Deep Learning Identifies Histopathologic Changes in Bladder Cancers associated with Smoke Exposure Status

Okyaz Eminaga*[1], Hubert Lau[2,3], Eugene Shkolyar[4], Eva Wardelmann[5], Mahmoud Abbas*[5]

[1] AI Vobis, Palo Alto, U.S.A.
[2] Department of Pathology and Laboratory Medicine, Veterans Affairs Palo Alto Health Care System

[3] Department of Pathology, Stanford University School of Medicine
[4] Department of Urology, Stanford University School of Medicine

[5] Department of Pathology, University Hospital of Muenster, Germany

Corresponding authors:
Okyaz Eminaga (okyaz.eminaga@aivobis.com)

Abstract

Smoke exposure is associated with bladder cancer (BC). However, little is known whether the histologic changes of BC can predict the smoke exposure status. Given this knowledge gap, the current study investigated the potential association between histology images and smoke exposure status. A total of 483 whole slide histology images of 285 unique cases of BC was available from multiple centers for BC diagnosis. A deep learning model was developed to predict the smoke exposure status and externally validated on BC cases. The development set consisted of 66 cases from two centers. The external validation consisted of 94 cases from remaining centers for patients who either never smoked cigarettes or were active smokers at the time of diagnosis. The threshold for binary categorization was fixed to the median confidence score (65) of the development set. On external validation, AUC was used to assess the randomness of predicted smoke status, we utilized latent feature presentation to determine common histologic patterns for smoke exposure status, and mixed effect logistic regression models determined the parameter independency from BC grade, gender, time to diagnosis and age at diagnosis. We used 2,000-times bootstrap resampling to estimate the 95% Confidence Interval (CI) on the external validation set. The results showed an AUC of 0.67 (95% CI: 0.58–0.76), indicating non-randomness of model classification, with a specificity of 51.2% and sensitivity of 82.2%. Multivariate analyses revealed that our model provided an independent predictor for smoke exposure status derived from histology images, with an odds ratio of 1.710 (95% CI: 1.148–2.54). Common histologic patterns of BC were found in active or never smokers. In conclusion, deep learning reveals histopathologic features of BC that are predictive of smoke exposure, and therefore may provide valuable information regarding smoke exposure status.

Introduction

Bladder cancer (BC) is the sixth most common malignancy in the United States, with an estimated 83,730 new diagnoses in 2021 [1]. Smoking is a well-accepted risk factor for developing BC, accounting for up to 50% of cases [2]. The association between BC and tobacco consumption [1, 3] is attributable to aromatic amines and heterocyclic compounds [4] that incite the carcinogenic molecular mechanisms in bladder urothelium [5]. Previous studies reported a potential correlation between histopathologic features and smoke exposure in BC [6, 7] as a consequence of its carcinogenic mechanisms. Additionally, intensity of cigarette consumption and timing of cessation have been linked to differing histopathologic subtypes and disease aggressiveness [8-11]. However, the existing data from previous studies are characterized by overall inconsistencies and a focus on the controversial association between BC grade and smoke exposure. As a result, it remains unclear whether there is indeed an association between specific histologic features in BC and smoke exposure. Given this knowledge gap, the current study investigated the potential association between histology images and smoke exposure status.

To further characterize the relationship between histopathology and smoke exposure, the current study used deep learning (DL) to examine histologic images of BC and investigated the predictive value of histopathology for smoke exposure status. DL is a broad family of machine learning methods within the artificial intelligence (AI) domain.

DL incorporates deep convolutional neural networks (CNN) [12, 13] which have demonstrated exceptional utility in computer vision due to strong performance in pattern recognition tasks such as in histopathology[14-16] or treatment response prediction from histology images [17, 18]. Herein we investigate the histopathological features associated with smoke exposure and characterize this relationship using DL. We developed a CNN model that predicts, scores, and extracts latent features associated with smoke exposure using histologic images of BC from the multi-center Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial.

Material and Methods

Study cohort

The PLCO Cancer Screening Trial is a multi-center, randomized trial designed to evaluate the impact of screening on cancer related outcomes [19, 20]. Although this trial did not screen for BC, it tracked diagnoses of BC during the trial period. Briefly, 154,900 participants from the general population aged 55 through 74 years were enrolled between 1993 and 2001 [19]. Only subjects without a history of prostate, lung, colorectal or ovarian cancer were enrolled. After enrollment, participants were randomized to "screening" or "no-screening" arms for these malignancies and followed through the end of 2008. Smoke exposure was regularly documented without applying any restriction to smoking exposure levels as part of an annual survey via mailed questionnaires during the follow-

up period[19]. Cancer diagnoses were confirmed by retrieving results and information from medical record systems and the cancer registry system. During the study follow-up period, 1,430 cases of BC were diagnosed.

