, since triplets predicted as low-grade is at index 0 and thus not adding to the sum. By dividing by the total number of triplets in the WSI, we get the ratio of high-grade triplets referred to as R_{high} in this paper:

$$R_{high} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} c_i \tag{2}$$

If R_{high} exceeds the decision threshold D_t, the slide is given the slide-level prediction of high-grade; else, it is considered low-grade.

$$\hat{Y} = \begin{cases} \text{High-grade}, & \text{if } R_{high} \geq D_t \\ \text{Low-grade}, & \text{otherwise} \end{cases} \tag{3}$$

Algorithm 1 describes how to find the optimal threshold value D_t. Y is considered the ground truth grading of a slide and consists of a single value, whereas $\mathcal{Y}$ is a list of all the ground truth values. The same holds for $\hat{Y}$ and $\hat{\mathcal{Y}}$, which holds a single slide-level prediction and a list of all the predictions, respectively. First, all WSIs are processed, and the ratio R_{high} for each WSI is appended to the list $\mathcal{R}$. The true grade Y of each WSI is also saved in the list $\mathcal{Y}$. All WSIs in the dataset are processed before proceeding to the next step. A set of candidate threshold values, D_c, between 0-50% are tested one at a time. For each candidate threshold, the slide-level prediction $\hat{Y}$ for all WSIs is saved to the list $\hat{\mathcal{Y}}$. The total accuracy score is then calculated for the dataset. The decision threshold D_t is chosen as the candidate threshold with the highest score, or, if more than one D_c value yielded the same maximum result, the average integer value is selected as the decision threshold D_t.

5) TRAINING PARAMETERS

The $\text{TRI}_{\text{WHO04}}$-model was trained using a stochastic gradient descent (SGD) optimizer with a learning rate of 1×10^{-3}, learning rate decay of 1×10^{-6}, and momentum set to 0.9. The batch size used during training was set to 128. Both dropout layers had a dropout rate of 0.5. The cross-entropy loss function was used to optimize the model during training.

TABLE 3. Slide-level prediction results for automatic WHO04 grading tested on the 50 WSIs of the test set. Precision, recall, and F1-score is the weighted average score for the two classes across all 50 WSIs in the test set. D_t is the decision threshold found using Algorithm 1. The column trained epochs show how many epochs each model was trained before the early stopping criteria were reached. Training times are shown as hours:minutes.

Model	Trained epochs	Time per epoch	Training time	Precision	Recall	F1-Score	D_t
TRI_{WHO04}-250	23	1:22	31:39	0.86	0.84	0.84	49
TRI_{WHO04}-250-AUG	33	1:19	43:53	0.89	0.86	0.85	47
TRI_{WHO04}-500	21	1:42	35:59	0.89	0.86	0.85	43
TRI_{WHO04}-500-AUG	21	1:44	36:39	0.77	0.76	0.76	49
TRI_{WHO04}-1000	15	1:52	28:03	0.83	0.82	0.82	49
TRI_{WHO04}-1000-AUG	18	1:55	34:45	0.80	0.80	0.80	49
TRI_{WHO04}-3000	15	3:24	51:01	0.89	0.86	0.85	49
TRI_{WHO04}-3000-AUG	12	3:29	41:54	0.78	0.78	0.78	49
TRI_{WHO04}-5000	16	4:10	66:42	0.85	0.84	0.84	48
TRI_{WHO04}-5000-AUG	17	5:18	90:20	**0.92**	**0.90**	**0.90**	48

The pre-trained weights of the VGG16 networks were held frozen during training. To avoid overfitting the models on the training set, an early-stopping rule monitored the validation loss and stopped the training when no improvements were seen for ten epochs. The best epoch was restored when testing the models on the test set.

To train the models, a program was written in Python 3.6 using Keras 2.2.4 together with the Tensorflow 1.14 as backend [34], [35]. The PyVips 2.1 library was used for handling the WSI [31], and Scikit-learn 0.19 for evaluation [36]. The models were training on a Ubuntu 18.04 server, running on dual Xeon E5-2650 v5 @ 2.2GHz with a total of 48 cores. An Nvidia Tesla P100 16GB GPU was used for the training. Training parameters for the TRI_{tissue}-model can be found in Wetteland *et al.* [13].

III. EXPERIMENTS

We have conducted two experiments, listed here.

Experiment 1: is for slide-level prediction of WHO04 grade and is tested on the test set of 50 WSIs. As training of the TRI_{WHO04}-model is very time-consuming, we wanted to see if it is preferable to utilize more of the available urothelium data from each WSI as training data at the cost of additional training time or if a smaller dataset could perform equally well. This is interesting, both for our research group as well as other researchers working with large WSI datasets. If the optimal number of tiles used from each WSI during training can be lowered, then time can be saved in future experiments. To investigate this, we created several datasets where we extracted N triplets per WSI, as shown in Table 2. In this experiment, ten versions of the TRI_{WHO04}-model, all with the same architecture, were trained on training sets of various sizes, listed in Table 2. The micro predictions from the individual triplets were aggregated into a slide-level prediction of the WHO04 grading. A decision threshold D_t was found for each model using Algorithm 1; then, equation 3 was used to provide the final predicted grade.

Experiment 2: is testing the tile-level prediction and compare that in detail with the 14 WSIs of the segmentation test set. This set contains pathologist annotated regions belonging to either low- or high-grade which are considered the ground truth. The best model from experiment 1 is used for this, and the model's performance will be visualized as heatmaps. Calculation of recall and F1-score will be presented for each WSI, in addition to a total score across all WSIs.

IV. RESULTS

In experiment 1, slide-level test results for the ten models are listed in Table 3, showing trained epochs, time, precision, recall, F1-score, and the threshold value D_t. For precision, recall, and F1-score, the weighted average score is presented as reported by the *classification report* function from the scikit-learn library [36].

For experiment 2, the TRI_{WHO04}-5000-AUG model was used, as it performed best in experiment 1. The predicted heatmaps for each WSI in the segmentation test set are shown in Fig. 6 together with the ground truth. Recall, and F1-score for each WSI is listed in Table 4. As each ground truth WSIs only contain annotations for one of the two classes, the precision score will always be 1.00 because whenever the model predicts the ground truth class, it will be correct. The precision column in Table 4 is thus discarded. The last row in Table 4 shows the average value of all scores for each class together with the standard deviation. Table 5 shows the total aggregated results for all 14 WSIs. Here, the predictions for all WSIs are accumulated before the score is calculated.

A slide-level comparison between the proposed TRI_{grade} system and the model presented in Jansen *et al.* [22] is shown in Table 6. The TRI_{grade} system consists of the TRI_{tissue}-model followed by the TRI_{WHO04}-5000-AUG model. Values for sensitivity, specificity, and accuracy are shown for easier comparison with the reported results from [22]. These values are unweighted and calculated using values for true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Note that these results are based on models trained and evaluated on different datasets.

V. DISCUSSION

The three VGG16 networks are identical copies as we have used frozen (pre-trained) weights in this work. Thus, it would be possible to use only one copy of the model, with the appropriate change in the architecture, keeping in mind that

TABLE 4. Tile-level prediction for each individual WSI in the segmentation test set, using the TRI_{WHO04}-5000-AUG model. The WSI numbering is referring to the WSIs in Fig. 6. The last row shows the average value and standard deviation for its respective column.

	Low-grade		High-grade	
WSI	Recall	F1-score	Recall	F1-Score
WSI A	-	-	0.79	0.88
WSI B	-	-	0.90	0.95
WSI C	-	-	0.86	0.92
WSI D	0.87	0.93	-	-
WSI E	-	-	0.94	0.97
WSI F	0.83	0.91	-	-
WSI G	0.86	0.92	-	-
WSI H	-	-	0.90	0.95
WSI I	0.85	0.92	-	-
WSI J	0.79	0.88	-	-
WSI K	-	-	0.92	0.96
WSI L	0.92	0.96	-	-
WSI M	0.68	0.81	-	-
WSI N	-	-	0.58	0.73
Average	0.83 ± 0.07	0.91 ± 0.04	0.84 ± 0.12	0.91 ± 0.08

the feature vectors from the different magnifications are concatenated before the classification network. However, utilizing three versions of the VGG16 network allows us to train the entire multiscale model end-to-end and allows unfreezing the weights if a larger training set is available. We have experimented with unfreezing weights, but we quickly get overfitting problems with the available data material, this is therefore omitted from the paper.

Experiment 1 was conducted using ten training sets with a different number of triplets extracted from the same 220 WSI. From the result in Table 3, we see that the best performing model is trained on the largest dataset. However, the other models are not far behind. Even with a small value of N, the models do a good job at correctly predicting the WHO04 grade of WSIs.

Regarding overfitting, we tried training the models using unfrozen weights in the VGG16 networks, but this led to instantaneous overfitting of the model and had no improvements on the validation set. However, by freezing the weights, we see that all models improve on the validation dataset before reaching a plateau and eventually triggering the early stopping trigger. E.g., as shown in Fig. 7, the best model, TRI_{WHO04}-5000-AUG, improved its performance for seven epochs before training stopped after epoch 17. The weights from epoch seven were restored when using the model on the test sets. The number of trained epochs before the early stopping criteria is triggered decreases as the training dataset increases. This can be explained by the models trained on the larger datasets having more parameter updates per epoch than that of the smaller dataset models, thus reaching the plateau faster. Similarly, we see that the duration of one epoch is increasing as the dataset size increases. There is about a 60-hour difference in the smallest and largest model by comparing the total training time. Even though we would advise utilizing the most data to train a production model, it could be helpful to do an extended hyperparameter search and train multiple models on a smaller dataset.

TABLE 5. Aggregated tile-level result for all WSIs in the segmentation test set using the TRI_{WHO04}-5000-AUG model.

	Precision	Recall	F1-score
Low-grade	0.83	0.79	0.81
High-grade	0.90	0.81	0.85
Weighted Average	0.87	0.80	0.83

TABLE 6. Comparison table for automatic slide-level grading between our proposed method and the method presented in Jansen *et al.* [22]. Note that these results are based on models trained and evaluated on different datasets.

Model	Sensitivity	Specificity	Accuracy
TRI_{grade}	0.85	1.00	0.90
Jansen *et al.* [22]	0.71	0.76	0.74

Experiment 2, tile-level prediction, was conducted using the TRI_{WHO04}-5000-AUG model, which had a slide-level F1-score of 0.90. As seen in Fig. 6, Table 4 and 5, the results are overall excellent. The model does a very good job at correctly identifying both the low-grade and high-grade regions in the different WSIs. Table 4 shows that the model achieved an average F1-score of 91% for both the low-grade and high-grade classes. The aggregated score for all WSIs in Table 5 shows a small decrease in performance, with an F1-score of 81% and 85% for the two classes, respectively.

The largest misclassification in Fig. 6 is one of the regions in WSI-N, where the ground truth is high-grade, but the model predicts low-grade. When reevaluated by the pathologist, the misclassified area was found to be heterogenous, showing mixed low- and high-grade features, consequently regarded as high-grade initially. This illustrates one of the challenges with automatic grading of urothelial carcinoma, that grading between low- and high-grade is not two distinct binary classes but rather a continuous spectrum with a floating transition, making it difficult to set a hard threshold between the two.

To correct such misclassifications, and also avoid the costly task of annotating a large dataset, one possible solution is human-assisted learning. For example, the proposed TRI_{grade} system could be used to find and predict urothelium regions into the low-grade and high-grade classes, e.g., like the regions seen in Fig. 6. Then, a pathologist could verify the regions in each WSI and correct misclassified regions. This way, a large, strongly labeled dataset could be created, and the TRI_{WHO04}-model could be fine-tuned on the new dataset.

A direct comparison of results with others reported in the literature is not straightforward, as the experiments performed in this paper are conducted on a private dataset, which is often the case in many medical applications. To our knowledge, there exists no publically available NMIBC dataset or any publically available models from other researchers that we can evaluate on our dataset. The work of Jansen *et al.* [22] is based on the same labels but evaluated on a private dataset using different methods. Unfortunately, their models are not available for us to evaluate, and we do not have access to labels to train a Unet segmentation model from scratch, hence we cannot test the same approach by training the models ourselves. However, even though the dataset or model used in

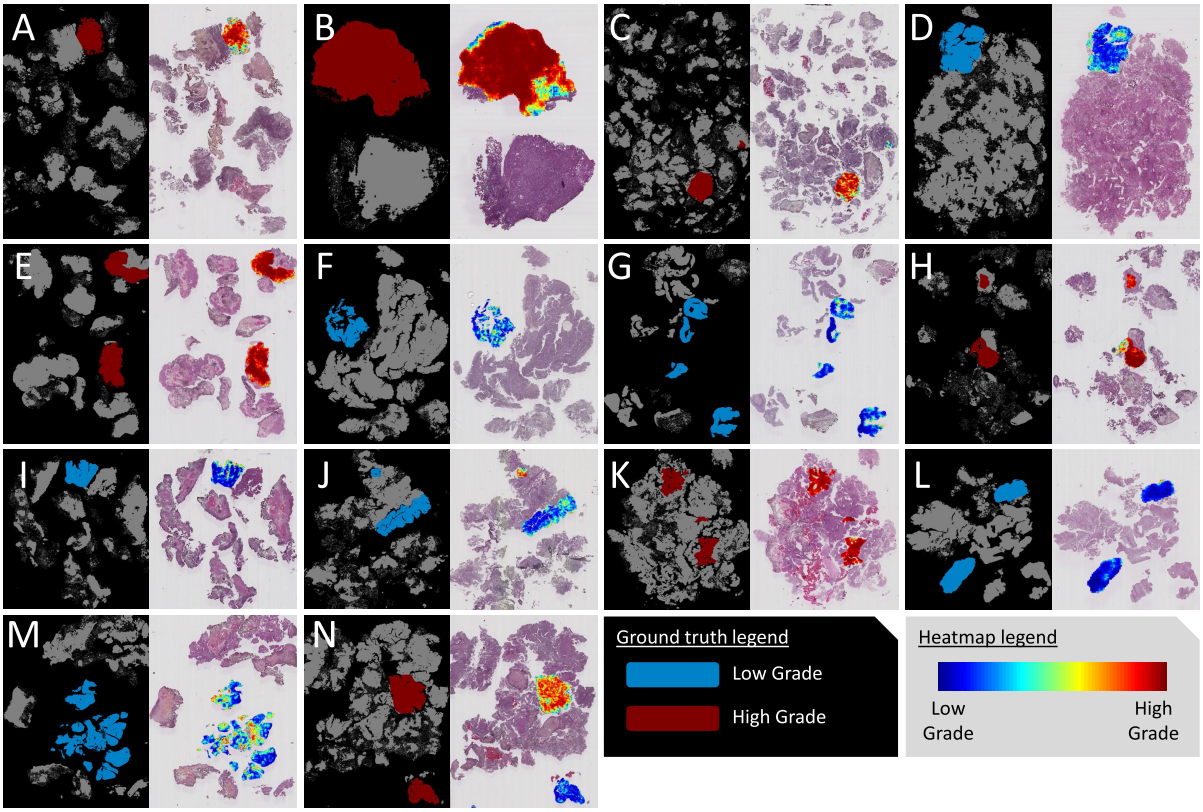

FIGURE 6. Ground truth annotations vs. model prediction. The WSI with a black background is the ground truth images with low- and high-grade annotations. The WSI with a grey background has superimposed a heatmap from the same area as the ground truth and highlights the predictions from the TRI_{WHO04}-model. For quantitative results, see Table 4 and 5.

Jansen *et al.* [22] are not publically available, a comparison is still included as both research results are based on an NMIBC dataset of similar size (328 WSIs from 232 patients vs. our dataset of 300 WSIs), a similar split of the dataset into training, validation, and test, and the use of the same labels (WHO04). The results in Table 6 compare the slide-level sensitivity, specificity, and accuracy for our proposed TRI_{grade} pipeline, to the results reported in table 3 from [22]. We achieve better results on all metrics, and with 45 of the 50 WSIs correctly predicted, we achieve an accuracy of 0.90.

Training and validation accuracy from the training of the TRI_{WHO04}-5000-AUG model is shown in Fig. 7. The model uses frozen pre-trained weights for the VGG16 networks, and only the last layers in the model have random weights which are being optimized. The model uses the largest training dataset from Table 2 with a mini-batch size of 128, resulting in a large number of weight updates per epoch, and the majority of the accuracy is achieved from the first epoch. After the initial epoch, the validation accuracy is not improving too much. This is most likely because the datasets use imprecise weak labels (e.g., all urothelium triplets extracted from a high-grade WSI will have the class label high-grade, but not all triplets from this WSI will represent high-grade tissue). Note also that all the urothelium triplets from all the WSIs in the validation set are predicted before Tensorflow computes the accuracy score for the validation set.

A. USAGE SCENARIOS

The automatic TRI_{grade} system presented in this paper has many potential applications. The tissue model we presented in Wetteland *et al.* [13] provides the tissue segmentation maps, which clinicians can use to discriminate urothelium regions from other tissue classes. This can be a valuable tool to aid pathologists in examining the whole-slide images by focusing their attention on the diagnostic relevant areas of the stained specimen. With the addition of the TRI_{WHO04}-model presented in this paper, the focus can not only be aimed towards the urothelium regions in general but be further narrowed down to the most *severe* urothelium regions.

The automated slide-level prediction can potentially be used to prioritize high-grade patients for earlier examination. Also, it can be used as input to an automatic prognostic tool and output a measure of the patient's overall clinical outcome, such as the risk of recurrence, 1-yr and 5-yr survival rate, and mortality. In the future, it is also a possibility to use it in an automatic system that predicts how a patient will respond to a given treatment and therapy program.

B. LIMITATIONS

In the paper, we train a model to classify urothelium tissue into two classes, low- and high-grade carcinoma. However, it is also a possibility that the urothelium tissue can be healthy non-cancerous tissue. Since our models are dependent on the weak slide-level label, and all cases in the data material are

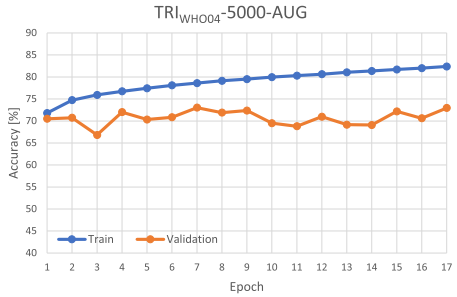

FIGURE 7. Training and validation accuracy for the TRI$_{WHO04}$-5000-AUG model. The model is trained on imprecise weak labels, using the largest training set in Table 2. Results are shown for tile-level prediction on the entire training and validation sets. Validation accuracy is computed at the end of each epoch.

diagnosed with cancer, we currently do not have any training material containing non-cancerous samples.

All WSIs in this study are collected from the same laboratory and consists of high quality with relatively small variations in stain colors and little blur. This is both a strength in the sense that we have produced good models and reliable predictions, but also a limitation in the sense that we do not know how the system will perform on slides of lower quality.

C. FUTURE WORK

In future work, preprocessing steps might be added to deal with color variations, blur, and folded tissue, or the tissue segmentation model can be updated with a new class for blur, providing a more generalized system.

From [13] it was concluded that for the tissue segmentation task, the multiscale TRI-25x-100x-400x model (which is used as the TRI$_{tissue}$-model in this work) provided the best performance. Following, a multiscale model was adopted for the grading task as well, with the masking of the urothelium tissue performed at the 400x level. However, the large field-of-view provided by the 25x and 100x magnification will bring neighboring tissue types into the triplet, like, for example, damaged tissue, which might affect the performance in such areas. In future work, we would like to use the tissue segmentation maps and not only extract the urothelium tissue but also mask out unwanted regions of damaged tissue and blood. Incorporating attention modules is also something we will try, which would further help explain what parts of the WSI are responsible for the predictions.

Cells of low-grade cancer often resemble that of non-cancerous cells, and high-grade cells have a more abnormal appearance and are disorganized. Thus, we expect that non-cancerous tissue would be predicted as low-grade carcinoma. However, this is our expectation as we do not have verified material to test this on. To better detect these non-cancerous regions in the future, we would have to expand our training dataset to include examples of non-cancerous urothelium. The TRI$_{WHO04}$-model architecture must be updated to include one additional class on the output and then be trained on the updated dataset.

