Bladder Cancer

Artificial Intelligence–based Detection of FGFR3 Mutational Status Directly from Routine Histology in Bladder Cancer: A Possible Preselection for Molecular Testing?

Chiara Maria Lavinia Loeffler a, Nadina Ortiz Bruechle b, Max Jung b, Lancelot Sellier b, Michael Rose b, Narmin Ghaffari Laleh a, Ruth Knuechel b, Titus J. Brinker c, Christian Trautwein a, Nadine T. Gaisa b, i, *, Jakob N. Kather a, d, e

a Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany; b Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany; c Digital Biomarkers for Oncology Group (DBO), National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany;d Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany; e Pathology & Data Analytics, Leeds Institute of Medical Research at St James’s, University of Leeds, Leeds, UK

Abstract

Background: Fibroblast growth factor receptor (FGFR) inhibitor treatment has become the first clinically approved targeted therapy in bladder cancer. However, it requires previous molecular testing of each patient, which is costly and not ubiquitously available.

Objective: To determine whether an artificial intelligence system is able to predict mutations of the FGFR3 gene directly from routine histology slides of bladder cancer.

Design, setting, and participants: We trained a deep learning network to detect FGFR3 mutations on digitized slides of muscle-invasive bladder cancers stained with hematoxylin and eosin from the Cancer Genome Atlas (TCGA) cohort (n = 327) and validated the algorithm on the “Aachen” cohort (n = 182); n = 121 pT2–4, n = 34 stroma-invasive pT1, and n = 27 noninvasive pTa tumors).

Outcome measurements and statistical analysis: The primary endpoint was the area under the receiver operating curve (AUROC) for mutation detection. Performance of the deep learning system was compared with visual scoring by an uropathologist.

Results and limitations: In the TCGA cohort, FGFR3 mutations were detected with an AUROC of 0.701 (p < 0.0001). In the Aachen cohort, FGFR3 mutants were found with an AUROC of 0.725 (p < 0.0001). When trained on TCGA, the network generalized to the Aachen cohort, and detected FGFR3 mutants with an AUROC of 0.625 (p = 0.0112). A subgroup analysis and histological evaluation found highest accuracy in papillary growth, luminal gene expression subtypes, females, and American Joint Committee on Cancer (AJCC) stage II tumors. In a head-to-head comparison, the deep learning system outperformed the uropathologist in detecting FGFR3 mutants.

Conclusions: Our computer-based artificial intelligence system was able to detect genetic alterations of the FGFR3 gene of bladder cancer patients directly from histological slides. In the future, this system could be used to preselect patients for further molecular testing. However, analyses of larger, multicenter, muscle-invasive bladder cancer cohorts are now needed in order to validate and extend our findings.

i These authors share co-senior authorship.
* Corresponding author. Institute of Pathology, RWTH Aachen University, Pauwelsstrasse 30, 52074 Aachen, Germany. Tel. +49-241-8036118; Fax: +49-241-8082439. E-mail address: ngaisa@ukaachen.de (N.T. Gaisa).

https://doi.org/10.1016/j.euf.2021.04.007

2405-4569/© 2021 The Author(s). Published by Elsevier B.V. on behalf of European Association of Urology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

Since its approval by the Food and Drug Administration (FDA) in 2019, fibroblast growth factor receptor (FGFR) inhibitor treatment has become the first targeted therapy in advanced muscle-invasive bladder cancer (MIBC). Approval by the European Medicines Agency (EMA) is awaited for 2022. However, treatment is restricted to patients with proven alterations of the FGFR3 gene—or the less frequently altered FGFR2 gene. Identification of these patients requires molecular testing of tumor tissue [1], which is not ubiquitously available. Therefore, easy, fast, and cheap preselection of patients with subsequent genetic testing for FGFR3 alterations is desirable.