For model development, in-training validation, and external validation, 285 cases were available from 9 centers (University of Colorado, Georgetown University, Pacific Health Research and Education Institute -Honolulu-, Henry Ford Health System, University of Minnesota, Washington University in St Louis, University of Pittsburgh, University of Utah, and Marshfield Clinic Research Foundation). A total of 483 paraffin-embedded hematoxylin and eosin (H&E)-stained whole slide images (WSI) captured at 20x magnification using a Leica Biosystem Scanner (Wetzlar, Germany) were included. All samples were originally obtained through transurethral resection of bladder tumors. The supplementary figure 1 provides the flowchart for the study cohort.

Data preprocessing

The development dataset consisted of 90 WSI of 55 cases from a single center (Marshfield Clinic Research Foundation) for training and 26 WSI of 16 cases from another center (Georgetown University) for model optimization (in-training validation). A total of 367 WSI from 214 cases from the remaining study centers were available for the external validation set. From these, we limited the validation dataset to cases selected either from never-smokers or active smokers, for a total of 155 WSI from 94 cases diagnosed with BC.

The purpose of excluding former smokers from external validation was to eliminate the potential effect of smoking cessation on histopathological appearance and consequently on downstream analyses.

Given the distinctive of hematoxylin-stained nuclei and contrasting appearance of urothelium from less cellular stromal tissue, we applied a color mask to the WSI thumbnails, whose pixels have hue, saturation, and value (HSV) colors located in the color range between (H:140, S:20, V:50) and (H:150, S:255, V:255) (**Supplementary Figure 1**). Background noise was then filtered from the WSI using the erosion function (kernel size: 2 x 2). Masked areas were expanded and missing portions were filled with the dilation function (kernel size: 5 x 5) provided by OpenCV library[21]. Masked areas were split into 10% overlapping squares and subsequently rescaled and remapped to the original WSI size. The masked areas were tiled on WSI at 10x magnification into small patches (512x512 pixels, one pixel = 1 μm) for processing due to memory constraints. The red, green, and blue (RGB) color space was set as the default color for model training and evaluation. All patches were labelled by one of three cigarette exposure statuses: "never smoker", "active smoker" and "former smoker" during the follow-up period up to BC diagnosis.
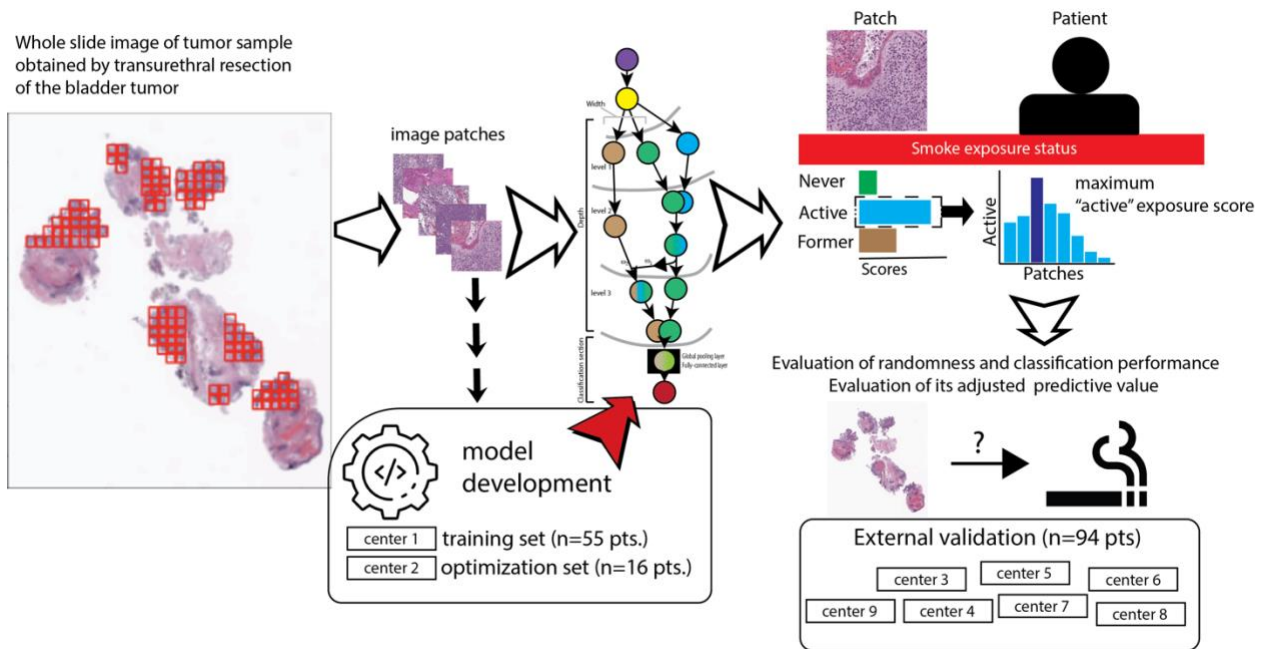
Figure 1 summarizes the workflow of model development and external validation as well as processing the whole slide image to obtain the exposure score for each patient. The whole slide images are tiled in small patches; the image patches from training or optimization aka in-training validation set were used to develop and optimize the model for the predication of the smoke exposure status (never, active and former smoker). After model development, the pipeline processed the whole slides and estimated the smoking exposure status for each patient. Here, the exposure scores derived from the confidence scores of "active smoker" category are first calculated for all image patches per patient and then assigned the maximum exposure score to the corresponding patient.