The proposed model uses three VGG16 networks as feature extractors. In the future, we would like to experiment with other deep learning networks for our base model. Newer deep learning models continuously improve the results on datasets like ImageNet, and could potentially improve feature extraction of urothelium tissue. We also plan to look into different ways of fusing the multiscale information, both for the tissue classifier (TRI$_{tissue}$) and grade-classifier (TRI$_{WHO04}$).

VI. CONCLUSION

In this paper, we have proposed a TRI$_{grade}$ pipeline for automatic grading of urothelial carcinoma slides based on the WHO04 grading system. First, the slide is segmented into the tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background). Next, tiles are extracted at three magnification levels (25x, 100x, and 400x) from the urothelium regions. The three tiles form a triplet, which is fed sequentially to a multiscale CNN-based WHO04 grading model.

The proposed method will generate a tissue segmentation map, helpful for the clinicians to easier find diagnostic relevant regions during an examination. The system will also output a WHO04 grade heatmap, highlighting the most severe urothelium tissue regions, beneficial for the pathologists who can focus their limited per-patient time on the most important regions in the WSI. Finally, the system produces a slide-level WHO04 grade that could potentially be used to prioritize high-grade patients for earlier examination, as well as suggest the diagnosis to the pathologist.

Ten WHO04 grade models were trained on datasets of varying sizes. Note that all the same number of WSI were used all the time, but a different number of triplets were extracted from each WSI, constituting the training set. The model trained on the largest training dataset achieved the best result, a weighted average F1-score of 0.90 on the test set. This model was further evaluated on a segmentation test set, where low- and high-grade regions were annotated by a pathologist. On this task, the model got an average F1-score of 0.91 on both the low-grade and high-grade classes.

The system as a whole can be used by clinicians and pathologists to potentially improve their decision-making and further help patients by receiving correct diagnoses and treatment.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA A, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] O. M. Mangrud, "Identification of patients with high and low risk of progresson of urothelial carcinoma of the urinary bladder stage Ta and T1," Ph.D. dissertation, Dept. Clin. Med., Fac. Med. Dentistry, Univ. Bergen, Bergen, Norway, 2014.

[3] K. D. Sievert, B. Amend, U. Nagele, D. Schilling, J. Bedke, M. Horstmann, J. Hennenlotter, S. Kruck, and A. Stenzl, "Economic aspects of bladder cancer: What are the benefits and costs?" *World J. Urol.*, vol. 27, no. 3, pp. 295–300, Jun. 2009, doi: 10.1007/s00345-009-0395-z.

[4] M. F. Botteman, C. L. Pashos, A. Redaelli, B. Laskin, and R. Hauser, "The health economics of bladder cancer," *PharmacoEconomics*, vol. 21, no. 18, pp. 1315–1330, Dec. 2003, doi: 10.1007/bf03262330.

[5] V. Kvikstad, O. M. Mangrud, E. Gudlaugsson, I. Dalen, H. Espeland, J. P. A. Baak, and E. A. M. Janssen, "Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas," *Diagnostic Pathol.*, vol. 14, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/s13000-019-0868-3.

[6] O. M. Mangrud, R. Waalen, E. Gudlaugsson, I. Dalen, I. Tasdemir, E. A. M. Janssen, and J. P. A. Baak, "Reproducibility and prognostic value of WHO1973 and WHO2004 grading systems in TaT1 urothelial carcinoma of the urinary bladder," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e83192, doi: 10.1371/journal.pone.0083192.

[7] L. Browning, E. Fryer, D. Roskell, K. White, R. Colling, J. Rittscher, and C. Verrill, "Role of digital pathology in diagnostic histopathology in the response to COVID-19: Results from a survey of experience in a UK tertiary referral hospital," *J. Clin. Pathol.*, vol. 74, no. 2, pp. 129–132, Feb. 2021, doi: 10.1136/jclinpath-2020-206786.

[8] M. G. Hanna, V. E. Reuter, O. Ardon, D. Kim, S. J. Sirintrapun, P. J. Schüffler, K. J. Busam, J. L. Sauter, E. Brogi, L. K. Tan, and B. Xu, "Validation of a digital pathology system including remote review during the COVID-19 pandemic," *Modern Pathol.*, vol. 33, no. 11, pp. 2115–2127, Nov. 2020, doi: 10.1038/s41379-020-0601-5.

[9] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *Lancet Oncol.*, vol. 20, no. 5, pp. e253–e261, 2019, doi: 10.1016/S1470-2045(19)30154-8.

[10] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology," *Nature Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, Nov. 2019, doi: 10.1038/s41571-019-0252-y.

[11] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Med. Image Anal.*, vol. 33, no. 6, pp. 170–175, 2016, doi: 10.1016/j.media.2016.06.037.

[12] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016, *arXiv:1606.05718*. [Online]. Available: http://arxiv.org/abs/1606.05718

[13] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. A. M. Janssen, "A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides," *Technol. Cancer Res. Treatment*, vol. 19, Jan. 2020, Art. no. 153303382094678, doi: 10.1177/1533033820946787.

[14] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers*, vol. 11, no. 9, p. 1235, Aug. 2019, doi: 10.3390/cancers11091235.

[15] S. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–8, Dec. 2020, doi: 10.1038/s41746-020-00324-0.

[16] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019, doi: 10.1038/s41591-019-0508-1.

[17] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101563, doi: 10.1016/j.media.2019.101563.

[18] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, "Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images," *Pattern Recognit. Lett.*, vol. 133, pp. 232–239, May 2020, doi: 10.1016/j.patrec.2020.03.011.

[19] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, and K. A. Iczkowski, "Pathologist-level grading of prostate biopsies with artificial intelligence," 2019, *arXiv:1907.01368*. [Online]. Available: http://arxiv.org/abs/1907.01368

[20] J. D. Ianni, R. E. Soans, S. Sankarapandian, R. V. Chamarthi, D. Ayyagari, T. G. Olsen, M. J. Bonham, C. C. Stavish, K. Motaparthi, C. J. Cockerell, T. A. Feeser, and J. B. Lee, "Tailored for real-world: A whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Dec. 2020, doi: 10.1038/s41598-020-59985-2.

[21] O.-J. Skrede, S. De Raedt, A. Kleppe, T. S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J. A. Nesheim, F. Albregtsen, and I. N. Farstad, "Deep learning for prediction of colorectal cancer outcome: A discovery and validation study," *Lancet*, vol. 395, no. 10221, pp. 350–360, 2020, doi: 10.1016/S0140-6736(19)32998-8.

[22] I. Jansen, M. Lucas, J. Bosschieter, O. J. de Boer, S. L. Meijer, T. G. van Leeuwen, H. A. Marquering, J. A. Nieuwenhuijzen, D. M. de Bruin, and C. D. Savci-Heijink, "Automated detection and grading of non–muscle-invasive urothelial cell carcinoma of the bladder," *Amer. J. Pathol.*, vol. 190, no. 7, pp. 1483–1490, Jul. 2020, doi: 10.1016/j.ajpath.2020.03.013.

[23] Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F. K. Khalil, S. I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, and L. Yang, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 236–245, May 2019, doi: 10.1038/s42256-019-0052-1.

[24] M. Lucas, I. Jansen, T. G. van Leeuwen, J. R. Oddens, D. M. de Bruin, and H. A. Marquering, "Deep learning–based recurrence prediction in patients with non–muscle-invasive bladder cancer," *Eur. Urol. Focus*, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2405456920303102, doi: 10.1016/j.euf.2020.12.008.

[25] K. Sirinukunwattana, N. K. Alham, C. Verrill, and J. Rittscher, "Improving whole slide segmentation through visual context—A systematic study," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 192–200, doi: 10.1007/978-3-030-00934-2_22.

[26] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi, "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3852–3861, doi: 10.1109/CVPR42600.2020.00391.

[27] X. Zhang, F. Dong, G. Clapworthy, Y. Zhao, and L. Jiao, "Semi-supervised tissue segmentation of 3D brain MR images," in *Proc. 14th Int. Conf. Inf. Visualisation*, vol. 2688, Jul. 2010, pp. 623–628. [Online]. Available: http://CEUR-WS.org/Vol-2688/paper14.pdf

[28] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. Janssen, "Multiclass tissue classification of whole-slide histological images using convolutional neural networks," in *Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods*, vol. 1, 2019, pp. 320–327, doi: 10.5220/0007253603200327.

[29] J. Urdal, K. Engan, V. Kvikstad, and E. A. M. Janssen, "Prognostic prediction of histopathological images by local binary patterns and RUSBoost," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 2349–2353, doi: 10.23919/EUSIPCO.2017.8081630.

[30] M. Babjuk, A. Böhle, M. Burger, O. Capoun, D. Cohen, E. M. Compérat, V. Hernández, E. Kaasinen, J. Palou, M. Rouprêt, B. W. G. van Rhijn, S. F. Shariat, V. Soukup, R. J. Sylvester, and R. Zigeuner, "EAU guidelines on non–muscle-invasive urothelial carcinoma of the bladder: Update 2016," *Eur. Urol.*, vol. 71, no. 3, pp. 447–461, Mar. 2017, doi: 10.1016/j.eururo.2016.05.041.

[31] K. Martinez and J. Cupitt, "VIPS—A highly tuned image processing software architecture," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Sep. 2005, pp. 2–574, doi: 10.1109/ICIP.2005.1530120.

[32] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019, doi: 10.1016/j.media.2019.03.009.

[33] R. Wetteland, K. Engan, and T. Eftestøl, "Parameterized extraction of tiles in multilevel gigapixel images," in *Proc. 12th Int. Symp. Image Signal Process. Anal. (ISPA)*, 2021.

[34] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[35] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

Paper I

Kvikstad et al. Diagnostic Pathology (2019) 14:90
https://doi.org/10.1186/s13000-019-0868-3

Diagnostic Pathology

RESEARCH Open Access

Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas

Vebjørn Kvikstad[1,2*†], Ok Målfrid Mangrud[3†], Einar Gudlaugsson[1], Ingvild Dalen[4], Hans Espeland[5], Jan P. A. Baak[1,6,7†] and Emiel A. M. Janssen[1,2†]

Abstract

Background: European treatment guidelines for pTa and pT1 urinary bladder urothelial carcinoma depend highly on stage and WHO-grade. Both the WHO73 and the WHO04 grading systems show some intra- and interobserver variability. The current pilot study investigates which histopathological features are especially sensitive for this undesired lack of reproducibility and the influence on prognostic value.

Methods: Thirty-eight cases of primary non-muscle invasive urothelial carcinomas, including thirteen cases with stage progression, were reviewed by three pathologists. Thirteen microscopic features were extracted from pathology textbooks and evaluated separately. Reproducibility was measured using Gwet's agreement coefficients. Prognostic ability regarding progression was estimated by the area under curve (AUC) of the receiver operating characteristics (ROC) function.

Results: The best reproducible features (Gwet's agreement coefficient above 0.60) were papillary architecture, nuclear polarity, cellular maturation, nuclear enlargement and giant nuclei. Nucleoli was the strongest prognostic feature, and the only feature with an AUC above 0.70 for both grading systems, but reproducibility was not among the strongest. Nuclear polarity also had prognostic value with an AUC of 0.70 and 0.67 for the WHO73 and WHO04, respectively. The other features did not have significant prognostic value.

Conclusions: The reproducibility of the histopathological features of the different WHO grading systems varied considerably. Of all the features evaluated, only nuclear polarity was both prognostic and significantly reproducible. Further validation studies are needed on these features to improve grading of urothelial carcinomas.

Keywords: Papillary urothelial carcinoma, Grading, Reproducibility, Prognosis

Background

Bladder cancer is the ninth most frequently diagnosed cancer worldwide. The incidence is highest in developed countries, and is the fourth most common cancer among men in Norway [1, 2]. Urothelial carcinoma accounts for about 90% of bladder cancers in industrialized countries [3], and 70–80% of these are non-muscle-invasive bladder cancers (NMIBC), pTa, pT1 or pTis, on first diagnosis. Among these 50–70% will recur, while only 15–25% will progress to a higher stage [4]. The follow-up of these patients is labor-intensive [5, 6], causing massive costs for the health care systems [7].

Papillary urothelial carcinomas are the most frequent in western countries and are graded based on the degree of anaplasia. In 1973 the World Health Organization (WHO) introduced a classification system, in which papillary carcinomas were divided into three groups; grades 1, 2 and 3 (WHO73). A new classification system was

* Correspondence: vebjorn.kvikstad@sus.no
†Jan P. A. Baak and Emiel A.M. Janssen are Both senior authors contributed equally.
[1]Department of Pathology, Stavanger University Hospital, Stavanger, Norway
[2]Department of Mathematics and Natural Science, University of Stavanger, Stavanger, Norway
Full list of author information is available at the end of the article

© The Author(s). 2019 **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.

introduced in the 2004 WHO Classification of tumours of the urinary system ("blue book"), following an International Society of Urological Pathology (ISUP) consensus conference in 1998 (WHO04). This grading system is maintained in the 4th.edition, 2016, of the WHO blue book. Currently, both systems are being used in routine diagnostics at pathology departments around the world [8]. The WHO04 classification system divided the papillary urothelial tumours into papillary urothelial neoplasm of low malignant potential (PUNLMP), low and high grade carcinomas. The histologic features are described in detail, aiming to improve reproducibility. However, several studies have shown considerable inter-observer variability for both classification systems [9–11]. In a recent review Soukup et al. [12] conclude, on behalf of the European Association of Urology (EAU), that the "Current grading classifications in NMIBC are suboptimal", both with regards to reproducibility (poor to fair) and with regards to prognostication.

Grading of papillary urothelial carcinomas according to the WHO73 and the WHO04 classification systems is based on a variety of histopathological features. However, these are not necessarily consciously and systematically analysed one-by-one in a routine diagnostic setting by diagnostic pathologists. Rather than a time consuming analytical approach, many pathologists make a first-glance low-magnification diagnosis, and zoom in on special areas or features to get their diagnosis confirmed. This is a quick, time-effective method but a drawback is lack of reproducibility, with classification shifts from one to other grades and hence prognostic variation as well.

The aim of this pilot study was to systematically analyse the reproducibility and prognostic value of each of the microscopic features. As far as we know, this has not been done before; although previous work on mitotic activity in urothelial carcinoma has found mitosis to be a prognostic factor [13, 14].

Methods

The study was approved by the Norwegian Regional Ethics Committee (#106/09). All patients with a primary non-muscle-invasive papillary urothelial carcinoma, at Stavanger University Hospital (SUH) from January 2002 to January 2007 were investigated ($N = 228$). All patients with urothelial carcinoma outside the urinary bladder (except for those with tumour in the pericollicular area in the urethra) were excluded. Thirty-five cases were excluded because of inadequate sample quality (necrotic tumour, fragmentation, thermal damage and insufficient material), leaving a total of 185 patients. Of these, 13 patients had stage progression; 12 within 5 years, and one after 5 years and 1 month.

In this pilot study we selected a group of 38 patients, including the 13 with progression and 25 without progression. Among the 13 patients with progression 10 were high grade and 3 were low grade according to WHO04. Patients without progression were randomly selected from the remaining 172 patients. There were no statistical significant differences between the grade, age, sex, recurrence or follow-up time of the selected 25 and the other 147 patients without progression.

Tumour tissue was obtained by transurethral resection or biopsy. Tissue was fixed in 4% buffered formaldehyde, dehydrated and embedded in paraffin. For microscopic evaluation four μm thick sections stained with haematoxylin-eosin-saffron (HES) were used.

The patients were treated according to the national guidelines at the time of diagnoses. The treatment consisted of transurethral resection (TUR), followed by a single instillation of a cytotoxic agent (epirubicin hydrochloride). Most patients defined as high risk patients were offered regular instillations with Bacillus Calmette Guérin (BCG), but some were offered alternative treatment with regular instillations containing a combination of epirubicin hydrochloride and interferon alpha. High risk patients included stage T1, grade 3 (WHO73), concurrent or later carcinoma in situ (pTis), three or more separate tumours diagnosed within 18 months or recurrences at multiple sites at first or second follow-up. Provided that the first follow-up cystoscopy was negative, patients with Ta grade 1 tumours would undergo control cystoscopies 3 months after initial diagnosis, 9 months later, and then annually for 5 years. All other patients would have cystoscopies every 3 months for the first 2 years, every 4 months for the 3rd year, every 6 month the 4th and 5th years, followed by annual cystoscopies thereafter.

Follow-up data were retrieved from the medical- and laboratory records at SUH. We defined progression as any advance in TNM stage, including both from pTa to pT1 or to pT2, and from pT1 to pT2. Progression to muscle invasive disease is clinically most relevant due to major differences in therapy. We also included cases with progression from pTa to pT1 as these tumours have gained the capability to infiltrate the stroma, a basic trait for progression.

The histopathological features constituting the grading systems were derived from urological pathology textbooks [15–17]. A list of the microscopic features and their interpretation, both for WHO73 and WHO04, is shown in Table 1. We extracted 13 features: papillae architecture, superficial layer, papillary fusion, nuclear polarity, cell maturation, cohesion, mitoses, nuclear enlargement, nuclear shape, nuclear hyperchromasia, chromatin pattern, nucleoli and giant nuclei.

All specimens were evaluated by three pathologists, focusing on grading criteria of the individual features, one at a time, for both WHO73 and WHO04. In tumours

Table 1 The microscopic features with descriptions for each grade (WHO73/ 04)

	WHO73			WHO04	
	Grade 1	Grade 2	Grade 3	Low grade	High grade
Architecture					
Papillae	Delicate	Varies	Broad, varies	Slender	Broad
Superficial layer (umbrella cell layer)	Usually present	Usually present	Partially or completely lost	Usually present	Partially or completely lost
Papillary fusion	Some	Varies	Common	Some	Varies
Nuclear arrangement					
Polarity	Preserved	Moderate loss	Lost	Preserved, moderate loss	Lost
Maturation	Normal	Some	Lost	Preserved, moderate loss	Lost
Cohesion	Normal	Some	Lost	Some	Lost
Proliferation					
Mitotic figures	Rare, basal	Lower half	Common, atypical	Rare	Common
Nuclear atypia					
Nuclear enlargement	Mild	Mild	Varies	Mild	Varies
Nuclear shape	Uniform	Moderate variation	Pleomorphic	Moderate variation	Pleomorphic
Nuclear hyperchromasia	Mild	Moderate	Varies	Mild to moderate	Varies
Chromatin pattern	Finely granular	Granular	Coarse	Fine	Coarse
Nucleoli	Occasional	Occasional	Common	Occasional	Common
Giant nuclei	No	No	Yes	No	Yes

with morphological heterogeneity the "worst" area was graded. The evaluations were done without any knowledge about the original diagnosis or the other pathologists' results. At a later stage, all three pathologists contributed to a consensus assessment for all the variables. Concerning the WHO04, only low grade and high grade were used as only three cases were classified as PUNLMP in our original cohort. In a previous study we found that recurrence and stage progression in the PUNLMPs and the low grade tumours by univariate survival analysis on our material were no different [18]. A later publication by Kim et al. [19] also showed no difference in progression between PUNLMP and low grade carcinomas.

Statistics

Reproducibility was measured using Gwet's AC_1 agreement coefficient [20] for features with two categories, and using Gwet's AC_2 agreement coefficient with quadratic weights for features with > 2 categories [21]. Fleiss' generalized kappa [22] is also reported for reference; however, due to its vulnerability to skewed marginal distributions [23], the focus in this paper is on Gwet's agreement coefficients. A coefficient of < 0.2 is defined as poor agreement, 0.2–0.4 fair agreement, 0.4–0.6 moderate agreement, 0.6–0.8 good agreement and > 0.8 as very good agreement [24]. Confidence intervals (CIs) for the reliability measures were based on the normal approximation [21].