Hematoxylin and eosin (H&E)-stained histopathological slides are part of the routine diagnostic workup of transurethral resections and cystectomy specimens of bladder cancer patients. These slides can be digitized, and the resulting images contain innumerable visual features. While pathologists routinely extract clinically relevant information from tissue slides, some visual patterns could be too subtle or scarce for reliable detection. Additionally, even in the case of clearly defined patterns with prognostic relevance, intra- and interobserver reliability of pathologists can be suboptimal [2,3]. Until a few years ago, computer-based analysis of digitized histological slides was confined to simple tasks such as tumor detection [4], but recently deep learning has enabled significant progress in this field [5]. Increasingly, deep learning is used for end-to-end studies, where a high-level concept such as a genetic mutation can be predicted directly from raw image data alone [6]. For example, in colorectal cancer, this has enabled robust detection of clinically relevant molecular alterations directly from H&E images [6,7]. In bladder cancer, there are not yet any clinically validated deep learning–based systems for the detection of molecular alterations.

Therefore, in this retrospective single-center study, we built upon our previous experience in other tumor types and aimed to develop a system, using deep learning, for prescreening of bladder tumors for FGFR3 mutations.

2. Patients and methods

2.1. The Cancer Genome Atlas bladder cancer data set

Data of the Cancer Genome Atlas (TCGA) bladder cancer (BLCA) data set were acquired from the Genomics Data Commons (GDC) data portal [8]. Formalin-fixed paraffin-embedded (FFPE) histological slides (labeled as “diagnostic slides”) and corresponding genetic data were available for 327 MIBC cases (see Fig. 1B). Mutational status was acquired through the cBioPortal database [9]. For further details, see the Supplementary material.

2.2. Aachen bladder cancer cohort

The Aachen cohort comprised the FFPE H&E slides of 121 muscle-invasive (pT2–4; n = 47 not otherwise specified [NOS], n = 41 urothelial squamous mixed, and n = 33 squamous carcinomas), 34 stroma-invasive pT1 NOS, and 27 noninvasive pTa NOS bladder cancers from the local pathology archive with known FGFR3 mutational status (for information on cohorts and publications, see Supplementary Table 1). Approval of the local ethics committee was obtained for all experiments (EK55-20). In total, 182 patients with 24% (n = 43) FGFR3 mutations were included (see Fig. 1B). They were either detected by whole exome sequencing [9] or identified by using the SNAPshot method for the simultaneous detection of hotspot mutations according to Hurst et al [10].

2.3. Visual judgment of FGFR3 mutations by an expert pathologist

To compare deep learning–based detection of FGFR3 mutants with the performance of a histopathologist, all cases (TCGA and Aachen) were visually scored for patterns that had been associated with FGFR3 mutational status in the literature [11]. For further details, see the Supplementary material and Supplementary Tables 2 and 3.

2.4. Artificial intelligence–based detection of FGFR3 mutations

All samples were preprocessed according to the “Aachen Protocol for Deep Learning Histology” [12]. In particular, tumor tissue was manually annotated in every single histological image (see Fig. 1A). This tumor-bearing region was subsequently tessellated into rectangular image tiles of 256 μm edge length, processed at 512 px edge length. A deep learning network (shufflenet with two output neurons) was trained on these image tiles as described before [13]. For all patients in the test set, a prediction was made for each image tile. Subsequently, the fraction of all tiles predicted to be FGFR3 mutated were used as an “artificial intelligence (AI) patient score” for each patient. Statistical significance was assessed by an unpaired two-tailed t test, which was run on the patient scores. Experiments were run on two NVidia RTX Titan graphics processing units using Matlab R2019b (Mathworks, Natick, MA, USA). Relevant deep learning hyperparameters are listed in Supplementary Table 4. All source codes are available on https://github.com/jnkather/DeepHistology. All methods were described previously in more detail and were extensively validated in other tumor types [13].
3. Results