Model development

Figure 1 summarizes the study's workflow for model development. All models were trained to classify patches into three classes of smoke exposure (i.e., never smoker, former smoker, and active smoker) to facilitate learning distinguishable patterns between different smoke exposure conditions.

Neural architecture search (NAS) with a predefined search space was employed with the training set[22] to determine optimal model morphology. PlexusNet was used to perform a grid search for determining the path, the width, and the depth of model architecture [23]. The grid search was preferred because the search space for architecture

hyperparameters was finite, limited, and searchable using the conventional grid search. The computation cost for grid search was also reasonable. The search space was defined by the following parameters.

$$w \in \{2,3,4\};$$

$$d \in \{3,4,5,6\};$$

$$c \in \{4,6,8\};$$

$$b \in \{\text{soft attention, inception, } vgg, resnet\};$$

$$j \in \{2,3\}$$

where w is the graph branching factor, d the graph level depth, c the initial channel number used to populate the channels in the first and consequence blocks, b the block type, and j the junction number between two paths (root and auxiliary paths). The hyperparameters for model training were prefixed and included the ADAM optimization algorithm[24] with a learning rate of 1e-4 (Beta1: 0.9 and beta2: 0.999 by default), the categorical cross entropy loss function, and a batch size of 32 patches. The evaluation metrics were area under the curve (AUC) of the receiver operating characteristic (ROC) curve and the classification accuracy (CA). Image augmentation was applied with 50% probability and generated a variety of patches to increase the likelihood of a generalizable training set and to regulate the model fitness. Image augmentation included random rotation (angle range: [-40,40]), manipulation of brightness (0.1), image resolution quality (i.e., JPEG image compression rates between 80 and 100%), color saturation (manipulation of HSV color space), and image contrast, as well as the random horizontal flip. NAS using

these search parameters and parallel computing generated and examined 288 models in 48 hours on a single GPU card; the optimal model was determined according to the highest classification accuracy by 2-fold cross validation after 10 training epochs.

After identifying the optimal model architecture, it was trained using the training set at the patch level, using the same augmentation and optimization configuration until model convergence. Model performance was evaluated per epoch using the in-training validation data set. The summation of AUC and CA on the in-training validation set was used to determine the best model at the patch level.

On the external validation set, we estimated and evaluated the exposure scores to determine the histopathologic association between "exposure score" and active smoker status. In order to identify cases with active cigarette smoke exposure from histology images, the maximum confidence score of the "active smoker" class generated from the available patches was applied to the corresponding case. This maximum confidence score is herein labeled as the "exposure score" of the corresponding case. Afterward, the median exposure score was estimated from the development data set and used as the threshold for conversion of the exposure score from a continuous to binarily categorized parameter, and this was locked as the threshold for the external validation data set.

To investigate the effect of the magnification, we repeated the previous steps (except the NAS step; here, we used the resulting model architecture from the first evaluation) to develop and validate a prediction model on images at 20x magnification. After fixing the optimal magnification level, we also examined the attention-based multi-instance learning defined according to Ilse et al. [25] as it has shown performance benefits for certain tasks in digital pathology [26, 27]. The backbone model was the PlexusNet model we defined earlier.

External validation

The external validation data set was used to determine the discriminatory accuracy and randomness of the exposure score for smoking status. Randomness was estimated using AUC [28]. Discriminatory accuracy was determined by classification accuracy, specificity, sensitivity, f1-score, recall, precision, negative and positive predictive values. The contribution of demographic and pathologic variables such as age, gender, BC grade, and time to diagnosis were also evaluated. To assess the independent variables, two mixed-effect regression models were built for prediction of smoke exposure. The base model considered gender and age at diagnosis as random effects, and BC grade and time to diagnosis as fixed effects. Exposure scores were included as a continuous parameter for the first model and as a binary categorization for the second model. The general equation of the mixed effect regression model is:

$$\hat{y}_{(smoke\ status)} = Exposure\ score \bullet BC\ grade + Time\ to\ diagnosis$$
$$+ (gender \mathbin{||} age\ at\ diagnosis)$$

The random effect incorporated gender and age. To avoid a singular model fit resulting from the limited random effect, we defined the interaction between exposure score and BC grade. Moreover, we assumed that BC grade and our exposure score are interacted given their information source (i.e., histology appearance).