Prognostic ability with regard to progression for the consensus classification of each feature was estimated by the area under curve (AUC) of the receiver operating characteristics (ROC) function, which is reported with a normal based confidence interval [25]. Statistical analysis was performed in R version 3.4.0 with syntax provided at http://www.agreestat.com/r_functions.html (downloaded 24.05.2018) and with package pROC [25].

Results

The median age at diagnosis was 72 years (range 56–87). Thirty patients were male (79%) and eight female (21%) (M:F ratio = 3.8). Median follow-up time was 73 months (range 5–168). Not all samples were regarded adequate for assessing all the microscopic features by all three pathologists. These cases were not included in the calculation of reliability for that particular feature (Table 2). At the consensus meeting, there was agreement that two cases could not be used to assess the feature "papillary fusion". There were also two cases in which "maturation" could not be reliably assessed, and in one case "superficial layer" could not be assessed. This left between 36 to 38 total cases for each of the different features.

The reproducibility varies among the different microscopic features according to the calculated Gwet's $AC_{1/2}$ agreement coefficient (Table 2). The values range from 0.47 for mitosis in the WHO73 system to 0.85 for giant nuclei. This corresponds to moderate to very good

Table 2 Reproducibility and prognostic value for each of the microscopic characteristics

Feature	n*	AC_1/AC_2 (95% CI)	Fleiss' κ (95% CI)	n**	Consensus grade (prob. of progression)	AUC_{ROC} (95% CI) ***
Papillae73	36	0.62 (0.42 to 0.82)	0.63 (0.45 to 0.82)	38	Delicate (1/10) Varies (4/11) Broad, varies (8/17)	0.67 (0.51 to 0.83)
Papilae04	36	0.61 (0.39 to 0.82)	0.59 (0.37 to 0.81)	38	Slender (3/16) Broad (10/22)	0.64 (0.49 to 0.80)
Superficial layer73/04	36	0.51 (0.30 to 0.73)	0.50 (0.29 to 0.72)	37	Usually present (4/12) Partially lost (8/25)	0.49 (0.33 to 0.66)
Papillary fusion73	34	0.64 (0.44 to 0.84)	0.67 (0.48 to 0.86)	36	Some (2/13) Varies (4/9) Common (7/14)	0.67 (0.50 to 0.84)
Papillary fusion04	34	0.53 (0.32 to 0.75)	0.53 (0.31 to 0.75)	36	Some (4/19) Varies (8/17)	0.67 (0.51 to 0.84)
Polarity73	38	0.68 (0.53 to 0.82)	0.70 (0.55 to 0.84)	38	Preserved (1/9) Moderate (4/14) Lost (8/15)	0.70 (0.54 to 0.86)
Polarity04	38	0.66 (0.47 to 0.86)	0.63 (0.43 to 0.84)	38	Preserved (5/23) Lost (8/15)	0.67 (0.50 to 0.83)
Maturation73	36	0.60 (0.43 to 0.78)	0.59 (0.42 to 0.76)	36	Normal (1/9) Some (5/14) Lost (6/13)	0.66 (0.49 to 0.83)
Maturation04	36	0.62 (0.42 to 0.82)	0.60 (0.40 to 0.81)	36	Some (6/23) Lost (6/13)	0.60 (0.43 to 0.78)
Cohesion73	37	0.57 (0.42 to 0.71)	0.47 (0.28 to 0.65)	38	Normal (1/12) Some (9/21) Lost (3/5)	0.71 (0.56 to 0.85)
Cohesion04	37	0.54 (0.30 to 0.77)	0.23 (−0.02 to 0.47)	38	Some (10/33) Lost (3/5)	0.58 (0.44 to 0.71)
Mitosis73	38	0.47 (0.23 to 0.71)	0.41 (0.18 to 0.64)	38	Rare, basal (8/31) Lower half (1/1) Common, atypical (4/6)	0.65 (0.50 to 0.80)
Mitosis04	38	0.64 (0.43 to 0.85)	0.49 (0.25 to 0.72)	38	Rare (9/32) Common (4/6)	0.61 (0.47 to 0.76)
Nuclear enlargement73/04	38	0.65 (0.45 to 0.85)	0.65 (0.45 to 0.84)	38	Mild (4/19) Varies (9/19)	0.65 (0.48 to 0.81)
Nuclear shape73	38	0.58 (0.41 to 0.74)	0.51 (0.32 to 0.69)	38	Uniform (3/10) Moderate (8/23) Pleomorphic (2/5)	0.53 (0.36 to 0.71)
Nuclear shape04	38	0.58 (0.34 to 0.81)	0.41 (0.21 to 0.61)	38	Moderate (11/33) Pleomorphic (2/5)	0.52 (0.40 to 0.64)
Nuclear hyperchromasia73	38	0.51 (0.38 to 0.65)	0.51 (0.35 to 0.68)	38	Mild (3/11) Moderate (6/17) Varies (4/10)	0.56 (0.37 to 0.74)
Nuclear hyperchromasia04	38	0.51 (0.28 to 0.74)	0.43 (0.21 to 0.65)	38	Mild to moderate (9/28) Varies (4/10)	0.53 (0.38 to 0.69)
Chromatin pattern73	38	0.51 (0.29 to 0.73)	0.46 (0.26 to 0.67)	38	Finely granular (7/25) Granular (4/10) Coarse (2/3)	0.60 (0.43 to 0.78)
Chromatin pattern04	38	0.66 (0.47 to 0.86)	0.55 (0.31 to 0.79)	38	Fine (9/31) Coarse (4/7)	0.59 (0.45 to 0.74)
Nucleoli73/04	38	0.54 (0.33 to 0.76)	0.54 (0.33 to 0.75)	38	Occasional (2/16) Common (11/22)	**0.70 (0.56 to 0.85)**
Giant nuclei	38	**0.85 (0.72 to 0.98)**	0.78 (0.59 to 0.98)	38	No (8/28) Yes (5/10)	0.59 (0.43 to 0.75)

AC_1/AC_2 Gwet's AC_1/AC_2 coefficient, *CI* Confidence interval, *AUC_{ROC}* Area under Receiver Operating Characteristics Curve.
* Number of cases evaluated by all three pathologists
** Number of cases for which consensus was reached

reproducibility. The other features yielded evenly distributed values, with papillae architecture, nuclear polarity, cell maturation, nuclear enlargement and giant nuclei as the most reproducible, all with Gwet's $AC_{1/2}$ agreement coefficient above 0.60 (=good agreement) for both grading systems. Several of the values have very wide

confidence intervals, making them less robust. For instance, for mitosis73 the confidence interval ranges from 0.23 to 0.71.

Prognostic ability for the different features, estimated by AUC, ranged from 0.49 for superficial layer, to 0.71 for cohesion in WHO73. To qualify as reliable, we wanted the features to be convincing (> 0.7) for both WHO73 and WHO04. For instance, cohesion generated an AUC of 0.58 for WHO04, and should therefore not be relied on in our material. Only nucleoli achieved an AUC above 0.7 for both WHO73 and WHO04, which is seen as an acceptable discrimination for progression or not. Polarity tends to show some prognostic information for both grading systems with AUC 0.70/ 0.67 for WHO73 and WHO04 respectively. These two features and papillary fusion gave estimated confidence intervals ≥0.5 for both grading systems. The other ten features showed no statistical significant prognostic value.

Nuclear polarity was the only feature with both reasonable reproducibility and prognostic value in this pilot study.

Discussion

Grade is seen as one of the most important prognostic factors in bladder cancer, with impact on treatment and patient follow-up. As reproducibility of both WHO73 and WHO04 is suboptimal, we systematically analysed the reproducibility and prognostic value of each of the microscopic features described as being part of grading. Each of the 13 features, which theoretically should be used to reach the final grade, carries its own uncertainty in terms of reproducibility and prognostic value.

In the absence of a formal prognostic decision tree of microscopic features in urinary bladder cancers, and lack of a descriptive atlas with typical pictures, pathologists will emphasize each feature differently while grading a urinary bladder tumour. The assessment of grade is therefore more or less based on intuition, as the features are not evaluated in a systematic manner, and only rarely truly quantitatively. This partially explains the considerable difficulty with reproducibility. Furthermore, the thresholds for the different subclasses of each of the included features are very subjective (example: the described thresholds for cohesion are: normal, some or lost). Such descriptive and subjective criteria lead to diagnostic confusion. In the process of grading, pathologists will also be challenged by laboratory variables like section thickness which might blur nuclear hyperchromasia or the introduction of artefacts that might mimic dyscohesiveness. The individual prognostic values of these features has never been analysed separately in urinary bladder tumours.

Before our analyses we expected mitoses to be a useful feature, as reported in a previous study on bladder cancer [13]. In the current analyses, mitosis was one of the least reproducible and prognostic features. However, mitotic activity in the current study was assessed in a semi-quantitative manner. Contrary, previous studies which reported mitoses as a strong prognostic factor, counted mitoses in a defined area by using the protocol for Mitotic Activity Index (MAI) as it is used and developed for breast cancer, and the final number of mitoses was used to categorize the tumours. When grading according to either of the WHO-systems, a rough mitotic impression, rather than a formalized mitotic count is used. This may explain the differences in prognostic value and reproducibility. Such a prognostic difference between mitotic activity as the MAI (truly quantitative) and mitotic impression (a rough estimate) has previously been shown in breast cancer [26], and may be true for urothelial carcinoma as well.

To be clinically useful, a grading system should be well reproducible to assure the intended sensitivity and specificity. As known the final grade is the sum of an evaluation of different microscopic features, therefore if one of these features is not truly quantitative, it inevitably will lack reproducibility and this will affect the final grade as well. Individual features may have a prognostic potential, which might be hidden by low overall reproducibility. It is crucial to minimalize the interobserver variability, making these features more reliable before extracting and emphasizing the features giving the best prognostic information. These features might be evaluated separately in a new grading system.

One way to improve reproducibility could be to provide pathologists with an image atlas with examples of the various features, facilitating comparison with the tumour to be graded. In prostate adenocarcinoma, the Gleason score has been well documented, tested and tried since its introduction in 1966 [27]. It has been claimed that the success of the system may in part be attributed to the ease of application and the simplicity of the original drawings [15]. Although the Gleason score has issues regarding reproducibility as well, especially when differentiating between Gleason grade group 2 and 3 [28, 29], the system as a whole has proven to be an important predictor of prognosis [30, 31]. A similar system with simplified, stylized illustrations may improve grading reproducibility in bladder cancer as well.

In this study nuclear polarity stands out as the most valuable histopathological feature in grading. This supports the current view that architectural and cytological order versus disorder decides whether a lesion should be regarded as low or high grade in the WHO04 grading system. Strict definitions will be necessary to further improve reproducibility of this feature as well. One approach could be to grade nuclear polarity according to how much the axis of the nuclei tends to deviate from a line perpendicular to the basement membrane (Fig. 1).

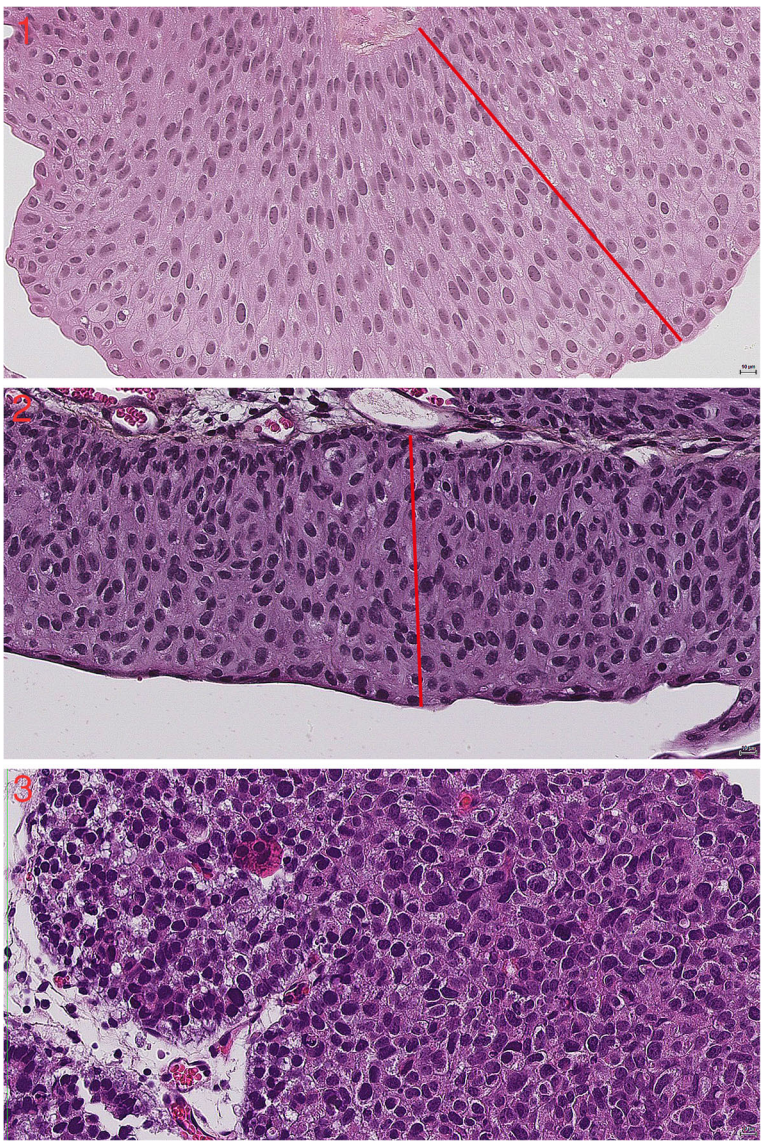

Fig. 1 The images 1–3 show decreasing nuclear polarity at 40 x magnification. The red line is for comparison with the axis of the nuclei

The introduction of digital pathology introduces a multitude of possibilities for measurement of structures like nuclei, nucleoli and papillae. This can be exploited in grading, in an attempt to achieve standardization. Digital images can be further analysed by computer based algorithms, thereby analysing features not easily measured directly, like polarity, nuclear shape and mitotic Figures. A first attempt, using a local binary pattern (LBP) and local variance (VAR) operators followed by a RUSboost classifier, on a small test set of 42 patients with NMIBC resulted in an accuracy of 70%, a sensitivity of 84% and a specificity of 45% for prediction of recurrences [32]. Although only performed using a small dataset these results show the potential of these methods. Further studies using bigger datasets are necessary to further investigate these new measurements.

The value of the data in this pilot study is limited by the small sample size, not allowing any final conclusions. Although, our data suggest a substantial variety among the different histopathological features when it comes to reproducibility. Also, the prognostic value is disappointing for most of the features. Our data calls for further validation studies to highlight the most reproducible and most prognostic microscopic features making up the current grading system. We hope this article will contribute to developing a new approach.

when it comes to grading of papillary urothelial carcinomas.

Conclusion

WHO grading is based on the use of 13 histopathological features, which in our material vary considerably in reproducibility and prognostic value. Of all the features evaluated in this small study, only nuclear polarity was both reasonably prognostic and reproducible. Further validation studies on the individual histopathological features are needed to improve the assessment of grade of urothelial carcinomas. A new grading system should be based upon more clear-cut definitions and features with true prognostic value.

Abbreviations

AUC: Area under curve; BCG: Bacillus Calmette Guérin; CI: Confidence interval; EAU: European association of urology; HES: Haematoxylin-eosin-saffron; ISUP: International Society of Urological Pathology; LBP: Local binary pattern; MAI: Mitotic Activity Index; NMIBC: Non-muscle-invasive bladder cancer; PUNLMP: Papillary urothelial neoplasia of low malignant potential; ROC: Receiver operating characteristics; SUH: Stavanger University Hospital; TUR: Transurethral resection; VAR: Local variance; WHO: World Health Organization; WHO04: The World Health Organization grading system from 2004; WHO73: The World Health Organization grading system from 1973

Acknowledgements

We would like to thank Bianca van Diermen Hidle, Melinda Lillesand, Eliza Peixoto Albernaz and Anne Elin Varhaugvik for technical assistance. We also want to thank the Department of Pathology at the Stavanger University Hospital for the opportunity to work on this project.

Authors' contributions

VK updated the data and wrote the article. OMM performed all histopathological evaluation and contributed in writing the article. EG performed histopathological evaluation. ID performed all the statistical analyses. HE provided clinical data. JPAB also performed histopathological evaluations and was involved in designing and supervising the study. Emiel A. M. Janssen designed and supervised the study and contributed to writing the paper. All authors critically evaluated the manuscript. All authors read and approved the final manuscript.

Funding

The authors have no support or funding to report.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

The study was approved by the Norwegian Regional Ethics Committee (#106/09). With approval from REK Vest, informed consent was not obtained as the tissue samples had already been removed for diagnostic and treatment purposes.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Pathology, Stavanger University Hospital, Stavanger, Norway. [2]Department of Mathematics and Natural Science, University of Stavanger, Stavanger, Norway. [3]Department of Pathology, Innlandet Hospital, Lillehammer, Norway. [4]Department of Research, Stavanger University Hospital, Stavanger, Norway. [5]Department of Urology, Stavanger University Hospital, Stavanger, Norway. [6]Medical practice Dr. med. Jan Baak AS, Tananger, Norway. [7]Department of TCM, Faculty of Sports and Health Sciences, Technical University Munich, Munich, Germany.