3.1. AI-based prediction of FGFR3 status

In the primary experiment (for the experimental design, see Fig. 1), we tested whether FGFR3-mutated bladder cancer cases were detectable from histology by deep learning in a three-fold cross-validated within-cohort experiment. Indeed, in the TCGA cohort, FGFR3 mutations were detectable from histology alone with an area under the receiver operating curve (AUROC) of 0.701 (p < 0.0001; see Fig. 2A). In a subgroup analysis of this cohort, we found highest accuracy in luminal mRNA expression subtypes (receiver operating curve [ROC] 0.73; see Fig. 3A), females (ROC 0.76; see Fig. 3C), and American Joint Committee on Cancer (AJCC) stage II tumors (ROC 0.85; see Fig. 3E). We then evaluated the performance of the same algorithm in a within-cohort experiment in the Aachen cohort. Again, this experiment was performed by a three-fold cross-validation, which strictly separates patients in the training and test sets. This experiment demonstrated that FGFR3 mutations can significantly (p < 0.0001 and p = 0.0028, respectively) be detected in the overall cohort and the “pT1 or higher, urothelial/mixed” subgroup with AUROCs of 0.725 and

![Experimental design](image)

Fig. 1 – Experimental design. (A) Workflow with multiple steps: annotation of tumor, tessellation of whole slide images in tiles and normalization, training of the neuronal network either within the cohort, and a three-fold cross validation or validation and prediction on the other cohort, that is, the cohort was randomly split into a training set (two-thirds of all cases) and a validation set (one-third of all cases). This was repeated three times, in order to test once on every patient. The primary endpoint was the area under the receiver operating curve, with the lower and upper bound of a ten-fold bootstrapped experiment used as confidence intervals as described before [13]. In addition to the ROC analysis, test statistics (sensitivity, specificity, positive predictive value, and negative predictive value) were assessed for a threshold value of 0.25. (B) Consort flow diagrams of our two cohorts. Icons made by Smashicons from [https://www.flaticon.com]. AC = Aachen; BLCA = bladder cancer; FGFR3 = fibroblast growth factor receptor 3; MUT = mutated; NOS = not otherwise specified; ROC = receiver operating curve; TCGA = The Cancer Genome Atlas.

Please cite this article in press as: Loeffler CML, et al. Artificial Intelligence-based Detection of FGFR3 Mutational Status Directly from Routine Histology in Bladder Cancer: A Possible Preselection for Molecular Testing?. Eur Urol Focus (2021), https://doi.org/10.1016/j.euf.2021.04.007
0.681, respectively (see Fig. 2D and 2E). Analysis of “pT2 or higher urothelial” and a mixed and pure squamous bladder cancer cohort did not show significant results, mainly due to low absolute numbers of mutated cases (see Table 1). In an external validation experiment, the highest methodical standard for deep learning classifiers, the TCGA-trained algorithm for validation on the Aachen cohort yielded a significant ($p = 0.0112$) detection of FGFR3 mutant cases with an AUROC of 0.625 (see Fig. 2F and Table 1). Confusion matrices in Supplementary Figures 1C and 1D showed that classifiers with high sensitivity and moderate specificity were attainable with this technology.

3.2. Comparison of the deep learning system with expert-based analysis

One of the key achievements of an unbiased AI-based analysis is the identification of subtle visual patterns beyond established morphological features. The expert pathologist’s judgment showed only moderate overall performance, with an AUROC of 0.607 for the Aachen cohort and 0.563 for the TCGA cohort. The performance of the pathologist’s score was assessed by confusion matrix analysis (Supplementary Fig. 1A and 1B). Applying a moderate stringent judgment of two positive criteria out of four, the pathologist correctly classified nine of 49 FGFR3-mutated samples of the TCGA cohort with sensitivity of 18% only, specificity of 94%, a positive predictive value (PPV) of 36%, and a negative predictive value (NPV) of 87%. Among the Aachen cohort, 13/43 FGFR3-mutated samples were classified correctly, resulting in sensitivity of 30%, specificity of 91%, a PPV of 50%, as well as an NPV of 81%.

3.3. Histological features critical for pathologists’ or AI-based classification

In the TCGA cohort, the most discriminating feature for correct AI-based classification (with a threshold value of 0.25) of true FGFR3-mutated samples (AI-based 32/49 true mutated) was papillary morphology (30/49). The same was found in the Aachen cohort (AI-based 36/43 true mutated), where 30 correctly identified samples showed papillary morphology (Supplementary Tables 2 and 3). To further clarify which visual features were detected by the AI-based analysis, the highest-scoring image tiles were assessed by a uropathologist. This confirmed papillary growth as the most evident feature associated with FGFR3 mutations. In addition, correct tile classification seemed to be associated with polar epithelial cell sheets, roundish areas, and densely packed, but homogenous, regular nuclei. Perinuclear clearing was seen in some of the top tiles, but irregular nuclei and distinct cell borders were not among the most evident features of the top tiles (see Fig. 2B and 2C). On a patient level, of the ten highest-scoring patients in the TCGA test set, nine showed papillary morphology.