In parallel we applied t-SNE (t-Distributed Stochastic Neighbor Embedding) to visualize the 3D space of the latent features resulted from the global pooling directly after the last convolutional layer of the model for all patch images[29]. Here, we labeled the data points according to the decile of the exposure score and the smoking status. Then, we searched for subspaces either dominated by active smokers or never smokers where their decile exposure score gradient is correlated with the smoking status. Here, the boundary definition for each subspace was explored on the 3D data points and fixed when at least 80% of a minimum 200 patch images were originated either from never smoker or active smoker patients. Finally, the patch images were clustered to identify similar patterns for each subspace, and we calculated the overall case number of active or never smokers in each space. Moreover, two genitourinary pathologists (MA and HDL) independently examined the image clusters for each subspace. They provided their overall impression relating to diagnosis and heterogeneity for each subspace and reported any

other notable findings. Both pathologists were blinded to the clinical information and no

time limitation for the examination was specified.

Statistical analyses

All analyses were performed using 2,000-times bootstrap resampling to estimate

the 95% confidence interval (CI) on the external validation set. A t-test was used for mean

comparison between two groups and the non-parametric Kruskal–Wallis test for

distribution evaluation between multiple groups. The reported *p*-values are two-tailed,

and statistical significance was considered when $p < 0.05$. The false discovery rate was

applied for test comparison and the statistical significance was considered when FDR <

0.1. The statistical robustness of AUC was estimated using the statistical Power. We set

the decision threshold for a robust statistical Power to 80% at an alpha level 0.05 [30-34].

Software and Hardware configuration

Statistical analysis was performed using Python 3.8 (Python Software Foundation,

Wilmington, DE) and R 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria). The

Keras library[35], a high-level wrapper of the TensorFlow framework (2.5), and Scikit-

learn[36], were used to develop the models and "lmer4" was used for mixed-effect

regression modeling [37]. Scikit-image (0.17.2), OpenCV (4.10), and matplotlib (3.3.2) were

used for image processing and plotting. All analyses were performed on a GPU machine

with a 44-core Intel processor with 64 GB RAM (Intel, Santa Clara, CA), from which an 8 GB virtual storage memory was allocated for fast read/write access to temporarily store patch images, 10 TB hard disks to store the whole slide images, and a single NVIDIA 1080 ti GPU with 11 GB VRAM.

Results

A total of 10,406 patches was generated for the training set, 3,272 patches for the in-training validation set, and 17,416 for the external validation set. **Table 1** summarizes the cohort description.

Table 1 describes the cohort characteristics of the current study. CI: Confidence interval; IQR: interquartile range; n.c.: not calculable. BC: bladder cancer. BC grade was judged by the criteria definition of PLCO study.

	All	Training set	In-training validation set	External validation set
Patients, n (%)	165 (100)	55 (42.86)	16 (12.38)	94 (44.76)
Age at diagnosis, years, median (IQR)	65 (60–69)	63 (58–68)	69 (65–71)	63 (60–67)
Gender, n (%)				
Male	125 (75.8)	44 (80.0)	13 (81.3)	68 (72.34)
Female	40 (24.2)	11 (20.0)	3 (18.7)	26 (27.66)
Smoking status				
Never smoker	63 (38.2)	9 (16.4)	5 (31.2)	49 (52.13)
Active smoker	57 (34.5)	11 (20.0)	1 (6.3)	45 (47.87)
Former smoker	45 (27.3)	35 (63.6)	10 (62.5)	0
Staging, n (%)				
Stage 0	103 (62.4)	37 (67.3)	13 (81.3)	53 (56.4)

Stage I	33 (20.0)	11 (20.0)	2 (12.5)	20 (21.3)
Stage II	18 (10.9)	4 (7.3)	1 (6.2)	13 (13.8)
Stage III	4 (2.4)	2 (3.6)	0	2 (2.1)
Stage IV	1 (0.6)	0	0	1 (1.1)
Unknown	6 (3.6)	1 (1.8)	0	5 (5.3)
BC grade, n (%)				
I (Well Differentiated)	49 (29.70)	16 (29.1)	9 (56.3)	24 (25.5)
II (Moderately Differentiated)	38 (23.03)	19 (34.55)	4 (25.9)	15 (16.0)
III (Poorly Differentiated)	70 (42.42)	19 (34.55)	3 (18.8)	48 (51.0)
IV (Undifferentiated)	2 (1.21)	1 (1.82)	0	1 (1.1)
Unknown	6 (3.64)	0	0	6 (6.4)
Time to diagnosis, years, median (IQR)	8 (4.67–11.67)	7.33 (4.25–11.12)	8.75 (5.67–11.83)	8.33 (5.25–10.75)
Cigarette exposure duration for active smokers in years, mean (95% CI)	44 (42 – 46)	46 (42 – 49)	n.c.	44 (41 – 46)
Pack-year, mean (95% CI)	55 (47 – 62)	60 (34 – 86)	n.c.	54 (46 – 61)
Number of patches				
Never smoker	12,641 (39.37)	2,008 (19.30)	819 (25.02)	9,814 (53.24)
Active smoker	11,106 (34.59)	2,485 (23.88)	3 (0.09)	8,618 (46.76)
Former smoker	8,363 (26.04)	5,913 (56.82)	2,450 (74.85)	0

The final model architecture was determined (architecture configuration: d=5; w=2; j=3, c=6; b=soft attention) to generate exposure scores for the external validation set. The

parameter capacity of the final model was 272,743 parameters, and the first fully connected layer or the global pooling layer had 64 dense units. The exposure scores of BC cases were equally distributed between the seven centers (Kruskal-Wallis chi-squared = 4.471, degree of freedom = 6, p = 0.6132), indicating no impact of center of diagnosis or variation in H&E staining protocols on the exposure scores.