Received: 25 March 2019 Accepted: 9 August 2019
Published online: 14 August 2019

References

1. Antoni S, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F. Bladder Cancer incidence and mortality: a global overview and recent trends. Eur Urol. 2017;71(1):96–108.
2. Norway Cro. Cancer in Norway 2016 - Cancer incidence, mortality, survival and prevalence in Norway. 2017.
3. Pasin E, Josephson DY, Mitra AP, Cote RJ, Stein JP. Superficial bladder cancer: an update on etiology, molecular development, classification, and natural history. Rev Urol. 2008;10(1):31–43.
4. Moch H, Humphrey PA, Ulbright TM, Reuter VE. World Health Organization Classification of tumours; 2016. p. 77–135.
5. Holmang S, Hedelin H, Anderstrom C, Johansson SL. The relationship among multiple recurrences, progression and prognosis of patients with stages ta and T1 transitional cell cancer of the bladder followed for at least 20 years. J Urol. 1995;153(6):1823–6 discussion 6-7.
6. Larsson P, Wijkstrom H, Thorstenson A, Adolfsson J, Norming U, Wiklund P, et al. A population-based study of 538 patients with newly detected urinary bladder neoplasms followed during 5 years. Scand J Urol Nephrol. 2003; 37(3):195–201.
7. Sievert KD, Amend B, Nagele U, Schilling D, Bedke J, Horstmann M, et al. Economic aspects of bladder cancer: what are the benefits and costs? World J Urol. 2009;27(3):295–300.
8. Babjuk M, Bohle A, Burger M, Capoun O, Cohen D, Comperat EM, et al. EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: update 2016. Eur Urol. 2017;71(3):447–61.
9. Bol MG, Baak JP, Buhr-Wildhagen S, Kruse AJ, Kjellevold KH, Janssen EA, et al. Reproducibility and prognostic variability of grade and lamina propria invasion in stages ta, T1 urothelial carcinoma of the bladder. J Urol. 2003; 169(4):1291–4.
10. Mangrud OM, Waalen R, Gudlaugsson E, Dalen I, Tasdemir I, Janssen EA, et al. Reproducibility and prognostic value of WHO1973 and WHO2004 grading systems in TaT1 urothelial carcinoma of the urinary bladder. PLoS One. 2014;9(1):e83192.
11. Yorukoglu K, Tuna B, Dikicioglu E, Duzcan E, Isisag A, Sen S, et al. Reproducibility of the 1998 World Health Organization/International Society of Urologic Pathology classification of papillary urothelial neoplasms of the urinary bladder. Virchows Arch. 2003;443(6):734–40.
12. Soukup V, Capoun O, Cohen D, Hernandez V, Babjuk M, Burger M, et al. Prognostic Performance and Reproducibility of the 1973 and 2004/2016 World Health Organization grading classification Systems in non-muscle-invasive Bladder Cancer: a European Association of Urology non-muscle invasive bladder Cancer guidelines panel systematic review. Eur Urol. 2017; 72(5):801–13.
13. Bol MG, Baak JP, Rep S, Marx WL, Kruse AJ, Bos SD, et al. Prognostic value of proliferative activity and nuclear morphometry for progression in TaT1 urothelial cell carcinomas of the urinary bladder. Urology. 2002; 60(6):1124–30.
14. Liukkonen T, Rajala P, Raitanen M, Rintala E, Kaasinen E, Lipponen P. Prognostic value of MIB-1 score, p53, EGFr, mitotic index and papillary status in primary superficial (stage pTa/T1) bladder cancer: a prospective comparative study. The Finnbladder Group. Eur Urol. 1999;36(5):393–400.
15. Cheng L, Lopez-Beltran A, MacLennan GT, Montironi R, Bostwick DG. Neoplasms of the urinary bladder. In: Bostwick DG, Cheng L, editors. Urological surgical pathology. 3rd ed. Philadelphia: Elsevier Saunders; 2014. p. 230–317.
16. Reuter VE, Algaba F, Amin MB, Cao D, Cheng L, Comperat E. Non-invasive urothelial lesions. In: Moch H, Humphrey P, Ulbright T, Reuter V, editors. WHO classification of tumours of the urinary system and male genital organs. 4th ed. Lyon: International agency for research on cancer; 2016. p. 99–108.
17. Ordóñez NG, Rosai J. Urinary tract. In: Rosai RJ, editor. Ackermans surgical pathology. 10th ed. Edinburgh: Mosby; 2011. p. 1102–286.
18. Mangrud OM, Gudlaugsson E, Skaland I, Tasdemir I, Dalen I, van Diermen B, et al. Prognostic comparison of proliferation markers and World Health Organization 1973/2004 grades in urothelial carcinomas of the urinary bladder. Hum Pathol. 2014;45(7):1496–503.
19. Kim JK, Moon KC, Jeong CW, Kwak C, Kim HH, Ku JH. Papillary urothelial neoplasm of low malignant potential (PUNLMP) after initial TUR-BT:

comparative analyses with noninvasive low-grade papillary urothelial carcinoma (LGPUC). J Cancer. 2017;8(15):2885–91.
20. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(1):29–48.
21. Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters: advanced analytics, LLC; 2014.
22. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378.
23. Quarfoot D, Levine RA. How robust are multirater interrater reliability indices to changes in frequency distribution? Am Stat. 2016;70(4):373–84.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
25. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77.
26. Skaland I, van Diest PJ, Janssen EA, Gudlaugsson E, Baak JP. Prognostic differences of World Health Organization-assessed mitotic activity index and mitotic impression by quick scanning in invasive ductal breast cancer patients younger than 55 years. Hum Pathol. 2008;39(4):584–90.
27. Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep. 1966;50(3):125–8.
28. Egevad L, Ahmad AS, Algaba F, Berney DM, Boccon-Gibod L, Comperat E, et al. Standardization of Gleason grading among 337 European pathologists. Histopathology. 2013;62(2):247–56.
29. Melia J, Moseley R, Ball RY, Griffiths DF, Grigor K, Harnden P, et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. Histopathology. 2006;48(6):644–54.
30. Chan TY, Partin AW, Walsh PC, Epstein JI. Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy. Urology. 2000;56(5):823–7.
31. Epstein JI, Amin M, Boccon-Gibod L, Egevad L, Humphrey PA, Mikuz G, et al. Prognostic factors and reporting of prostate carcinoma in radical prostatectomy and pelvic lymphadenectomy specimens. Scand J Urol Nephrol Suppl. 2005;216:34–63.
32. Urdal J, Engan K, Janssen EA. Prognostic prediction of histopathological images by local binary patterns and RUSBoost. In: Signal Processing Conference (EUSIPCO), 2017 25th European. Kos, Greece: IEEE; 2017.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Paper II

PLOS ONE

RESEARCH ARTICLE

Mitotic activity index and CD25+ lymphocytes predict risk of stage progression in non-muscle invasive bladder cancer

Melinda Lillesand[1☯]*, Vebjørn Kvikstad[1,2☯], Ok Målfrid Mangrud[3], Einar Gudlaugsson[1], Bianca van Diermen-Hidle[1†], Ivar Skaland[1], Jan P. A. Baak[1,4‡], Emiel A. M. Janssen[1,2‡]

1 Department of Pathology, Stavanger University Hospital, Stavanger, Norway, **2** Department of Mathematics and Natural Science, University of Stavanger, Stavanger, Norway, **3** Department of Pathology, Innlandet Hospital, Lillehammer, Norway, **4** Jan Baak AS, Tananger, Norway

☯ These authors contributed equally to this work.
† Deceased.
‡ These authors also contributed equally to this work.
* melinda.lillesand@sus.no

Check for updates

OPEN ACCESS

Citation: Lillesand M, Kvikstad V, Mangrud OM, Gudlaugsson E, van Diermen-Hidle B, Skaland I, et al. (2020) Mitotic activity index and CD25+ lymphocytes predict risk of stage progression in non-muscle invasive bladder cancer. PLoS ONE 15 (6): e0233676. https://doi.org/10.1371/journal.pone.0233676

Editor: Jason Chia-Hsun Hsieh, Chang Gung Memorial Hospital at Linkou, TAIWAN

Received: December 20, 2019

Accepted: May 10, 2020

Published: June 2, 2020

Copyright: © 2020 Lillesand et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Patient data cannot be shared publicly because of ethical and legal concerns. Anonymized data are available from the Stavanger University Hospital Institutional Data Access / Ethics Committee (contact via email: rek-vest@uib.no, REK vest, Rogaland, Vestland, Norway) for researchers who meet the criteria for access to confidential data. The data underlying the results presented in the study are available from the Biobank-coordinator, Stavanger University Hospital, Norway (email: forskning@sus.no).

Abstract

In urothelial cell type non-muscle invasive urinary bladder carcinoma, TNM stage and WHO grade are widely used to classify patients into low and high-risk groups for prognostic and therapeutic decision-making. However, stage and grade reproducibility and prediction accuracy are wanting. This may lead to suboptimal treatment. We evaluated whether proliferation features, nuclear area of the epithelial cancer cells and the composition of stromal and tumor infiltrating lymphocytes have independent prognostic value. In 183 primary non-muscle invasive bladder cancer patients with long follow-up (median for stage progression cohort: 119 months, range 5-173; median for tumor recurrence cohort: 82, range 3-165) proliferation features Ki67, PPH3 and Mitotic Activity Index (MAI), Mean Nuclear Area (MNA), lymphocyte subsets (CD8+, CD4+, CD25+) and plasma cells (CD138+) were assessed on consecutive sections. Post-resection instillation treatments (none, mitomycin, BCG) were strictly standardized during the intake period. Risk of recurrence was associated with expression of Ki67 ($\leq$ 39 vs. > 39) and Multifocality ($p = 0.01$). Patients with low Ki67 had a higher recurrence rate than those with high Ki67. Lymphocyte composition did not predict recurrence. Stage progression was strongly associated with high values for MAI (>15) and CD25+ (>0.2%). In a multivariate analysis the combination of MAI and CD25+ was the single most prognostic feature ($p<0.001$). Validation of these results in additional, independent studies is warranted.

Introduction

Urothelial cell carcinoma (UCC) is the most common type of carcinoma of the urinary bladder, accounting for about 90% of cases in Western Europe [1]. About 75 to 80% of the UCC are non-muscle invasive bladder cancer (NMIBC) at the time of diagnosis [2] and

Funding: The author(s) received no specific funding for this work. Jan Baak AS did not provide support in the form of salaries for authors and did not have any additional role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Jan Baak AS, Norway commercial affiliation not alter our adherence to PLOS ONE policies on sharing data and materials.

approximately 5 to 10% of these progress to muscle invasive disease [3, 4]. In addition to tumor stage (TNM-classification), grading based on the degree of anaplasia is used as an important prognostic factor. A recent review reported that WHO1973 and WHO04/WHO16 grading systems are suboptimal concerning both reproducibility and prognostic value [5, 6]. On the other hand, earlier reports stated that proliferation features Ki67, PPH3 and Mitotic Activity Index (MAI), as well the nuclear feature Mean Nuclear Area of the largest 10 nuclei (MNA), were strongly predictive, prognostic and cost effective markers, overriding both TNM stage and grade in NMIBC [7–9]. Currently, intravesical immunotherapy (BCG) is the standard treatment for intermediate or high-risk patients according to European Association of Urology guidelines [10]. Due to the intensive follow-up and treatment side effects, the costs per bladder cancer patient are high compared to other cancer types [11]. As such better predictive and prognostic markers than stage and grade are warranted [12].

Several publications have shown that an increased level of tumor infiltrating lymphocytes (TILs) is associated with a better overall survival [13] in patients with urothelial carcinomas [14]. More specifically, CD8+ cytotoxic T cells were related to a more favorable clinical outcome in both invasive bladder cancer [15] and many other tumor types. Also, the ratio of CD4 + and CD8+ TILs showed an altered pattern in recurrent and non-recurrent tumors in patients with NMIBC [16]. In addition, a strong association between increased numbers of regulatory T cells (Tregs) and bladder tumor recurrence, metastasis and stage has been reported [17]. Similarly, the ratio between tumor infiltrating effector T cells and Tregs was inversely related with tumor recurrence in invasive urothelial carcinomas [18].

Unfortunately, most of these studies have included small numbers of patients and mixed NMIBC and higher stage muscle invasive bladder cancer (MIBC) [16]. Furthermore, follow-up was often short, in spite of the fact that recurrences can also occur after many years. Another issue is that the selection of positive and negative cells were not random in the measurement procedure. This may have caused serious selection bias and erroneous results. At present, there are no reliable data regarding the significance of stromal and tumor infiltrating lymphocytes on prognosis in pTa pT1 urothelial carcinomas.

The aim of the present study is to investigate, whether CD8+, CD4+, CD25+ lymphocytes, and CD138+ plasma cells (immune cell markers) have additional prognostic value for recurrence and stage progression in a homogeneous cohort of pTa-pT1 tumors with long follow-up. We followed a fully randomized selection procedure for the measurement of Ki67, MNA and immune cell markers within the least differentiated area of the tumors. We hypothesize that adaptive immune cell composition in addition to proliferation features and MNA can have an additional value to predict, recurrence and stage progression.

Material and methods

This study was approved by the Norwegian Regional Ethics Committee (REK Vest, #106/09) before the start of the study. With approval from REK Vest, informed consent was not obtained as the tissue samples had already been removed for diagnostic and treatment purposes. In the period between January 1, 2002 and December 31, 2007, 249 patients were diagnosed with primary NMIBC, at the Department of Pathology, Stavanger University Hospital (SUH). Sixty-six cases were lost to follow-up or had inadequate sample quality for further analysis, leaving 183 patients to be included in the study (Table 1). Tissue samples were obtained by TURB or biopsies from the urinary bladder mucosa. After TURB, most patients underwent a single installation of the cytotoxic agent mitomycin C, while primary, high-risk patients (13%) classified according to national guidelines (bladdercalculator.no) were treated with BCG immunotherapy. All specimens were staged and graded by four experienced pathologists

Table 1. Exclusion criteria, number of excluded and included patients.

Primary pTaT1 urothelial carcinomas at SUH 2002–2006	**249**
Insufficient material	21
Thermal damage	11
Fragmented specimen	1
Necrotic specimen	2
Sarcomatoid differentiation	1
Previous urothelial carcinoma (on review of clinical notes)	1
cT3 or cT4 (on review of clinical notes)	3
pT2 at re-TURB	2
pT2 at review	1
Clinical metastasis at time of diagnosis	2
Lost to follow-up	11
Insufficient material for quantification of immune cell markers	2
Metastases at renal pelvis, ureter and urethra	8
Included in study	**183**

https://doi.org/10.1371/journal.pone.0233676.t001

(VK, OM, EG and JPAB) according to the WHO73 and WHO04 grading systems [19]. Tumor recurrence was defined as the presence of (a) tumor(s) in the bladder mucosa more than 3 months after the primary diagnosis. Total follow-up time, registered for statistical analyses of recurrence, was defined as the time from primary diagnosis until last control with cystoscopy. Stage progression was defined as recurrent tumor with pT2 or higher stage or confirmed metastasis, more than 3 months after primary diagnosis. Follow-up time registered for statistical analyses of stage progression was from primary diagnosis until death, or last known contact with the health care system. For stage progression, clinical follow-up was regarded as enough, but for recurrence, regular scheduled follow-up with cystoscopies according to guidelines was considered necessary. For the calculation of tumor recurrence only 177 patients were included as 6 more patients were lost to follow-up, as patients had to keep their urinary bladders and go through cystoscopy at least 3 months after primary diagnosis. Follow-up data were retrieved from medical hospital records, with last registration on June 30, 2016.

Immunohistochemistry (IHC)

TURB and biopsies were fixed in 10% neutral buffered Formalin®, dehydrated, and embedded in paraffin. Sections for assessment of MAI, MNA and histology were stained by Hematoxylin, Erythrosine & Saffron (HES). Adjacent to the HES stained sections, consecutive 4 μm paraffin sections were mounted onto Superfrost Plus® slides (Menzel, Braunschweig, Germany) and dried overnight at 37˚C followed by 1 h at 60˚C. Deparaffinization was performed stepwise by xylene, thereafter rehydration through decreasing concentrations of alcohol solutions. Heat-mediated antigen retrieval was performed with a computerized retrieval system (Immuno-Prep©; Instrumec, Oslo, Norway) using TRIS (10 mM) - EDTA (1 mM) antigen retrieval buffer (pH 9). The deparaffinized sections were first heated for 3 min at 110˚C and thereafter incubated for 10 min at 95˚C and finally cooled to 20˚C [20], in a pressure cooker. For the elimination of nonspecific background, a Tris-Buffered Saline Solution (DAKO, Glostrup, Denmark, S1968), containing 0.05% Tween 20, was used as a wash buffer (pH 7.6). Endogenous peroxidase activity was inactivated by the incubation of tissue sections in the peroxidase-blocking reagent (DAKO, Glostrup, Denmark; S2001) for 10 min. Immunostaining of CD4+, CD8+, CD25+ T lymphocyte subsets, CD138+ plasma cells and proliferation marker Ki67 was performed using an Autostainer (DAKO, Glostrup, Denmark). The tissue sections

were incubated with the monoclonal antibodies using the following dilutions: CD4 (Novocastra, Newcastle upon Tyne, UK; clone 1F6, 1:20); CD8 (DAKO, Glostrup, Denmark; clone C8/144B, 1:50); CD25 (Novocastra, Newcastle upon Tyne, UK; clone 4C9, 1:150); CD138 (Serotec, Kidlington, UK; clone B-B4, 1:200) and Ki67 (DAKO, Glostrup, Denmark; clone MIB-1, 1:100). An antibody diluent (DAKO, Glostrup, Denmark; S0809) was used for the preparation of primary and secondary antibody dilutions. The immune complex was visualized by peroxidase/DAB (DAKO, Glostrup, Denmark; EnVision Detection System, K5007) with incubation of Envision/HRP, rabbit/mouse antibody (ENV) for 30 min and DAB-chromogen with hematoxylin counterstain for 10 min. Thereafter the tissue sections were dehydrated and mounted [7, 21].

Quantitative image analysis

In each case the least differentiated area was carefully selected and demarcated on HES stained sections by a pathologist, based on the degree of cellular anaplasia. All assessments were done in this demarcated area. MAI, PPH3 and MNA were assessed as previously described [7–9]. Highly reproducible, semi-automated quantification of all immune cell markers and Ki67 was performed by using the motorized semi-automated QPRODIT (version 6.1) interactive image analysis system (Leica, Cambridge, UK). Immune cell quantification was performed at a final magnification of 400X using a 6-line grid in 150 random fields of vision (FOV) within the measurement area. In each FOV, the same endpoints of six electronic gridlines were used to register both immunohistochemically stained (IHC) positive-, IHC-negative immune cells and other cell types. Consequently, the sampling procedure within the measurement area was fully at random, which is essential for getting a well reproducible and prognostically accurate results (see Fig 1). Quantification and percentage calculation of Ki67-positivity was performed as earlier described [7–9]. The percentage of immune cell marker positivity was defined as: ((IHC positive immune cells) / (Total numbers of cells within the measurement area)) x 100. The average total counted cells within the measurement area was 210 cells (range 30-396). The average measurement area was 15 mm^2 (range 2-91).

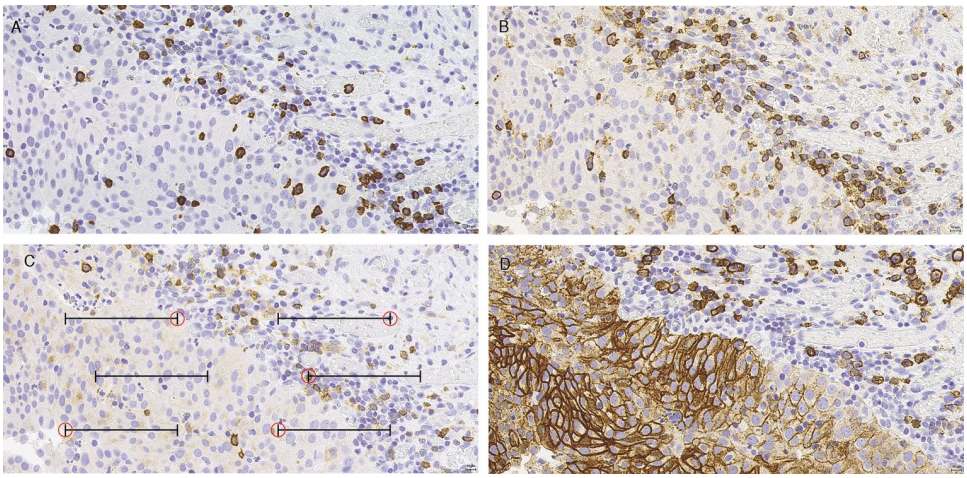

Fig 1. Immune cell markers CD8, CD4, CD25 and CD138 in representative urothelial bladder cancer tissue. (A) CD8 IHC stain (400X magnifications), (B) CD4 IHC stain (400X magnifications), (C) CD25 IHC stain (400X magnifications) (D) CD138 IHC stain (400X magnifications). In each field of vision positively stained immune cells were quantified as "positive counts" and negatively stained immune cells and other cell types were quantified as "negative counts". Counts were registered using six electronic grids with five endpoints. Scale bar 10 μm.

https://doi.org/10.1371/journal.pone.0233676.g001

Statistics

Statistical analysis was performed by using SPSS, version 21 (SPSS Inc., Chicago, IL, USA) and MedCalc Statistical Software version 19.1 (MedCalc Software BV, Ostend, Belgium; https://www.medcalc.org; 2019). The immunological markers, proliferation features and MNA were continuous variables, whilst stage progression, recurrence, grade and stage were categorical variables. Different percentiles (medians, tertiles and quartiles) and ROC curve analyses were used to determine the optimal prognostic thresholds of the continuous variables. Proliferation features and MNA were dichotomized by using previously published prognostic thresholds in both recurrence and stage progression cohorts [7–9]. In addition, proliferation features and MNA were dichotomized by using median values as well in the recurrence cohort. Univariate, nonparametric Kruskal-Wallis and Mann Whitney-U tests were performed to compare differences of continuous variables in the independent groups. A log rank test was run to determine if there were differences in the survival distributions of recurrence or stage progression for the two subgroups of immune cell markers, proliferation features and MNA or clinical and histopathological parameters. The differences between the subgroups were considered significant if the probability of no difference (p value) was <0.05. Univariate Cox proportional hazard ratios (HR) with 95% confidence intervals (CI) were also calculated. Multivariate Cox survival analysis was performed to evaluate the best prognostic combination of both continuous and categorical variables.

Results

The median follow-up time of the 177 patients available for recurrence analysis, was 82 months (range 3 to 165). From these, 105 patients (60%) experienced tumor recurrence. When analyzing for stage progression, the median follow-up time was 119 months (range 5 to 173). From these 183 patients, 13 patients (7%) experienced stage progression. In both groups, the gender distribution of the patients was 76% men and 24% women, and median age at first diagnosis was 74 years (range 39 to 95). According to the TNM-classification in both groups, 80% of the tumors presented as stage pTa and 61% were classified as WHO04 low-grade urothelial carcinoma. The distribution of WHO73 classification G1, G2 and G3 was 23%, 51% and 26% respectively.