Fig. 2 – AI-based prediction of FGFR3 mutational status: (A) AUROC TCGA, (B) top tiles, TCGA cohort FGFR3 mutated, (C) top tiles, TCGA cohort FGFR3 wild type, (D) AUROC AC—within all, (E) AUROC AC—pT1 or higher, and (F) AUROC train on TCGA validate on AC—all. AC = Aachen bladder cancer cohort; AI = artificial intelligence; AUROC = area under the receiver operating curve; FGFR3 = fibroblast growth factor receptor 3; FPR = false positive rate; MUT = FGFR3 mutated; TPR = true positive rate; TCGA = The Cancer Genome Atlas bladder cancer cohort; WT = FGFR3 wild type.
In the Aachen cohort, a single case with previously known intratumoral heterogeneity of FGFR3 mutational status was included (described by Heide et al [9]). In this unusual case, the deep learning system correctly predicted the intratumor heterogeneity: the papillary-invasive part of the tumor was correctly predicted to be FGFR3 mutated with high tile-level prediction scores (FGFR3 mutated, S249C and R248C). Conversely, a subjacent diffuse infiltrating tumor region was correctly predicted to be FGFR3 wild type with low tile-level prediction scores (Fig. 4).

4. Discussion

In this study, we used an AI-based algorithm to predict FGFR3 mutational status directly from histological H&E slides. We believe that digital pathology will markedly change routine pathology within the next 10 yr. In this light, preselection of patients for further molecular analyses will be an important issue, since it will be easily applicable on top of digital pathology workflows. So far, AI-based studies in bladder cancer used end-to-end deep learning to predict molecular subtypes of MIBC [14], but without focusing on a specific molecular alteration. Other related studies in bladder cancer analyzed the classification of MIBC and non-MIBC [15], grading of non-MIBC [16], recurrence prediction of non-MIBC [17], and tumor budding/staging of MIBC [18]. However, to the best of our knowledge, our histology-based deep learning approach is the first study identifying clinically relevant molecular treatment targets in bladder cancer samples in two independent patient cohorts.

When analyzing real-world patient cohorts, meticulous attention should be paid at clinicopathological features. Here, we used our in-house patient cohort in addition to the publicly available TCGA bladder cancer data set. The TCGA cohort (training set) comprised patients with MIBC only, which most reliably reflects advanced bladder cancers, for which anti-FGFR therapy is currently approved in the USA. However, among the TCGA slides, there were lots of papillary tumors, which did not show (muscle) invasive tumor parts on the whole slide images (but the clinical data indicated stage pT2–4). To make the cohorts comparable regarding papillary morphology and the number of positive events (FGFR3 mutations), we decided to also
include pTa and pT1 tumors in our analysis. Moreover, squamous differentiated carcinomas were incorporated for better reflection of the TCGA bladder cancer cohort.

After homogenizing the cohorts, the algorithm was trained and validated on the MIBC TCGA samples and also validated on the Aachen cohort. Like any classifier, our algorithm shows a tradeoff of sensitivity and specificity. Using a high-sensitivity operating point, the algorithm could be used as a low-specificity prescreening tool. This would enable detection of most truly mutated cases, while false positive cases would be removed in subsequent genetic testing. Overall, such a two-step process would reduce sequencing requirements and could save costs. However, future prospective trials are needed to demonstrate the clinical feasibility of such a two-step diagnostic approach.