The exposure score achieved an AUC of 0.64 (95% CI: 0.53 – 0.76) for distinguishing never-smokers from active smokers in the external validation set (**Figure 2**). As a binary category using a threshold of 65, the exposure score provided a balanced accuracy of 68% with an AUC of 0.67 (95% CI: 0.58 – 0.75), a sensitivity of 82% (95% CI: 71 – 93) and specificity of 51% (95% CI: 37–65). The F1-score was 0.70 (95% CI: 0.59 – 0.79) with a precision of 0.61 (95% CI: 0.49 – 0.72) and a recall rate of 0.82 (95% CI: 0.69 – 0.91). Table 2 shows the confusion matrix for smoke exposure status. The statistical power of the study was determined to be 82.2%, surpassing the commonly accepted threshold of 80% at an alpha level of 0.05.

Table 2 confusion matrix for the smoke exposure status determination.

Smoke exposure status		Predicted	
		Low exposure score, n (%)	High exposure score, n (%)
Ground	Never smoker, n (%)	25 (51.0)	24 (49.0)
truth	Active smoker, n (%)	8 (17.8)	37 (82.2)
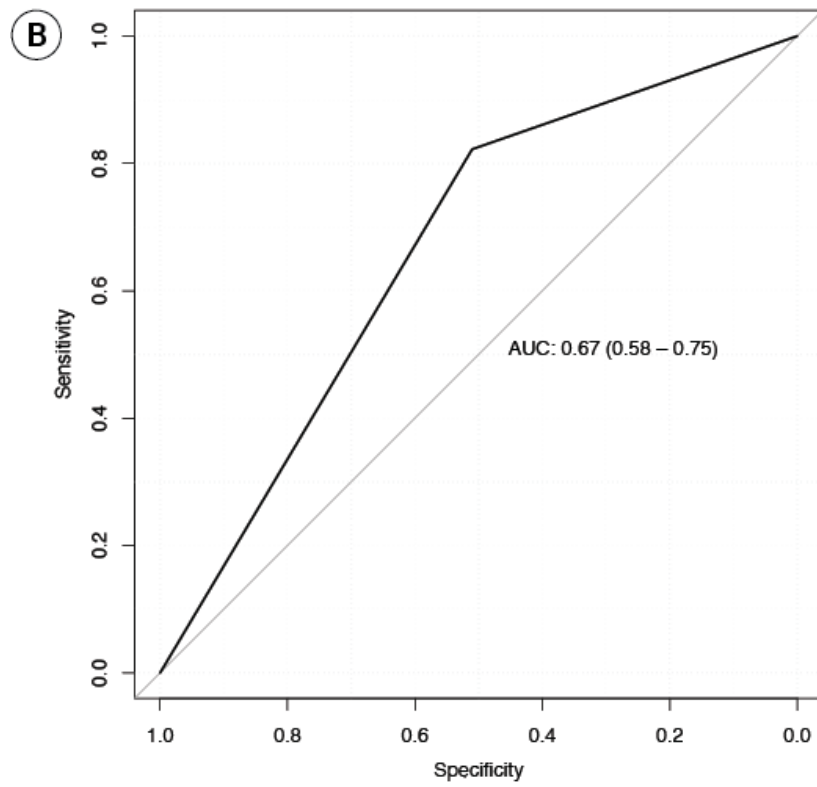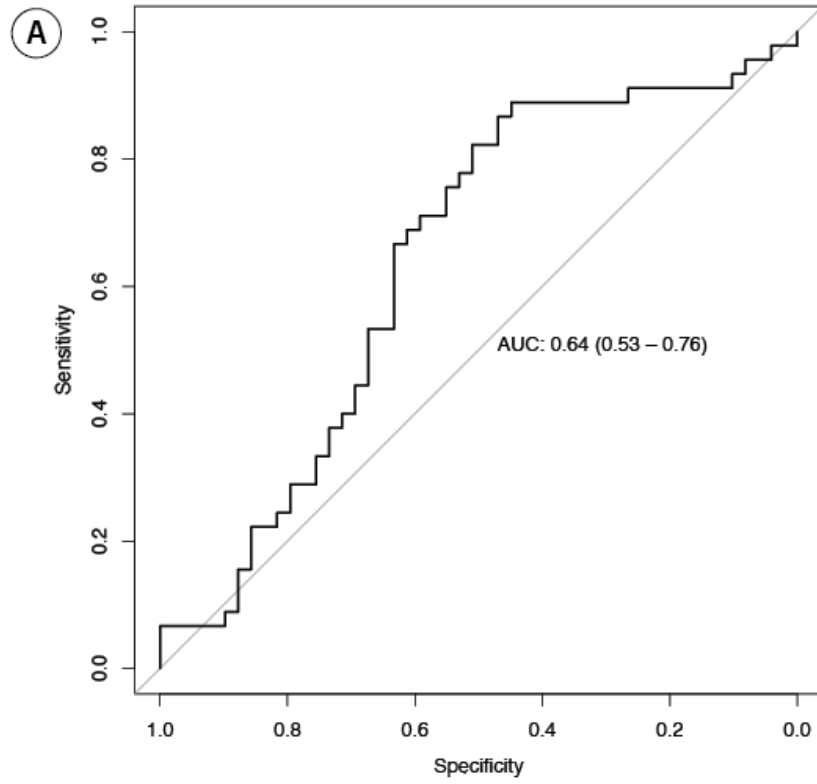
A