Recurrence analysis

In total 173/177 and 150/177 patients were available for statistical analyses of Ki67 and Multifocality respectively. Out of all investigated immune cell markers, proliferation and nuclear features, clinical and histopathological parameters only Ki67 (threshold 39%, HR: 0.61, 95% CI, 0.4-0.9; $p = 0.05$) and Multifocality (HR: 1.8, 95% CI, 1.2-2.7; $p = 0.01$) showed significant association with tumor recurrence. Fig 2 shows the Kaplan-Meier curves for recurrence free survival for the two subgroups of Ki67. The group with low Ki67 had significantly shorter recurrence free survival, than the group with higher values. The presence of Multifocality correlated with a shorter recurrence free survival (Fig 3). There were no statistically significant differences between the median values of the immune cell markers in patients with or without recurrence. Median values and range as well threshold values; and hazard ratio, CIs, and p values for histopathological characteristics, proliferation features, MNA and immune cell markers were calculated by univariate recurrence free survival analyses summarized in Table 2.

Stage progression analysis

In our stage progression cohort, 183 NMIBC patients were available, 146 with pTa tumors and 37 with pT1 tumors. Interestingly, in pT1 tumors the percentages of CD25+, CD4+ and CD138+ cells were significantly higher than in pTa tumors ($p<0.001$, $p = 0.01$ and $p = 0.01$

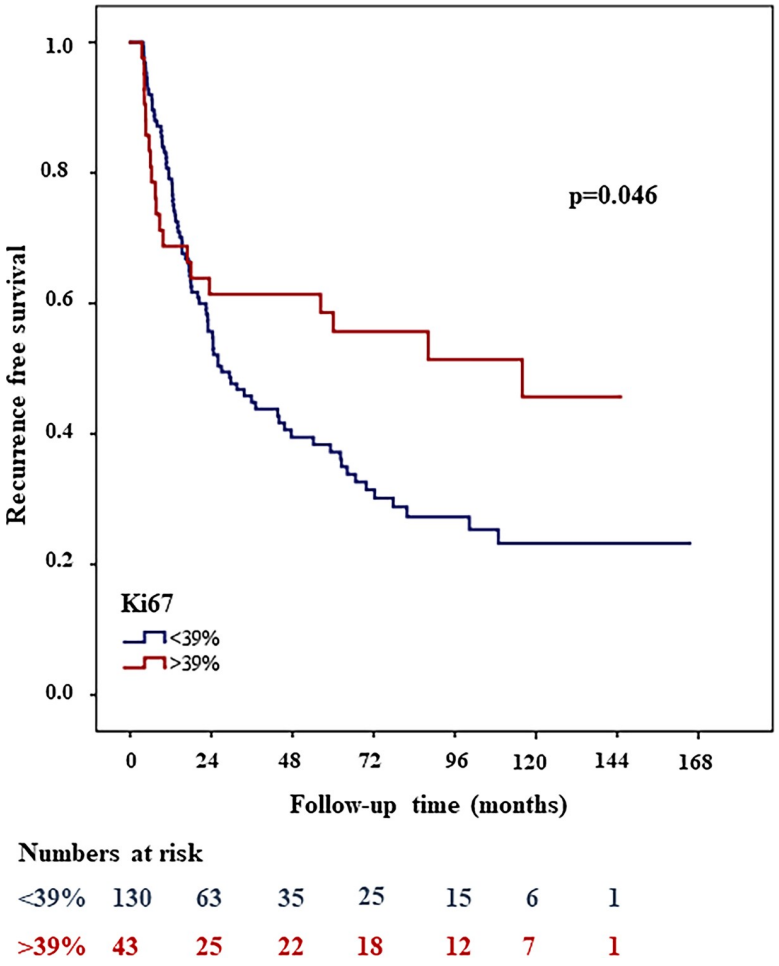

Fig 2. Low Ki67 (≤39%) associated with shorter recurrence free survival in Kaplan Meier survival analysis.

https://doi.org/10.1371/journal.pone.0233676.g002

respectively). When examining all 183 pTaT1 patients, out of all immune cell subsets only CD25+ showed significant association with stage progression (HR: 13.8, 95% CI, 1.8-106.2; $p = 0.001$) (Fig 4). Median values and range as well as threshold values; and hazard ratio, CIs, and p values for histopathological characteristics, proliferation features, MNA and immune cell markers were calculated by univariate recurrence free survival analyses summarized in Table 3. Patients with higher stage, grades and concomitant carcinoma in situ (CIS) had significantly higher stage progression risks ($p<0.05$).

Multivariate Cox proportional hazard analysis of all 183 patients, including age, WHO1973 and WHO2004 grade, KI67 (threshold≥18), PPH3, MAI and CD25+, showed that MAI (threshold >15) was the strongest single predictor for stage progression (HR: 8.6, 95% CI, 2.6-28.5; $p<0.001$). When the combination of MAI and CD25+ was also included, the MAI CD25 + combination was an even better predictor for stage progression (HR, 95% CI could not be computed, $p<0.001$). With Kaplan-Meier survival analyses, 26% of patients experienced stage progression in the high MAI high CD25+ group and 0% experienced stage progression in the low MAI low CD25+ group. In the mixed group (low MAI high CD25+, high MAI low CD25 +) 8% of patients progressed. Fig 5 shows the stage progression free survival curves for the two subgroups of MAI and Fig 6 shows the progression free survival curves for the two subgroups of MAI CD25+ combination.

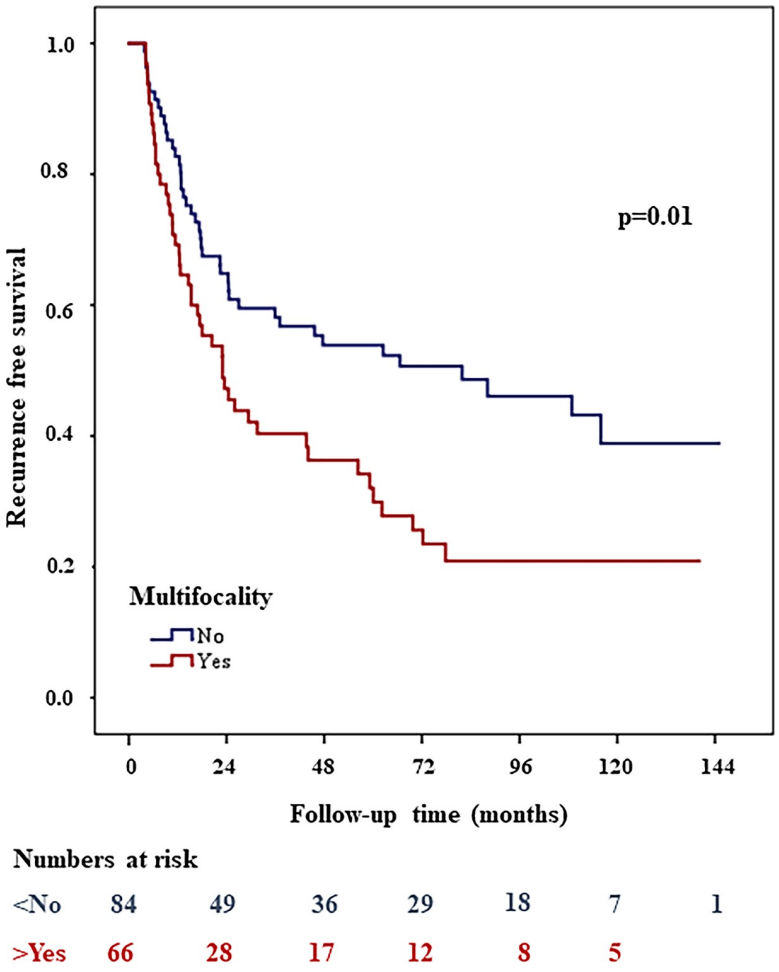

Fig 3. High CD25+ (>0.2%) associated with shorter stage progression free survival in Kaplan Meier survival analysis.

https://doi.org/10.1371/journal.pone.0233676.g003

Discussion

In the last decades, several studies have analyzed the association between different subgroups of lymphocytes and clinical outcome in bladder cancer. However, both concordant and conflicting results are observed when comparing our findings. One explanation might be the variation between the investigated areas within the bladder mucosa. Alternatively, numerous subsets of CD25+, CD4+ and CD8+ lymphocytes may coexist in the tumor microenvironment, which could differ in phenotypes, functions and locations in different time perspectives. One study demonstrated that increased numbers of CD4+ predict a lower 5-year overall survival (OS) in NMIBC [22]. On the other hand, it was also reported that increased numbers of CD4+ cells are related to a prolonged recurrence free survival in high-risk NMBIC [23]. Others published that CD8+ TILs are associated with better disease-free and overall survival in more advanced tumors [15, 24]. However Zhang *et al*. demonstrated that higher numbers of CD8 + TILs were related to a more unfavorable clinical outcome in pTa-pT2 (organ confined) tumors [25]. Regarding Tregs, one study reported that high FOXP3+/CD3+ and FOXP3 +/CD8+ cell ratios predict poorer overall survival in pT1-pT4 tumors [26]. Controversially, another study, observing pT1-pT4 tumors as well, demonstrated that high numbers of FOXP3 + lymphocytes were correlated with better survival [27].

Table 2. Univariate analyses for recurrence free survival of histopathological characteristics, immune cell markers, proliferation features and MNA.

Characteristics	Tumor recurrence cohort (105/177)			
	Event/at risk (%)	Log rank P value	HR	95% CI
Age at diagnosis				
<74	51/87 (59)	0.12	1.4	0.9-2.0
≥74	54/90 (60)			
(range 39-95)				
Gender				
Male	77/135 (57)	0.73	1.1	0.7-1.7
Female	28/42 (66)			
WHO1973 grade				
1	27/41 (66)	0.39		
2	50/90 (56)		0.8	0.5-1.2
3	28/46 (61)		1.0	0.6-1.7
WHO2004 grade				
Low	67/108 (62)	0.92	1.0	0.7-1.5
High	38/69 (55)			
Stage				
Ta	85/142 (60)	0.63	1.1	0.7-1.8
T1	20/35 (57)			
Multifocality				
No	42/84 (50)	**0.01**	1.8	1.2-2.7
Yes	47/66 (71)			
CIS				
No	92/156 (59)	0.63	1.2	0.6-2.1
Yes	13/21 (62)			
CD25+ (%)				
≤0.2	52/89 (58)	0.51	1.1	0.8-1.7
>0.2	53/88(60)			
(range 0-10)				
CD8+ (%)				
<3.0	53/89 (60)	0.74	1.1	0.7-1.6
≥3.0	52/88 (59)			
(range 0-28)				
CD4+ (%)				
<4.5	55/89 (62)	0.83	1.0	0.7-1.4
≥4.5	50/88 (57)			
(range 0–57)				
CD138+ (%)				
<1.4	53/89 (60)	0.62	1.1	0.8-1.6
≥1.4	52/88 (59)			
(range 0-20)				
Ki67 (%)				
≤39	83/130 (64)	**0.05**	0.6	0.4-0.9
>39	20/43 (47)			
(range 1-82)				
Ki67 (median)				

(*Continued*)

Table 2. (Continued)

Characteristics	Tumor recurrence cohort (105/177)			
	Event/at risk (%)	Log rank P value	HR	95% CI
<18	53/86 (62)	0.95	1.0	0.7-1.5
≥18	50/87 (57)			
(range 1-82)				
PPH3				
<40	79/131 (60)	0.52	0.86	0.5-1.4
≥40	25/44 (57)			
(range 0-137)				
PPH3 (median)				
<17	53/84 (63)	0.93	1.0	0.7-1.5
≥17	51/91 (56)			
(range 0-137)				
MAI				
≤15	80/138 (58)	0.56	1.1	0.7-1.8
>15	25/39 (64)			
(range 0-46)				
MAI (median)				
<4	51/83 (61)	0.86	1.0	0.7-1.5
≥4	54/94 (57)			
(range 0-46)				
MNA (mm^2)				
≤170	90/154 (58)	0.26	1.4	0.8-2.4
>170	15/23 (65)			
(range 49-488)				
MAI CD25+ (%)				
Low-Low*	45/77 (58)	0.6		
Mixed*	42/73 (57)		1.0	0.7-1.6
High-High*	18/27 (67)		1.3	0.8-2.3

HR Hazard Ratio, *CI* Confidence interval

**Low-Low*: low CD25 low MAI, *Mixed*: low CD25 high MAI and high CD25 low MAI, *High-High*: high CD25 high MAI

https://doi.org/10.1371/journal.pone.0233676.t002

An important issue when comparing these studies is the fact that many have included patients ranging from pTa-pT4. Furthermore, different methods and measurement areas have been used (some only intrastromal, others within the tumor urothelium as well) to quantitate the different subsets of lymphocytes. Moreover, non-random selection methods for the measurements have been described, which can be a serious cause for biased results. In the present study we have prevented these methodological challenges (consecutive sections, only pTa pT1 tumors, random selection methods for quantification) and analyzed the association between CD8+, CD4+, CD25+, CD138+, proliferation and nuclear features, histopathological and clinical parameters and outcomes (recurrence and stage progression free survival).

Doing so, the results for recurrence were different from those before. Low percentage of Ki67 was associated with a high risk for tumor recurrence. These results are unexpected and contrary to a recent meta-analysis, which showed that in 34 studies, high Ki67 was related to poor recurrence free survival [28]. One of the explanation could be, that in this meta-analysis

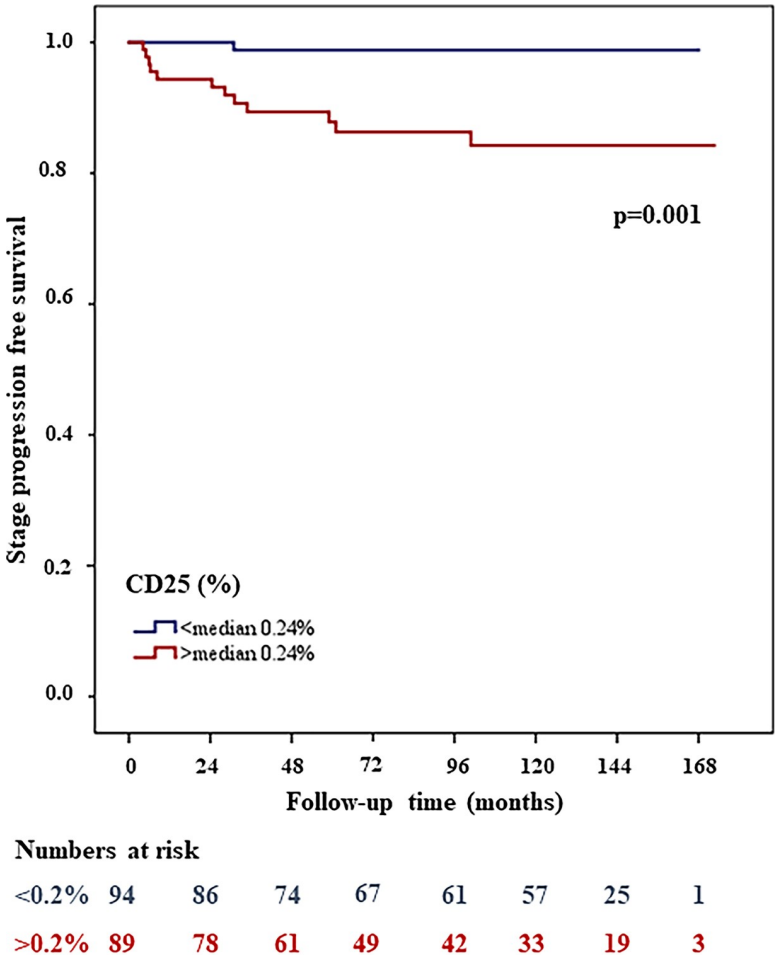

Fig 4. High MAI (>15) associated with shorter stage progression free survival in Kaplan Meier survival analysis.

https://doi.org/10.1371/journal.pone.0233676.g004

Table 3. Univariate analyses for stage progression free survival of clinical and histopathological characteristics, immune cell markers, proliferation features and MNA.

Characteristics	Stage Progression cohort (13/183)			
	Event/at risk (%)	Log rank P value	HR	95% CI
Age at diagnosis				
<74	3/90 (3)	**0.02**	4.0	1.1-14.7
≥74	10/93(11)			
(range 39–95)				
Gender				
Male	11/140 (8)	0.50	0.6	0.1-2.7
Female	2/43 (5)			
WHO1973 grade				
1	1/41 (2)	**0.04**		
2	5/94 (5)		2.3	0.3-19.3
3	7/48 (15)		6.8	0.8-55.6
WHO2004 grade				

(*Continued*)

Table 3. (Continued)

Characteristics	Stage Progression cohort (13/183)			
	Event/at risk (%)	Log rank P value	HR	95% CI
Low	4/113 (3)	**0.01**	4.1	1.3-13.2
High	9/70 (13)			
Stage				
Ta	4/146 (3)	<**0.001**	10.8	3.3-35.3
T1	9/37 (24)			
Multifocality				
No	3/87 (3)	0.10	3.0	0.8-11.5
Yes	7/68 (10)			
CIS				
No	9/162 (6)	**0.01**	4.3	1.3-14.0
Yes	4/21 (19)			
CD25+ (%)				
≤0.2	1/94 (1)	**0.001**	13.8	1.8-106.2
>0.2	12/89 (13)			
(range 0–10)				
CD8+ (%)				
<3.0	7/92 (8)	0.87	0.9	0.3-2.7
≥3.0	6/91 (7)			
(range 0–28)				
CD4+ (%)				
<4.5	5/92 (5)	0.31	1.8	0.6-5.4
≥4.5	8/91 (9)			
(range 0–57)				
CD138+ (%)				
<1.4	5/92 (5)	0.33	1.7	0.6-5.3
≥1.4	8/91 (9)			
(range 0–20)				
Ki67 (%)				
≤39	7/135 (5)	0.12	2.4	0.8-7.6
>39	5/44 (11)			
(range 1–82)				
Ki67 (median)				
<18	1/89 (1)	**0.003**	11.5	1.5-88.9
≥18	11/90 (12)			
(range 1–82)				
PPH3				
<40	4/136 (3)	<**0.001**	7.3	2.3-23.8
≥40	9/45 (20)			
(range 0–137)				
MAI				
≤15	5/143 (3)	<**0.001**	6.8	2.2-20.7
>15	8/40 (20)			
(range 0–46)				
MNA (mm2)				
≤170	10/159 (6)	**0.20**	2.3	0.6-8.2
>170	3/24 (12)			

(*Continued*)

Table 3. (Continued)

Characteristics	Stage Progression cohort (13/183)			
	Event/at risk (%)	Log rank P value	HR	95% CI
(range 49–488)				
MAI CD25+ (%)				
Low-Low*	0/81 (0)	**<0.001**		
Mixed*	6/75 (8)		-	-
High-High*	7/27 (26)		-	-

HR Hazard Ratio, *CI* Confidence interval

**Low-Low*: low CD25 low MAI, *Mixed*: low CD25 high MAI and high CD25 low MAI, *High-High*: high CD25 high MAI

https://doi.org/10.1371/journal.pone.0233676.t003

Ki67 threshold varied between 5% and 25%, in our study it was much higher, 39% (previously published prognostic threshold was used) [7]. In addition Ki67 positive cell quantification procedures differed from our highly reproducible semi-automated quantification method [7–9].