We also compared the performance of the AI-based mutation detection with the H&E-based visual prediction of an experienced pathologist. The specialist judged all TCGA and Aachen slides for four different “FGFR3-mutated features” proposed by Al-Ahmadie et al. [11]. The most reliable discriminating feature for correct FGFR3-mutated classification was papillary growth. Other features stated by Al-Ahmadie et al. [11] were less efficient, and both the pathologist and the AI-based algorithm failed several times. Re-evaluation of the whole slide images revealed certain constraints and explanations for this: (1) the substantial part of squamous differentiation in the TCGA cohort (and ours accordingly) is challenging since cell borders are quite distinct, nuclei are quite small, and the surrounding cytoplasm is quite clear; (2) roundish areas of papillary growth patterns can be imitated by a nodular infiltration pattern; (3) infiltrating immune cells could mask condensed nuclei; and (4) due to fixation artifacts, a substantial part of the whole slide images shows clear cytoplasm with retraction gaps, which can be misinterpreted as perinuclear clearing. However, the neural network showed a higher AUROC and therefore better performance for the prediction of FGFR3 mutations in bladder cancer than an experienced pathologist (AUROC = pathologist: Aachen cohort = 0.607 and TCGA = 0.563 vs AUROC = deep learning: Aachen cohort = 0.725 and TCGA = 0.701). This supports that our method will be useful in clinical practice for identifying patients for further molecular screening.

Most impressively, the deep learning algorithm was also able to predict correctly intratumor heterogeneity of FGFR3 mutations in a patient with mutations exclusive to the papillary tumor parts and lack of mutations in the diffuse infiltrating areas. Although this analysis was restricted to a single case, this provides a proof of principle for the capacity of AI-based systems to successfully visualize genetic intratumor heterogeneity although being trained only on bulk sequencing data. This is especially important since in large tumors, the analyzed area might be important. Pouessel et al. [19] showed a frequent discrepancy (50%) of the mutational FGFR3 status between two distant areas of invasive pT2 tumors, that is, between the papillary part oriented toward the lumen and a deeper area toward the invasion front. Thus, preselective studies on the heterogeneity of actionable molecular targets may avoid therapy failure, which otherwise may foster intrinsic mechanisms of resistance. Deep learning could also prescreen various slides in order to both quantify the mutated proportion of the tumor and select the optimal area for subsequent molecular screening. However, the cutoff for effective therapy based on the mutated proportion of the tumor still is subject of debate and remains to be investigated clinically. Since we had only limited samples with multiregion sequencing data (n = 10 patients [9]) and TCGA does not include spatial information of mutations, we could not further validate the superiority of AI in terms of intratumoral heterogeneity in this study.

In addition, future studies may not only address intratumor mutational heterogeneity, but also further improve AI-based genotyping of MIBC by addressing the following limitations: beyond FGFR3 mutations, therapeutically relevant events of rare FGFR gene fusions and FGFR2 gene alterations should also be considered in order to represent
the full actionable range. In addition, since bladder cancer shows striking differences in histological differentiation, growth, and clinical course, diverse histological cohorts should be analyzed. All efforts require large and well-characterized patient cohorts in multicenter settings. By showing a proof of concept for AI-based molecular diagnosis in MIBC, our study establishes a baseline for such follow-up studies.

5. Conclusions

In this study, we show that deep learning can predict FGFR3 mutational status of patients directly from bladder cancer H&E slides. In particular, this is possible in pT2–4 tumors, where FGFR inhibition is currently a clinically approved treatment option. Analyses of larger, multicenter MIBC cohorts are needed to further validate and improve these findings. After prospective validation, deep learning could be used as an inexpensive diagnostic tool in digital pathology workflows.

Financial disclosures: Nadine T. Gaisa certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (e.g., employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

Funding/Support and role of the sponsor: Jakob N. Kather is funded by the Max-Eden-Programme of the German Cancer Aid (Bonn, Germany; grant #70113864) and the German Ministry of Health (“Förderung aufgrund eines Beschlusses des Deutschen Bundestages durch die Bundesregierung”; grant DEEP LIVER, 2MVII–2520DAT111). Nadine T. Gaisa is funded by the German Research Foundation (Bonn, Germany; grant GA 1384/3-1).

Acknowledgments: The authors appreciate the mutational verification of Robert Stoehr done for previous studies.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.euf.2021.04.007.

References