AUC: 0.64 (0.53 – 0.76)

B

AUC: 0.67 (0.58 – 0.75)

When comparing models trained on histology images at 10x magnification versus those trained on images at 20x magnification, we observed performance degradation on the external validation set (AUROC: 0.528; 95% CI: 0.432 – 0.624), indicating that magnification does impact performance. In addition to that, MIL failed to provide a non-random prediction for the smoke exposure status (AUROC: 0.561; 95% CI: 0.437 – 0.685).

The average smoking intensity measured by pack-year (PY) and duration (years) was higher for active smokers with exposure scores above 65 compared to active smokers with exposure scores equal to or below 65 (54.8 vs. 47.7 for PY; 44.4 vs. 40.3 for smoking duration in years). However, these were not statistically significant (p=0.5783 for PY; p=0.324 for smoking duration).

The multivariate mixed-effect logistic regression model identified exposure score as the only predictor for active smoking after adjusting for gender, age at diagnosis, time to diagnosis and BC grade (Table 3).

Table 3: Multivariate mixed-effect regression analyses. ++ originated from a base model with categorized exposure score. FDR: false discovery rate.

Parameters	Odd ratio (95% CI)	*FDR*
Exposure score (continuous)	1.009 (1.002–1.02)	0.044
BC grade	1.088 (0.942 – 1.26)	0.259
Time to diagnosis	0.998 (0.996 – 1.00)	0.208

Exposure score • BC grade	0.844 (0.668 – 1.07)	0.208
Exposure score (categorized)++	1.710 (1.148 – 2.54)	0.044