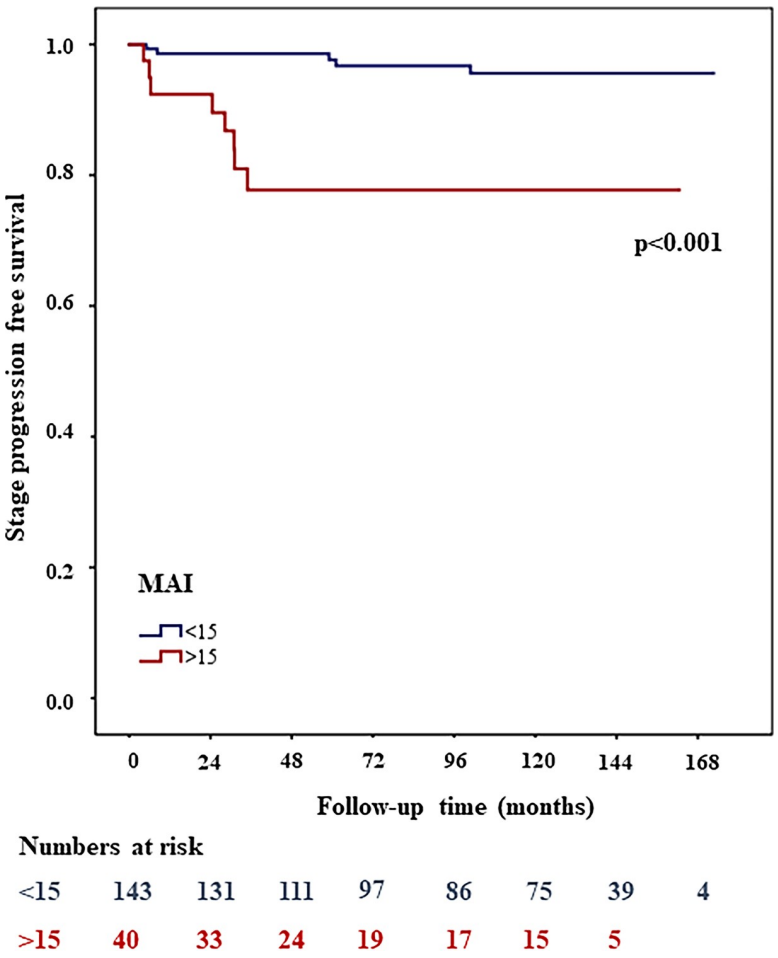

Fig 5. The combination of MAI and CD25+ stratifies patients into three groups. Patients with both low MAI and CD25+ values showed a better outcome than those with high amount of MAI and/or CD25+.

https://doi.org/10.1371/journal.pone.0233676.g005

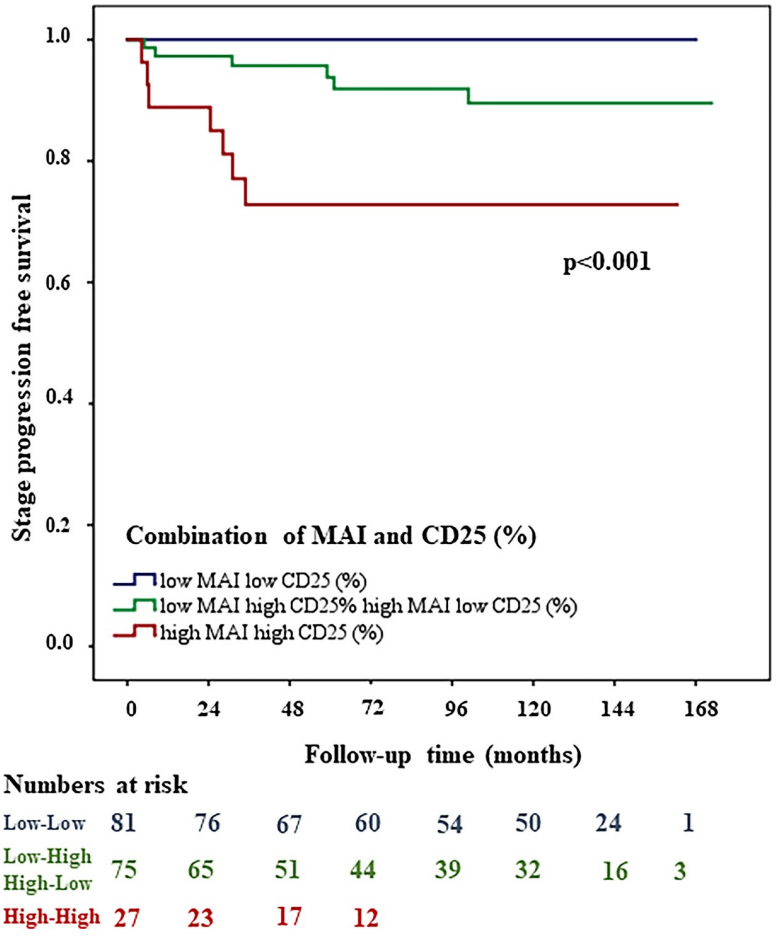

Fig 6.

https://doi.org/10.1371/journal.pone.0233676.g006

On the other hand the previously described threshold was calculated in correlation to progression, another threshold might be necessary for the prediction of recurrence. When using the median value for any of the proliferation markers no significant correlation was found with recurrence (Table 2). As such we need to validate the prognostic value of Ki67 in larger, independent, cohorts.

In our investigation we were not able to confirm a connection between the amount of CD8+, CD4+ and CD138+ cells, tumor recurrence or stage progression, nor the association between the amount of CD25+ cells and tumor recurrence in patients with NMIBC as described by other authors. Although our findings on CD4+ cells, are similar to the results of Kripna *et al*., who also reported no significant difference in the number of CD4+ cells between the recurrent and non-recurrent group in low-grade papillary urothelial carcinoma. However, they found an association between an increased number of CD8+ TILs and tumor recurrence [16]. Furthermore, Pichler *et al*. demonstrated that high FOXP3+CD25+ Tregs was associated with shorter recurrence free survival in NMIBC, which we could not confirm in our recurrence cohort [23].

As to stage progression prediction, we show that the number of CD25+ cells differed strongly between pTa and pT1, and is also significantly associated with stage progression. Loskog *et al*. previously published a similar finding that tumor infiltrating CD4+CD25+ T cells show a regulatory phenotype in human bladder cancer biopsies which was strongly associated with tumor progression [29]. Our results as well suggest a systemic suppression of immune

response by CD4+CD25+ Tregs in the bladder tumor tissue. Furthermore, the combination of MAI and CD25+ was the strongest predictor for tumor stage progression and strongly associated to TNM stages. The combination of MAI and CD25+ could define patient groups even better and in a more standardized and reproducible manner and can identify a large group of patients with a (nearly) 100% stage progression-free survival. Based on our current data and previously published data, we hypothesize that the CD25+ cells are Tregs, which increase in numbers following the development from superficial to more advanced stages, and as such tumors develop a gradually more immunosuppressive and more heterogeneous/proliferative tumor microenvironment.

A weakness of our study is, that in spite of the large number of patients and long follow-up, the number of patients with stage progression is still limited (n = 13). Mangrud *et al*. published, that the threshold estimation of Ki67 and other proliferation markers (MAI and PPH3) was sensitive to the number of patients [7]. Therefore external validation of our results is essential. Another issue is, that threshold values for Ki67 differed between previously published cohorts as well [28]. One of the explanations could be the lack of standardization of Ki67 antibodies, which makes the interpretation of true positive and negative cells difficult. In addition we used the same threshold values in both recurrence and stage progression cohorts. On the other hand Kaplan Meier survival plots and ROC curve analyses could not estimate an optimal threshold value in the recurrence cohort. Furthermore, tumor size of the patients (>3 cm) were not available in our retrospective cohort, which is an important factor regarding tumor recurrence. Although our quantification method for the immunohistochemical markers is highly reproducible, the method is very labor intensive. Therefore independent studies are needed to validate our results using more sophisticated and fully automated digital image analyses such as artificial intelligence. Bunimovich-Mendrazitsky recently published a mathematical dynamic model as a powerful tool, which could be used to analyze the interactions between stromal and tumor infiltrating lymphocytes and tumor cells [30] and to develop a standardized immunoscore [31] for predicting clinical outcome of patients with NMIBC.

Overall, the findings of our study show that a combination of MAI and CD25+ have overridingly strong prognostic value to predict stage progression and are worth validating in a well-defined, larger cohort.

Acknowledgments

We would like to thank Marit Nordhus for excellent technical assistance, Emma Rewcastle for the great discussions and proofreading the article. Furthermore, we are grateful for all the support from the Department of Pathology at the Stavanger University Hospital. Bianca van Diermen Hidle passed away before the submission of the final version of this manuscript. Melinda Lillesand accepts responsibility for the integrity and validity of the data collected and analyzed.

Author Contributions

Conceptualization: Melinda Lillesand, Vebjørn Kvikstad, Jan P. A. Baak, Emiel A. M. Janssen.

Data curation: Melinda Lillesand, Vebjørn Kvikstad, Ok Målfrid Mangrud, Einar Gudlaugsson, Bianca van Diermen-Hidle, Jan P. A. Baak.

Formal analysis: Melinda Lillesand.

Investigation: Melinda Lillesand, Vebjørn Kvikstad, Ok Målfrid Mangrud.

Methodology: Melinda Lillesand, Bianca van Diermen-Hidle, Ivar Skaland, Jan P. A. Baak, Emiel A. M. Janssen.

Project administration: Jan P. A. Baak, Emiel A. M. Janssen.

Resources: Melinda Lillesand, Vebjørn Kvikstad, Ok Målfrid Mangrud, Bianca van Diermen-Hidle, Ivar Skaland.

Supervision: Jan P. A. Baak, Emiel A. M. Janssen.

Validation: Melinda Lillesand, Vebjørn Kvikstad, Ok Målfrid Mangrud, Einar Gudlaugsson, Jan P. A. Baak.

Visualization: Melinda Lillesand.

Writing – original draft: Melinda Lillesand, Vebjørn Kvikstad.

Writing – review & editing: Ok Målfrid Mangrud, Einar Gudlaugsson, Ivar Skaland, Jan P. A. Baak, Emiel A. M. Janssen.

References

1. Miyazaki J, Nishiyama H. Epidemiology of urothelial carcinoma. Int J Urol. 2017; 24(10):730–4. https://doi.org/10.1111/iju.13376 PMID: 28543959
2. Grivas PD, Day M, Hussain M. Urothelial carcinomas: a focus on human epidermal receptors signaling. American journal of translational research. 2011; 3(4):362–73. PMID: 21904656
3. Cambier S, Sylvester RJ, Collette L, Gontero P, Brausi MA, van Andel G, et al. EORTC Nomograms and Risk Groups for Predicting Recurrence, Progression, and Disease-specific and Overall Survival in Non-Muscle-invasive Stage Ta-T1 Urothelial Bladder Cancer Patients Treated with 1–3 Years of Maintenance Bacillus Calmette-Guerin. Eur Urol. 2016; 69(1):60–9. https://doi.org/10.1016/j.eururo.2015.06.045 PMID: 26210894
4. van Kessel KEM, van der Keur KA, Dyrskjot L, Algaba F, Welvaart NYC, Beukers W, et al. Molecular Markers Increase Precision of the European Association of Urology Non-Muscle-Invasive Bladder Cancer Progression Risk Groups. Clin Cancer Res. 2018; 24(7):1586–93. https://doi.org/10.1158/1078-0432.CCR-17-2719 PMID: 29367430
5. Soukup V, Capoun O, Cohen D, Hernandez V, Babjuk M, Burger M, et al. Prognostic Performance and Reproducibility of the 1973 and 2004/2016 World Health Organization Grading Classification Systems in Non-muscle-invasive Bladder Cancer: A European Association of Urology Non-muscle Invasive Bladder Cancer Guidelines Panel Systematic Review. Eur Urol. 2017; 72(5):801–13. https://doi.org/10.1016/j.eururo.2017.04.015 PMID: 28457661
6. Kvikstad V, Mangrud OM, Gudlaugsson E, Dalen I, Espeland H, Baak JPA, et al. Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas. Diagn Pathol. 2019; 14(1):90. https://doi.org/10.1186/s13000-019-0868-3 PMID: 31412916
7. Mangrud OM, Gudlaugsson E, Skaland I, Tasdemir I, Dalen I, van Diermen B, et al. Prognostic comparison of proliferation markers and World Health Organization 1973/2004 grades in urothelial carcinomas of the urinary bladder. Hum Pathol. 2014; 45(7):1496–503. https://doi.org/10.1016/j.humpath.2014.03.001 PMID: 24796506
8. Bol MG, Baak JP, de Bruin PC, Rep S, Marx W, Bos S, et al. Improved objectivity of grading of T(A,1) transitional cell carcinomas of the urinary bladder by quantitative nuclear and proliferation related features. J Clin Pathol. 2001; 54(11):854–9. https://doi.org/10.1136/jcp.54.11.854 PMID: 11684720
9. Bol MG, Baak JP, Rep S, Marx WL, Kruse AJ, Bos SD, et al. Prognostic value of proliferative activity and nuclear morphometry for progression in TaT1 urothelial cell carcinomas of the urinary bladder. Urology. 2002; 60(6):1124–30. https://doi.org/10.1016/s0090-4295(02)01906-4 PMID: 12475695
10. Zuiverloon TC, Nieuweboer AJ, Vekony H, Kirkels WJ, Bangma CH, Zwarthoff EC. Markers predicting response to bacillus Calmette-Guerin immunotherapy in high-risk bladder cancer patients: a systematic review. Eur Urol. 2012; 61(1):128–45. https://doi.org/10.1016/j.eururo.2011.09.026 PMID: 22000498
11. Decaestecker K, Oosterlinck W. Managing the adverse events of intravesical bacillus Calmette-Guerin therapy. Research and reports in urology. 2015; 7:157–63. https://doi.org/10.2147/RRU.S63448 PMID: 26605208
12. Sievert KD, Amend B, Nagele U, Schilling D, Bedke J, Horstmann M, et al. Economic aspects of bladder cancer: what are the benefits and costs? World J Urol. 2009; 27(3):295–300. https://doi.org/10.1007/s00345-009-0395-z PMID: 19271220

13. Huang HS, Su HY, Li PH, Chiang PH, Huang CH, Chen CH, et al. Prognostic impact of tumor infiltrating lymphocytes on patients with metastatic urothelial carcinoma receiving platinum based chemotherapy. Sci Rep. 2018; 8(1):7485. https://doi.org/10.1038/s41598-018-25944-1 PMID: 29748589

14. Wang B, Wu S, Zeng H, Liu Z, Dong W, He W, et al. CD103+ Tumor Infiltrating Lymphocytes Predict a Favorable Prognosis in Urothelial Cell Carcinoma of the Bladder. J Urol. 2015; 194(2):556–62. https://doi.org/10.1016/j.juro.2015.02.2941 PMID: 25752441

15. Sharma P, Shen Y, Wen S, Yamada S, Jungbluth AA, Gnjatic S, et al. CD8 tumor-infiltrating lymphocytes are predictive of survival in muscle-invasive urothelial carcinoma. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(10):3967–72. https://doi.org/10.1073/pnas.0611618104 PMID: 17360461

16. Krpina K, Babarovic E, Dordevic G, Fuckar Z, Jonjic N. The association between the recurrence of solitary non-muscle invasive bladder cancer and tumor infiltrating lymphocytes. Croat Med J. 2012; 53 (6):598–604. https://doi.org/10.3325/cmj.2012.53.598 PMID: 23275325

17. Zhu X, Ma LL, Ye T. Expression of CD4(+)CD25(high)CD127(low/-) regulatory T cells in transitional cell carcinoma patients and its significance. J Clin Lab Anal. 2009; 23(4):197–201. https://doi.org/10.1002/jcla.20331 PMID: 19623656

18. Parodi A, Traverso P, Kalli F, Conteduca G, Tardito S, Curto M, et al. Residual tumor micro-foci and overwhelming regulatory T lymphocyte infiltration are the causes of bladder cancer recurrence. Oncotarget. 2016; 7(6):6424–35. https://doi.org/10.18632/oncotarget.7024 PMID: 26824503

19. Mangrud OM, Waalen R, Gudlaugsson E, Dalen I, Tasdemir I, Janssen EA, et al. Reproducibility and prognostic value of WHO1973 and WHO2004 grading systems in TaT1 urothelial carcinoma of the urinary bladder. PLoS One. 2014; 9(1):e83192. https://doi.org/10.1371/journal.pone.0083192 PMID: 24409280

20. Skaland I, Janssen EA, Gudlaugsson E, Klos J, Kjellevold KH, Soiland H, et al. Validating the prognostic value of proliferation measured by Phosphohistone H3 (PPH3) in invasive lymph node-negative breast cancer patients less than 71 years of age. Breast Cancer Res Treat. 2009; 114(1):39–45. https://doi.org/10.1007/s10549-008-9980-x PMID: 18373192

21. Ovestad IT, Gudlaugsson E, Skaland I, Malpica A, Kruse AJ, Janssen EA, et al. Local immune response in the microenvironment of CIN2-3 with and without spontaneous regression. Mod Pathol. 2010; 23(9):1231–40. https://doi.org/10.1038/modpathol.2010.109 PMID: 20512116

22. Zhang Q, Hao C, Cheng G, Wang L, Wang X, Li C, et al. High CD4(+) T cell density is associated with poor prognosis in patients with non-muscle-invasive bladder cancer. Int J Clin Exp Pathol. 2015; 8 (9):11510–6. PMID: 26617883

23. Pichler R, Fritz J, Zavadil C, Schafer G, Culig Z, Brunner A. Tumor-infiltrating immune cell subpopulations influence the oncologic outcome after intravesical Bacillus Calmette-Guerin therapy in bladder cancer. Oncotarget. 2016; 7(26):39916–30. https://doi.org/10.18632/oncotarget.9537 PMID: 27221038

24. Faraj SF, Munari E, Guner G, Taube J, Anders R, Hicks J, et al. Assessment of tumoral PD-L1 expression and intratumoral CD8+ T cells in urothelial carcinoma. Urology. 2015; 85(3):703 e1–6.

25. Zhang S, Wang J, Zhang X, Zhou F. Tumor-infiltrating CD8+ lymphocytes predict different clinical outcomes in organ- and non-organ-confined urothelial carcinoma of the bladder following radical cystectomy. PeerJ. 2017; 5:e3921. https://doi.org/10.7717/peerj.3921 PMID: 29043112

26. Horn T, Laus J, Seitz AK, Maurer T, Schmid SC, Wolf P, et al. The prognostic effect of tumour-infiltrating lymphocytic subpopulations in bladder cancer. World J Urol. 2016; 34(2):181–7. https://doi.org/10.1007/s00345-015-1615-3 PMID: 26055646

27. Winerdal ME, Marits P, Winerdal M, Hasan M, Rosenblatt R, Tolf A, et al. FOXP3 and survival in urinary bladder cancer. BJU Int. 2011; 108(10):1672–8. https://doi.org/10.1111/j.1464-410X.2010.10020.x PMID: 21244603

28. Ko K, Jeong CW, Kwak C, Kim HH, Ku JH. Significance of Ki-67 in non-muscle invasive bladder cancer patients: a systematic review and meta-analysis. Oncotarget. 2017; 8(59):100614–30. https://doi.org/10.18632/oncotarget.21899 PMID: 29246006

29. Loskog A, Ninalga C, Paul-Wetterberg G, de la Torre M, Malmstrom PU, Totterman TH. Human bladder carcinoma is dominated by T-regulatory cells and Th1 inhibitory cytokines. J Urol. 2007; 177(1):353–8. https://doi.org/10.1016/j.juro.2006.08.078 PMID: 17162090

30. Bunimovich-Mendrazitsky S, Claude Gluckman J, Chaskalovic J. A mathematical model of combined bacillus Calmette-Guerin (BCG) and interleukin (IL)-2 immunotherapy of superficial bladder cancer. J Theor Biol. 2011; 277(1):27–40. https://doi.org/10.1016/j.jtbi.2011.02.008 PMID: 21334346

31. Taube JM, Galon J, Sholl LM, Rodig SJ, Cottrell TR, Giraldo NA, et al. Implications of the tumor immune microenvironment for staging and therapeutics. Mod Pathol. 2018; 31(2):214–34. https://doi.org/10.1038/modpathol.2017.156 PMID: 29192647

Paper III

THE JOURNAL of UROLOGY®
www.auajournals.org/journal/juro

OPEN

T1 Substaging of Nonmuscle Invasive Bladder Cancer is Associated with bacillus Calmette-Guérin Failure and Improves Patient Stratification at Diagnosis

Florus C. de Jong, Robert F. Hoedemaeker, Vebjørn Kvikstad, Jolien T. M. Mensink, Joep J. de Jong, Egbert R. Boevé, Deric K. E. van der Schoot, Ellen C. Zwarthoff, Joost L. Boormans and Tahlita C. M. Zuiverloon*

From Department of Urology (FCdJ, JJdJ, JLB, TCMZ), Erasmus MC Cancer Institute, Rotterdam, The Netherlands, Pathan BV (RFH), Pathological Laboratory, Rotterdam, The Netherlands, Department of Pathology (VK), Stavanger University Hospital, Stavanger, Norway, Department of Mathematics and Natural Science (VK), University of Stavanger, Stavanger, Norway, Department of Pathology (JTMM, ECZ), Erasmus MC Cancer Institute, Rotterdam, The Netherlands, Department of Urology (ERB), Franciscus Gasthuis & Vlietland, Rotterdam, The Netherlands, Department of Urology (DKEvdS), Amphia, Breda, The Netherlands

Purpose: Currently, markers are lacking that can identify patients with high risk nonmuscle invasive bladder cancer who will fail bacillus Calmette-Guérin treatment. Therefore, we evaluated the prognostic value of T1 substaging in patients with primary high risk nonmuscle invasive bladder cancer.