We visualized the 3D space of the latent features for the whole patch images in the external validation set using t-SNE (**Figure 3**). In the spherical 3D space with an uneven surface, we identified one subspace dominated by never smokers and two feature subspaces dominated by active smokers. The first latent feature space had a total of 820 patch images from 48 patients; of those, 697 patch images (85%) were from 41 never smoker (85%). The second subspace was defined by the latent features of 217 patch images that originated from 52 patients; of those, 173 patch images (80%) were from all 45 active smokers (87%). The third latent feature subspace included a total of 214 patch images from 22 patients, from which 204 patch images (95.3%) were originated from 19 smokers (86%). **Figure 3** provides representative patch images for these subspaces whereas the complete patch images for all subspaces are available in the supplementary material.
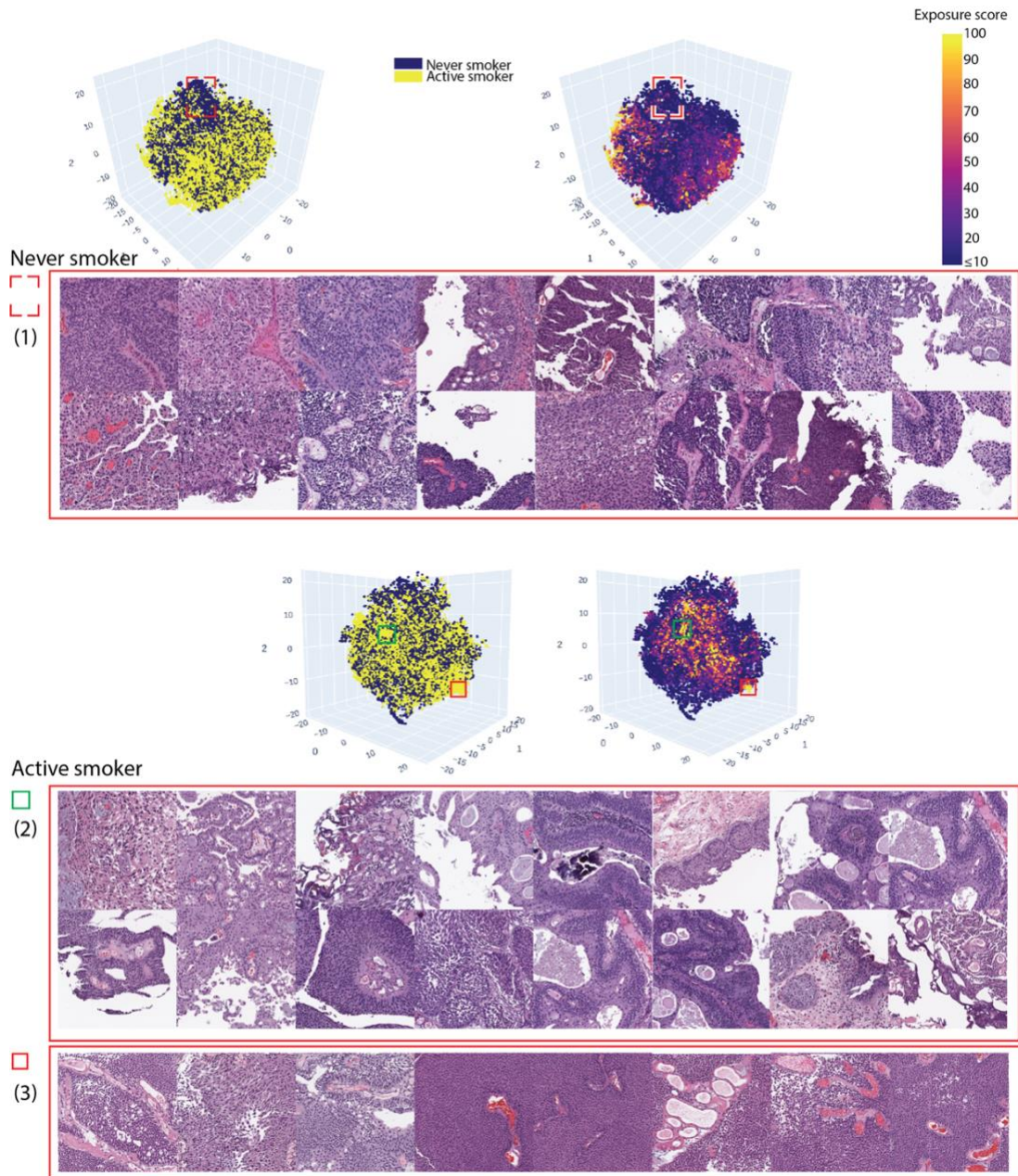
Figure 3 reveals the 3D latent feature representation of the external validation set using t-SNE. We identified three subspaces that met the criteria described in Material and Method section and highlighted by bounding boxes. Representative patch images are provided for each subspace. 1) the first latent feature space has a total of 820 patch images from 48 patients; of those, 697 patch images (85%) are from 41 never smoker (85%); 2) the second subspace is defined by the latent features of 217 patch images that are originated from 52 patients; of those, 173 patch images (80%) are from all 45 active smokers (45 of 52: 87%); 3) the third latent feature subspace includes a total of 214 patch images from 22 patients; of those, 204 patch images (95.3%) are from 19 active smokers (86%). The patch scale: ~512μm at 100× objective magnification. The supplementary section provides the clustered patch images for each subspace after stratifying by smoking status (active vs. never smoker). The supplementary video 1 provides a 3D navigation through the feature space labeled by the tobacco exposure status.

Evaluation of the subspaces by two genitourinary pathologists revealed trends in BC grade

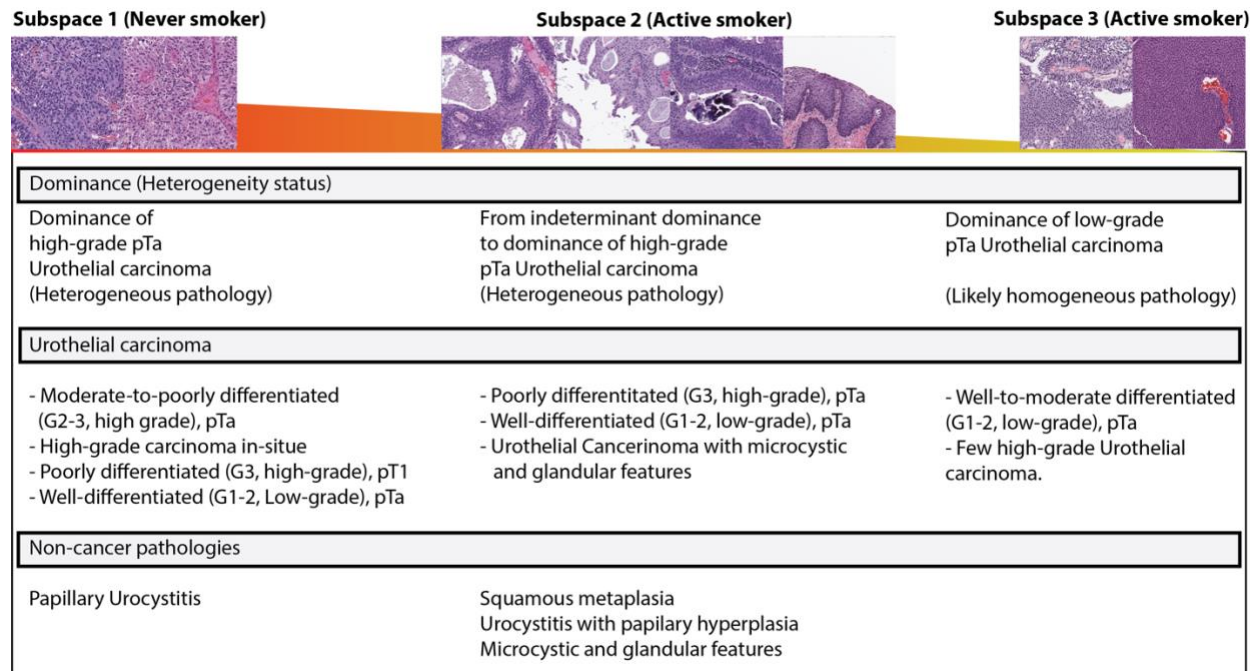and heterogeneity, as shown in **Figure 4** and **supplementary table 1**.