Materials and Methods: Patients with primary high risk nonmuscle invasive bladder cancer who received ≥5 bacillus Calmette-Guérin induction instillations were included. All tumors were centrally reviewed, which included T1 substaging (microinvasion vs extensive invasion of the lamina propria). T1 patients were stratified into high risk or highest risk subgroups according to major urology guidelines. Primary end point was bacillus Calmette-Guérin failure, defined as development of a high grade recurrence. Secondary end points were high grade recurrence-free survival, defined as time from primary diagnosis to biopsy-proven high grade recurrence and progression-free survival. Time-to-event analyses were used to predict survival.

Results: A total of 264 patients with high risk nonmuscle invasive bladder cancer had tumor invasion of the lamina propria, of which 73% were classified as extensive invasion and 27% as microinvasion. Median followup was 68 months (IQR 43–98) and bacillus Calmette-Guérin failure was more common among patients with extensive vs microinvasive tumors (41% vs 21%, p=0.002). The 3-year high grade recurrence-free survival (defined as bacillus Calmette-Guerin failure) for patients with extensive vs microinvasive tumors was 64% vs 83% (p=0.004). In multivariate analysis, T1 substaging was an independent predictor of high grade recurrence-free survival (HR 3.2, p=0.005) and progression-free survival (HR 3.0, p=0.009). Patients with highest risk/microinvasive disease have an improved progression-free survival as compared to highest risk/T1e disease (p.adj=0.038).

Conclusions: T1 substaging provides important prognostic information on patients with primary high risk nonmuscle invasive bladder cancer treated with

Abbreviations and Acronyms

BCG = bacillus Calmette-Guérin
CIS = carcinoma in situ
DSS = disease-specific survival
HG = high grade
HG-RFS = high grade recurrence-free survival defined as BCG failure
HPF = (microscopic) high-powered field (objective 40×)
HR-NMIBC = high risk nonmuscle invasive bladder cancer
LVI = lymphovascular invasion
MIBC = muscle invasive bladder cancer
MM-VP = muscularis mucosae-vascular plexus
p.adj = adjusted p value
PFS = progression-free survival
RC = radical cystectomy
re-TURBT = repeated transurethral resection of bladder tumor
T1e = T1 extensive invasion of the lamina propria
T1m = T1 microinvasion of the lamina propria
TaHG = pTa high grade
TILs = tumor infiltrating lymphocytes
VH = variant histology

Accepted for publication September 7, 2020.
Funding was obtained from Erasmus MC Medical Research Advisory Committee Erasmus (MRACE) Grant 107477. MRACE has no role whatsoever in design, collection, management, analysis, interpretation, preparation, review or approval of this manuscript.
This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND), which permits downloading and sharing the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.
* Correspondence: Erasmus MC Cancer Institute, Erasmus University Medical Center, Dr. Molewaterplein 40, Room Be-304, 3015 GD, Rotterdam, The Netherlands (telephone: +31-107043059; FAX: +31-107044762; email: t.zuiverloon@erasmusmc.nl).

0022-5347/21/2053-0701/0
THE JOURNAL OF UROLOGY®
© 2021 The Author(s). Published on behalf of the American Urological Association, Education and Research, Inc.
https://doi.org/10.1097/JU.0000000000001422
Vol. 205, 701-708, March 2021
Printed in U.S.A.

bacillus Calmette-Guérin. The risk of bacillus Calmette-Guérin failure is higher in extensive vs microinvasive tumors. Substaging of T1 high risk nonmuscle invasive bladder cancer has the potential to guide treatment decisions on bacillus Calmette-Guérin vs alternative strategies at diagnosis.

Key Words: BCG vaccine; urinary bladder neoplasms; prognosis; carcinoma, transitional cell; neoplasm staging

NONMUSCLE invasive bladder cancer accounts for ~75% of newly diagnosed bladder cancer cases.[1] In case of lamina propria (T1) invasion, patients are at high risk for recurrent and progressive disease.[2,3] Patients with T1 high risk nonmuscle invasive bladder cancer are treated with transurethral resection of the bladder tumor and adjuvant intravesical bacillus Calmette-Guérin induction instillations.[4] In 30%-50% of patients with HR-NMIBC, BCG therapy fails and these patients develop high grade recurrences or progression to muscle invasive bladder cancer. In case of progression, neoadjuvant chemotherapy followed by radical cystectomy is the standard of care.[4] Despite treatment, 50%-70% of progressed patients die within 5 years after diagnosis.[5]

No markers are available to predict which patients will benefit from BCG treatment.[6] Repeated BCG instillations in nonresponders cause a delay in RC and a recent study showed that progression to MIBC is associated with worse overall survival compared to patients with primary MIBC at diagnosis.[5] Furthermore, the ongoing global BCG shortage demands selective use of limited resources. Thus, there is a clinical need for markers to identify patients who will benefit from BCG and patients who should receive other treatments.[7] To improve patient stratification, guidelines use the presence of aggressive clinicopathological features to identify a subgroup of HR-NMIBC patients at the highest risk of progression.[4,8] For these patients, both American and European guidelines strongly recommend to consider an immediate RC.[4,8] Despite this substratification of HR-NMIBC patients at the highest risk of progression, performing immediate RC in all of these patients results in overtreatment.[2,3]

Over the years, T1 substaging has been investigated as a prognostic tool in HR-NMIBC. Several methods have been described to assess depth and extent of tumor invasion into the lamina propria, which was associated with an increased risk of progression and death.[9–15] Some evidence showed that deeper invasion was also associated with recurrent disease.[16,17] T1 substaging is recommended for pathologists since the 2016 WHO classification.[18] However, the most optimal T1 substaging system remains to be defined.[19,20] T1 substaging by evaluation of muscularis mucosae-vascular plexus (T1a/b MM-VP) invasion is challenging due to difficult assessment of the MM-VP and T1 metric substaging by (optical) micrometers is impractical and time-consuming.[12,16] T1 microinvasive vs extensive substaging, in which tumor invasion should not exceed 1 HPF, is easy to use and proved more accurate than MM-VP substaging in earlier studies.[9,21]

Currently, it is unknown if T1 substaging is associated with treatment response in HR-NMIBC. Furthermore, it is unclear if T1 substaging has the potential to guide treatment decisions. Here, we investigated whether T1 HPF substaging was associated with BCG failure and if this substaging method can be used to improve patient stratification at diagnosis.

MATERIALS AND METHODS

Patients and Pathology

All patients with a primary diagnosis of HR-NMIBC (Tis or Ta/T1HG urothelial carcinoma), who underwent transurethral resection of the bladder tumor with or without re-TURBT and who received ≥5/6 BCG induction instillations were retrospectively included at 3 Dutch (Erasmus MC; Franciscus Gasthuis & Vlietland and Amphia) and 1 Norwegian hospital (Stavanger University Hospital) from 2000–2017. Additional information on patient inclusion can be found in the supplementary methods (https://www.jurology.com). The study was approved by the Erasmus MC Medical Ethics Committee (MEC-2018-1097). Hematoxylin and eosin (HE) slides from all primary tumors, re-TURBTs and recurrent tumors were centrally reviewed by a uropathologist who was blinded for clinical information. Assessment included T stage, tumor grade (WHO 1973/2016), presence of concomitant carcinoma in situ, T1 HPF substaging, lymphovascular invasion, variant histology, tumor infiltrating lymphocytes (TILs) and tumor necrosis (TN). T stage, tumor grade, CIS, LVI, VH and TN were scored according to standard WHO criteria.[18] T1 HPF substaging was performed as described previously (in this manuscript referred to as T1 substaging).[9] Briefly, if a single focus of lamina propria invasion with a maximum diameter of 0.5mm (ie 1 HPF, objective 40×) was observed, the tumor was defined as T1m. If tumor invasion was >0.5 mm or when more than 1 invasive focus was observed, the tumor was defined as T1e. TILs were scored as either absent/sparse vs marked within the tumor area.[22] Patients for whom T1 disease was confirmed in either the primary or re-TURBT specimen, and with identifiable detrusor, were included in the

RIGHTSLINK

Table 1. *Baseline study characteristics of 264 patients with primary T1 high risk nonmuscle invasive bladder cancer*

	No. (%)
Age at diagnosis (yrs):	
Median (IQR)	71 (66—77)
Gender:	
Male	215 (81)
Female	49 (19)
Substaging:	
T1 microinvasion	72 (27)
T1 extended invasion	192 (73)
Tumor grade (WHO 1973):	
2	6 (2)
3	258 (98)
Smoking:	
No	85 (32)
Yes/stopped	165 (63)
Missing	14 (5)
Concomitant CIS:	
No	205 (78)
Yes	59 (22)
Tumor focality:	
Unifocal	129 (49)
Multifocal	132 (50)
Missing	3 (1)
Tumor size (cm):	
<3	51 (19)
≥3	43 (16)
Missing	170 (65)
Tumor infiltrating lymphocytes:	
No	74 (28)
Yes	190 (72)
Tumor necrosis:	
No	244 (92)
Yes	20 (8)
Lymphovascular invasion:	
No	249 (94)
Yes	15 (6)
Variant histology:	
No	216 (82)
Diffuse	21 (8)
Micropapillary	12 (4)
Glandular	8 (3)
Squamous	4 (1.5)
Neuroendocrine	1 (0.5)
Sarcamatoid	1 (0.5)
Other	1 (0.5)
Re-TURBT performed:	
No	51 (19)
Yes	213 (81)
Risk classification at start of BCG:	
High risk	90 (34)
Highest risk	174 (66)
Adequate BCG:*	
No	27 (10)
Yes	237 (90)
Median BCG maintenance instillations (IQR)	12 (6—18)
BCG maintenance completed:	
1 Yr (≥3 cycles)	173 (66)
3 Yrs (≥9 cycles)	52 (20)
Median total BCG instillations (IQR)	18 (12—24)
BCG failure:†	
No	171 (65)
Yes	93 (35)
BCG failure (characteristics):	
1. MIBC as first recurrence	20 (8)
2. T1, grade 3/HG after BCG induction	19 (7)
3. HG recurrence after adequate BCG	54 (20)
Progression (MIBC, lymph node disease and metastases):	
No	201 (76)
Yes	63 (24)
Median time to progression (IQR)	18 (10—48)
Lymph node metastases:	
N0	247 (93)
N1-3	17 (6)
Distant metastases:	
M0	240 (91)
M1	24 (9)
Death from bladder cancer at last followup	41 (15)
Death of other cause:	67 (26)
Unknown	6 (2)
Alive	150 (57)
Median total mos followup (IQR)	68 (43—98)
Median followup BCG responders (IQR)	71 (55—99)
Median mos time to BCG failure (IQR)	7 (5—16)

Data in table 1 are also summarized in table 2, stratified by T1 substaging.
* Defined as ≥5/6 inductions + ≥2/3 maintenance instillations.
† Specified by major urology guidelines, which include patients with muscle invasive recurrences, T1HG after BCG induction and high grade recurrences after adequate BCG therapy.

analyses. After review, patients were stratified into a high risk or highest risk subgroup, according to the AUA/SUO nonmuscle invasive bladder cancer algorithm and EAU risk stratification.[4,8] Highest risk clinicopathological features were: T1G3/HG with concomitant CIS, lymphovascular invasion or VH, T1G3/HG with prostatic urethra involvement or multifocal and/or large (≥3 cm) T1G3/HG.[4,8]

Definitions, End Points and Statistics

Primary end point was BCG failure. BCG failure was defined as biopsy proven T1HG disease after ≥5 BCG induction instillations, HG disease after adequate BCG therapy or recurring muscle invasive disease.[4,8] Adequate BCG consists of ≥5/6 BCG induction instillations plus ≥2/3 BCG maintenance instillations.[23] Secondary end points were 3-year HG recurrence-free survival, progression-free survival and disease-specific survival. Time-to-recurrence was defined as the moment from primary T1 diagnosis until a biopsy-proven HG recurrence occurred (BCG failure). Three-year HG-RFS was selected because duration of the BCG regimen is 3 years.[4] Patients with only HG cytology or low grade (LG) biopsy recurrences were not considered BCG failures.[4,23] Further details on definitions and statistical analyses can be found in the supplementary methods (https://www.jurology.com).

RESULTS

Study Population

The study population consisted of 535 primary HR-NMIBC patients who received ≥5 BCG induction instillations. After pathology review, 26 cases were excluded because of the following reasons: up staging to muscle invasion in 4, downgrading to G2/LG in 5, and 17 cases had degraded hematoxylin and eosin slides and tissue blocks. Of the remaining 509 HR-NMIBC patients, 264 were included based on the presence of lamina propria invasion (T1) in the primary and/or re-TURBT specimen. Clinicopathological

RIGHTSLINK

Table 2. *Main study variables and outcome parameters stratified for T1 substaging in 72 patients with T1m vs 192 with T1e disease*

	T1m	T1e
Median yrs age at diagnosis (IQR)	73 (64–80)	70 (63–75)
No. female gender (%)	21 (29)	28 (15)
No. active smoker (%)	15 (20)	62 (33)
No. re-TURBT (%)	57 (79)	156 (81)
No. WHO grade 3 (%)	69 (96)	189 (98)
No. CIS (%)	11 (15)	48 (25)
No. multifocal (%)	34 (47)	98 (51)
No. variant histology (%)	3 (4)	45 (23)
No. TILs (%)	30 (42)	160 (83)
No. LVI (%)	2 (3)	13 (7)
No. tumor necrosis (%)	4 (6)	16 (8)
No. BCG failure (%)	15 (21)	78 (41)
No. progression (%)	9 (13)	54 (28)
No. death from BC (%)	6 (8)	35 (18)

characteristics of patients with T1 disease at diagnosis is depicted in table 1. Median age was 67 years (IQR 62–71), 81% of patients were male, and the median time from T1 diagnosis to BCG induction was 8 weeks. Median followup for the entire cohort was 70 months (IQR 48–90).

T1 Substaging is Associated with BCG Failure

T1 substaging was assessed in all tumors; T1m was present in 72 (27%) and T1e in 192 (73%). An overview of main study variables and outcome parameters according to T1 substaging is included in table 2. Patients with T1m vs T1e disease underwent the same number of re-TURBTs (79% vs 81%, p=0.731). Adequate BCG was administered in 237/264 (90%) patients. Reasons for not having received adequate BCG were incorrect planning/BCG intolerance in 7/264 (3%) and discontinuation of BCG treatment due to persistent ≥T1 disease following BCG induction in 20/264 (8%). Of the patients 93/321 developed BCG failure; 20 had a muscle invasive recurrence, 19 had recurrent T1HG disease after BCG induction and 54 patients had a HG recurrence after adequate BCG. The median time to BCG failure was 7 months (IQR 5–16 months). BCG failure occurred more often in patients with T1e HR-NMIBC than in patients with T1m disease: 41% vs 21%, p=0.002. Furthermore, patients with T1e disease had significantly worse 3-year HG-RFS than patients with T1m tumors: 64% vs 83% (p=0.004), worse PFS (p=0.014), but not DSS (p=0.08) (fig. 1, *A* to *C*). In multivariate analysis, T1 substaging was an independent predictor of HG-RFS (HR 3.2, p=0.005), PFS (HR 3.0, p=0.009) and DSS (HR 3.1, p=0.031) (HG-RFS in table 3; PFS and DSS in supplementary table 1, https://www.jurology.com).

Understaging of T1 HR-NMIBC may occur in case a re-TURBT is not performed. Hence, to exclude the possibility that understaging of T1 disease could have caused our observed association between T1e substaging and poor clinical outcome, the analyses were repeated in 213/264 patients who also received a re-TURBT. BCG failure occurred more often in patients with T1e vs T1m disease (39% vs 18%, p=0.005). In addition, the 3-year HG-RFS (p=0.011) and PFS (p=0.028) remained worse in patients with T1e tumors, in contrast to the DSS (p=0.12) (supplementary fig. 1, *A* to *C*, https://www.jurology.com). In multivariate analysis, T1 substaging remained an independent predictor of HG-RFS (HR 3.2, p=0.016) and PFS (HR 2.9, p=0.025), yet not for

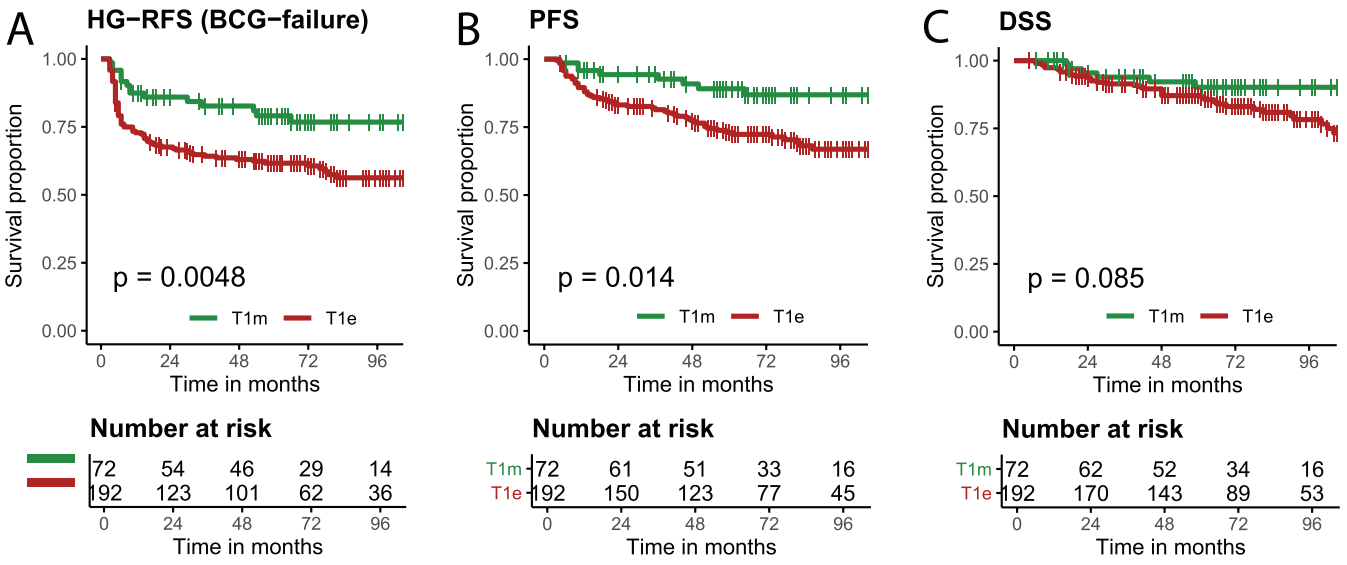

Figure 1. Kaplan-Meier estimates of clinical outcome in 264 patients with primary T1 high risk nonmuscle invasive bladder cancer stratified by T1 high-powered field substaging and with visible detrusor in specimen. *A*, HG-RFS (BCG failure). *B*, PFS. *C*, DSS. p Value is determined by log-rank test.

RIGHTSLINK

Table 3. *Univariate and multivariate Cox proportional hazard analyses of high grade recurrence-free survival in 264 primary T1 high risk nonmuscle invasive bladder cancer patients*

	HG-RFS Univariate		HG-RFS Multivariate	
	HR (95% CI)	p Value	HR (95% CI)	p Value
Age at diagnosis	1.0 (0.99–1.0)	0.255	1.0 (1.0–1.1)	0.047
Female gender	0.7 (0.4–1.3)	0.229	0.7 (0.3–1.6)	0.401
Smoking (active)	1.1 (0.7–1.7)	0.645	0.6 (0.3–1.1)	0.095
Pos re-TURBT	0.7 (0.4–1.0)	0.074	1.0 (0.5–2.1)	0.977
Substage T1e	2.2 (1.3–3.8)	0.006	3.2 (1.4–7.3)	0.005
Grade 3	2.3 (0.3–16)	0.409	0.7 (0.1–5.9)	0.803
Pos CIS	1.7 (1.1–2.7)	0.015	1.8 (0.96–3.4)	0.068
Size ≥3 cm	1.2 (0.6–2.4)	0.608	-	-
Multifocal	1.9 (1.3–2.9)	0.003	1.8 (0.98–3.3)	0.060
Pos variant histology	1.1 (0.7–1.9)	0.728	0.5 (0.2–1.2)	0.110
Pos TILs	1.1 (0.7–1.7)	0.727	0.8 (0.4–1.5)	0.438
Pos LVI	2.5 (1.3–4.9)	0.006	4.4 (2.1–9.4)	<0.001
Pos tumor necrosis	1.0 (0.5–2.2)	0.951	1.8 (0.7–4.5)	0.223

DSS (HR 2.8, p=0.08) (supplementary table 2, *A* and *B*, https://www.jurology.com). To confirm that the difference between T1m and T1e disease is not due to up staging of Ta to T1m disease at central review, we analyzed T1m/T1e vs TaHG disease. In total, 37 patients were down staged (T1 >Ta) and 12 patients were up staged (Ta >T1). Patients with Ta disease had a similar HG-RFS (p=0.592), PFS (p=0.828) and DSS (p=0.798) as patients with T1m disease (supplementary fig. 2, *A* to *C*, https://www.jurology.com). Importantly, patients with Ta disease had a favorable HG-RFS and PFS as compared to patients with T1e disease (both p <0.01). Lastly, we investigated whether T1e disease correlated with other characteristics. Patients with T1e tumors had more TILs (OR: 6.9, p <0.001) and VH (OR: 5.0, p=0.012) (supplementary table 3, https://www.jurology.com). Interestingly, all 12 patients with clinically unfavorable micropapillary VH had T1e disease.

T1 Substaging Improves Stratification of HR-NMIBC Patients

We determined if T1 substaging could improve accuracy of the current HR-NMIBC risk stratification for PFS. To this end, patients were stratified into the high risk (90, 34%) or highest risk (174, 66%) subgroup. Highest risk patients had a higher risk of developing progression than high risk patients (HR 2.1, p=0.001) and a worse PFS (fig. 2, *A*, p=0.017). Patients were stratified by T1 substaging and no difference was found in PFS within the high risk group (T1m vs T1e) (fig. 2, *B*, p.adj=0.754). Importantly, patients with highest risk/T1m disease had a comparable PFS to high risk patients (T1m/T1e)

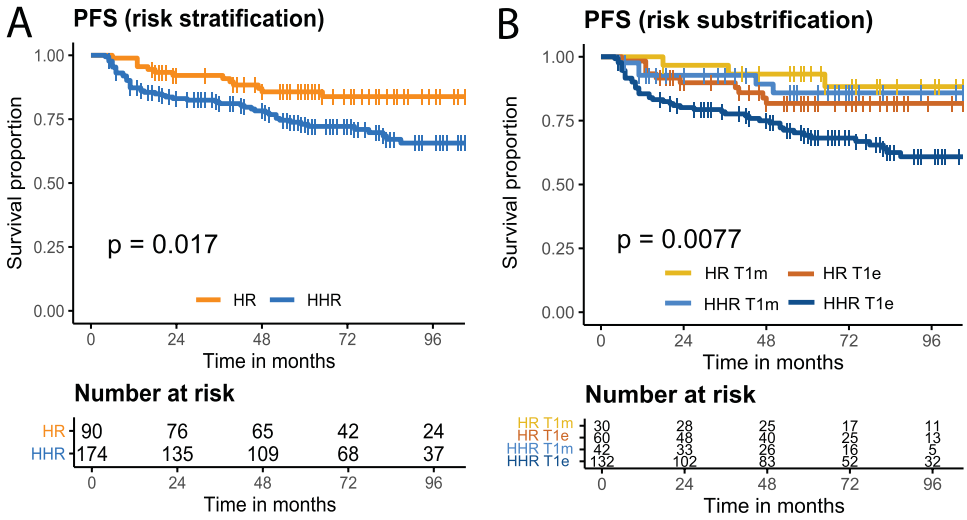

Figure 2. Kaplan-Meier estimates of progression-free survival according to substratification of T1 disease and T1 high-powered field substaging in 264 patients with primary T1 high risk nonmuscle invasive bladder cancer and with visible detrusor in specimen. *A*, PFS high risk vs highest risk subgroup. *B*, PFS high risk vs highest risk subgroups according to T1 substaging. *HR*, high risk patients. *HHR*, highest risk patients. p Value is determined by (pooled) log-rank test.

RIGHTSLINK

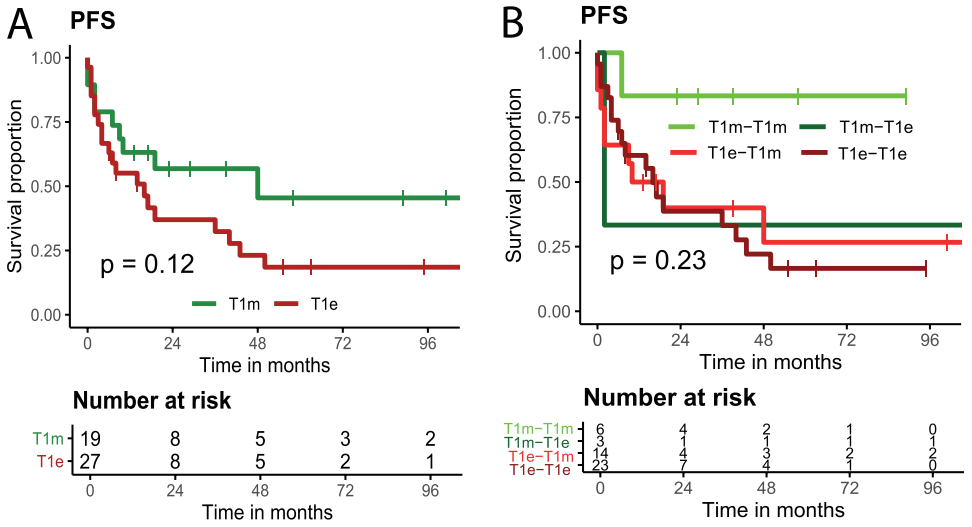

Figure 3. Kaplan-Meier estimates of progression-free survival in 46 patients with primary T1 high risk nonmuscle invasive bladder cancer who developed T1 recurrence, stratified by T1 high-powered field substaging and with visible detrusor in specimen. *A*, PFS of primary T1 patients with T1 recurrences stratified according to T1 substaging. *B*, PFS in primary-recurrent T1 combinations both stratified by T1 substaging. p Value is determined by (pooled) log-rank test.

(p.adj = 0.823). Patients with highest risk/T1e disease had a significantly worse PFS than highest risk/T1m disease (p.adj = 0.038).

Recurrent T1e Disease is Associated with a Very High Risk of Disease Progression

Lastly, we determined whether T1 substaging of recurrences was associated with PFS. Of 264 patients 84 developed nonmuscle invasive recurrences, ie Tis/Ta in 38 and T1 in 46. Patients with T1 recurrences had a worse PFS than patients with Ta/Tis recurrences (HR 4.1, p <0.001). Within the group of patients with T1 recurrences, 19/46 (41%) had T1m and 27/46 (59%) had T1e disease. Patients with a T1e recurrence had a nonsignificant increased risk of progression, compared to patients with a T1m recurrence (HR 1.8, p = 0.12, fig. 3, *A* and supplementary table 4, https://www.jurology.com). Patients with primary and recurrent T1e disease (T1e-T1e) had a worse PFS compared to patients with primary and recurrent T1m disease, but the difference was not statistically significant after multiple testing correction (T1m-T1m; p.adj = 0.19; fig. 3, *B*).

DISCUSSION

T1 substaging is easy to use and predictive of outcome in HR-NMIBC, but it is unknown whether T1 substaging improves patient stratification in the context of current guidelines.[4,8,17,21,24] Thus, we investigated if T1 substaging was associated with BCG failure and whether T1 substaging could be used as a tool to refine risk stratification in a cohort of BCG treated HR-NMIBC patients.

Patients with T1e HR-NMIBC were more likely to fail BCG, suggesting that they should be surveilled with vigilance. In a previous study (79 patients), T1e vs T1m was associated with a worse 5-year RFS (29% vs 64%), but treatment information was unavailable.[25] Rouprêt et al showed a worse RFS in T1b disease (ie below muscularis mucosae in T1 MM-VP substaging).[16] However, analysis did not include important predictive variables such as CIS, LVI and VH. Moreover, by selecting RFS as an end point, LG recurrences, which are not considered BCG failures, were included as events.[4] Therefore, we selected HG-RFS to define BCG failure as our primary end point, since a HG recurrence will affect therapeutic decision making.

In the real-world situation, it may occur that a re-TURBT is not performed, especially when detrusor muscle was visible. Most studies investigating T1 substaging showed an association with PFS and DSS. None of the 36 studies included in a recent meta-analysis took into account the impact of I) adequate BCG treatment, II) a re-TURBT before BCG induction, III) detrusor muscle in the transurethral resection/re-TURBT specimen to prevent the risk of understaging of T1 disease and IV) a comparison of T1m vs TaHG disease.[9,17,20,21,24,26] Therefore, we also performed 2 subanalyses in patients who received a re-TURBT and compared T1m vs TaHG disease to investigate up staging at centralized review.

Guidelines recommend considering an immediate RC in HR-NMIBC patients with highest risk features, but this may lead to overtreatment by RC.[4,8] In line with previous studies, we observed a higher progression rate in patients with highest risk prognostic factors.[2,3,27] To our knowledge, we are the first to demonstrate that T1 substaging

RIGHTSLINK

improves stratification of highest risk patients. Interestingly, highest risk/T1m patients had a risk of progression comparable to patients with high risk disease, indicating that a bladder sparing approach is worthwhile investigating. Hence, prospective trials are needed to assess the safety of bladder-sparing approaches in highest risk/T1m patients. Patients with highest risk/T1e disease had the worst outcome and for these patients immediate RCs should be considered.

T1e disease was associated with the presence of VH and TILs. T1e disease is more invasive than T1m disease, which may clarify the association with VH that is frequently found in advanced disease.[28] In contrast to T1e tumors, T1m disease rarely shows inflammatory features such as TILs.[20] In line with recent findings, we found that TILs were not associated with clinical outcome, possibly because not the overall presence, but specific T cell subsets predict BCG failure.[29]

The main limitation of this study is its retrospective design, which led to missing data on tumor size and therefore we had to exclude this parameter from our analyses. Nonetheless, in a multivariate analysis of 110 patients for whom tumor size was available, T1 substaging was predictive of HG-RFS, PFS and DSS (data not shown). Moreover, it is unlikely that highest risk patients were misclassified as high risk due to missing tumor size, as 51/110 (46%) of the patients with a reported size had a large tumor ($\geq$3 cm), which far exceeds the expected 18% patients with large tumors in European Organization for Research and Treatment of Cancer (EORTC) studies (ie reporting bias in favor of large tumors).[3] The prevalence of LVI varies considerably in literature, yet our cohort showed a relatively low prevalence (5%).[30] LVI scoring was based on hematoxylin and eosin slides, without the use of endothelial markers to facilitate diagnosis.[30] VH pointed towards a nonsignificant favorable outcome, yet results should be treated with caution due to a low number of cases, heterogeneity in variant types and selection bias, as T1 tumors with aggressive VH may have been treated with immediate RC instead of BCG therapy.[4,8] We selected T1 HPF substaging as it was shown that T1 metric substaging (using micrometers) is time-consuming and T1 MM-VP substaging is more difficult and less predictive than T1 HPF substaging.[10,12,21] Additionally, a low interobserver variability has been reported for T1 HPF substaging.[10] T1 HPF substaging is easy to use, can be implemented in every clinical practice without additional costs, is reproducible with a proven prognostic value and has a 100% evaluation rate.[9,10,12,15,21]

CONCLUSIONS

T1 HR-NMIBC patients with T1e tumors were at higher risk of BCG failure compared to both T1m and TaHG tumors, while T1m and TaHG tumors have a similar risk of BCG failure. T1 substaging has potential to guide treatment decisions on BCG vs alternative treatments. A prospective trial is needed to investigate whether bladder-sparing approaches are safe in patients with highest risk/T1m disease. In contrast, for patients with the highest risk/T1e disease early RC should be considered.

REFERENCES

1. Antoni S, Ferlay J, Soerjomataram I et al: Bladder cancer incidence and mortality: a global overview and recent trends. Eur Urol 2017; **71:** 96.
2. Gontero P, Sylvester R, Pisano F et al: Prognostic factors and risk groups in T1G3 non-muscle-invasive bladder cancer patients initially treated with Bacillus Calmette-Guérin: results of a retrospective multicenter study of 2451 patients. Eur Urol 2015; **67:** 74.
3. Sylvester RJ, van der Meijden AP, Oosterlinck W et al: Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol 2006; **49:** 466.
4. Babjuk M, Burger M, Comperat EM et al: European Association of Urology guidelines on non-muscle-invasive bladder cancer (TaT1 and carcinoma in situ)—2019 update. Eur Urol 2019; **76:** 639.
5. Pietzak EJ, Zabor EC, Bagrodia A et al: Genomic differences between "primary" and "secondary" muscle-invasive bladder cancer as a basis for disparate outcomes to cisplatin-based neoadjuvant chemotherapy. Eur Urol 2019; **75:** 231.
6. Zuiverloon TC, Nieuweboer AJ, Vekony H et al: Markers predicting response to bacillus Calmette-Guerin immunotherapy in high-risk bladder cancer patients: a systematic review. Eur Urol 2012; **61:** 128.
7. Martin-Doyle W, Leow JJ, Orsola A et al: Improving selection criteria for early cystectomy in high-grade t1 bladder cancer: a meta-analysis of 15,215 patients. J Clin Oncol 2015; **33:** 643.
8. Chang SS, Boorjian SA, Chou R et al: Diagnosis and treatment of non-muscle invasive bladder cancer: AUA/SUO guideline. J Urol 2016; **196:** 1021.
9. van Rhijn BW, van der Kwast TH, Alkhateeb SS et al: A new and highly prognostic system to discern T1 bladder cancer substage. Eur Urol 2012; **61:** 378.
10. van der Aa MN, van Leenders GJ, Steyerberg EW et al: A new system for substaging pT1 papillary bladder cancer: a prognostic evaluation. Hum Pathol 2005; **36:** 981.
11. Orsola A, Trias I, Raventos CX et al: Initial high-grade T1 urothelial cell carcinoma: feasibility and prognostic significance of lamina propria invasion microstaging (T1a/b/c) in BCG-treated and BCG-non-treated patients. Eur Urol 2005; **48:** 23.
12. Leivo MZ, Sahoo D, Hamilton Z et al: Analysis of T1 bladder cancer on biopsy and transurethral resection specimens: comparison and ranking of T1 quantification approaches to predict progression to muscularis propria invasion. Am J Surg Pathol 2018; **42:** e1.

RIGHTSLINK

13. Patriarca C, Hurle R, Moschini M et al: Usefulness of pT1 substaging in papillary urothelial bladder carcinoma. Diagn Pathol 2016; **11:** 6.

14. Brimo F, Wu C, Zeizafoun N et al: Prognostic factors in T1 bladder urothelial carcinoma: the value of recording millimetric depth of invasion, diameter of invasive carcinoma, and muscularis mucosa invasion. Hum Pathol 2013; **44:** 95.

15. van Rhijn BW, Liu L, Vis AN et al: Prognostic value of molecular markers, sub-stage and European Organisation for the research and treatment of cancer risk scores in primary T1 bladder cancer. BJU Int 2012; **110:** 1169.

16. Rouprêt M, Seisen T, Compérat E et al: Prognostic interest in discriminating muscularis mucosa invasion (T1a vs T1b) in nonmuscle invasive bladder carcinoma: French national multicenter study with central pathology review. J Urol 2013; **189:** 2069.

17. Kardoust Parizi M, Enikeev D, Glybochko PV et al: Prognostic value of T1 substaging on oncological outcomes in patients with non-muscle-invasive bladder urothelial carcinoma: a systematic literature review and meta-analysis. World J Urol 2019; **38:** 1437.

18. Humphrey PA, Moch H, Cubilla AL et al: The 2016 WHO classification of tumours of the urinary system and male genital organs-part B: prostate and bladder tumours. Eur Urol 2016; **70:** 106.

19. Colombo R, Hurle R, Moschini M et al: Feasibility and clinical roles of different substaging systems at first and second transurethral resection in patients with T1 high-grade bladder cancer. Eur Urol Focus 2018; **4:** 87.

20. Raspollini MR, Montironi R, Mazzucchelli R et al: pT1 high-grade bladder cancer: histologic criteria, pitfalls in the assessment of invasion, and substaging. Virchows Arch 2020; **477:** 3.

21. Fransen van de Putte EE, Otto W, Hartmann A et al: Metric substage according to micro and extensive lamina propria invasion improves prognostics in T1 bladder cancer. Urol Oncol 2018; **36:** 361.

22. Hendry S, Salgado R, Gevaert T et al: Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. Adv Anat Pathol 2017; **24:** 311.

23. Kamat AM, Sylvester RJ, Bohle A et al: Definitions, end points, and clinical trial designs for non-muscle-invasive bladder cancer: recommendations from the International Bladder Cancer Group. J Clin Oncol 2016; **34:** 1935.

24. Bertz S, Denzinger S, Otto W et al: Substaging by estimating the size of invasive tumour can improve risk stratification in pT1 urothelial bladder cancer-evaluation of a large hospital-based single-centre series. Histopathology 2011; **59:** 722.

25. Nishiyama N, Kitamura H, Maeda T et al: Clinicopathological analysis of patients with non-muscle-invasive bladder cancer: prognostic value and clinical reliability of the 2004 WHO classification system. Jpn J Clin Oncol 2013; **43:** 1124.

26. Gontero P, Sylvester R, Pisano F et al: The impact of re-transurethral resection on clinical outcomes in a large multicentre cohort of patients with T1 high-grade/grade 3 bladder cancer treated with bacille Calmette-Guérin. BJU Int 2016; **118:** 44.

27. Willis DL, Fernandez MI, Dickstein RJ et al: Clinical outcomes of cT1 micropapillary bladder cancer. J Urol 2015; **193:** 1129.

28. Baumeister P, Zamboni S, Mattei A et al: Histological variants in non-muscle invasive bladder cancer. Transl Androl Urol 2019; **8:** 34.

29. Rouanne M, Betari R, Radulescu C et al: Stromal lymphocyte infiltration is associated with tumour invasion depth but is not prognostic in high-grade T1 bladder cancer. Eur J Cancer 2019; **108:** 111.

30. Mari A, Kimura S, Foerster B et al: A systematic review and meta-analysis of the impact of lymphovascular invasion in bladder cancer transurethral resection specimens. BJU Int 2019; **123:** 11.

RIGHTSLINK

Paper IV