Subspace 1 (Never smoker)	Subspace 2 (Active smoker)	Subspace 3 (Active smoker)
Dominance (Heterogeneity status)		
Dominance of high-grade pTa Urothelial carcinoma (Heterogeneous pathology)	From indeterminant dominance to dominance of high-grade pTa Urothelial carcinoma (Heterogeneous pathology)	Dominance of low-grade pTa Urothelial carcinoma (Likely homogeneous pathology)
Urothelial carcinoma		
- Moderate-to-poorly differentiated (G2-3, high grade), pTa - High-grade carcinoma in-situ - Poorly differentiated (G3, high-grade), pT1 - Well-differentiated (G1-2, Low-grade), pTa	- Poorly differentitated (G3, high-grade), pTa - Well-differentiated (G1-2, low-grade), pTa - Urothelial Cancerinoma with microcystic and glandular features	- Well-to-moderate differentiated (G1-2, low-grade), pTa - Few high-grade Urothelial carcinoma.
Non-cancer pathologies		
Papillary Urocystitis	Squamous metaplasia Urocystitis with papilary hyperplasia Microcystic and glandular features	

Figure 4 summarizes the overall impressions of the subspaces made by two genitourinary pathologists. Each subspace is dominated either by never or active smoker.

Discussion

Chronic tobacco exposure leads to chemical irritation and chronic inflammation in

many organs including the urinary bladder [38-40]. Moreover, tobacco exposure is a well-

known risk factor for BC [41]. We utilized AI to examine whether there is an association

between smoke exposure and histopathologic changes in BC. We found that

histopathologic features are indeed independent predictors for smoke exposure in BC

and that histopathologic changes resulting in distinctly clustering latent features were

specific for smoke exposure status, comparable to work done by Auerbach et al in 1989

which showed histopathologic changes in 88% of BC with similar smoke intensity as our cohort[6].

Although the widespread presence of common histologic features among cases is expected, the 3D feature exploration recognizes three distinctive feature subspaces according to smoke exposure status. Further evaluation of these subspaces unveils trends in malignancy degree (i.e., BC grade and stage) and tumor heterogeneity across these feature subspaces. This observation implies that predictable alterations in the tissue morphology of BC conceivably exist based on smoke exposure status, especially when comparing active smokers to never smokers. The histologic changes may be associated with the carcinogenic effects of cigarette smoke, which impairs DNA-repair activity [42, 43], and causes epigenetic changes[44], epithelial-mesenchymal transition [45, 46], and hypoxic changes [47], consequently altering urothelial cell[48] and tissue morphology [49, 50].

A feature space dominated by active smokers included a subset of images with keratinizing squamous metaplasia, which is an uncommon, potentially pre-malignant condition resulting from chronic irritation of the urothelium [51, 52]. Although multiple studies have associated the development of squamous metaplasia of the respiratory tract with active smokers [53, 54], such an association is not well established for squamous metaplasia of the urinary bladder [55]. Studies of risk factors for bladder cancer found cigarette exposure increases the risk for squamous cell carcinoma of the bladder, a rare

subtype of bladder cancer that may arise from underlying squamous metaplasia [56]. These findings suggest the need for further investigation of the association between squamous metaplasia of the bladder and smoke exposure.

Epidemiologic studies have shown contradicting results regarding the association between smoking status and BC grade. Some studies find an increased risk for high-grade non-muscle invasive bladder cancers in active smokers [8, 9, 53], whereas other studies did not find this association with active smokers [44, 57]. Another study led by Sturgeon et al reported that active smokers with BC diagnosis are rather associated with low-grade BC[58]. The current study did not find any significant association between smoking status and BC grade. One possible explanation is the cancer dynamics of BC in active smokers results in both low- and high-grade tumors. Interestingly, the feature subspaces dominated by active smokers in this study showed different dominant tumor grades, mirroring the contradictory results found in the literature.

Our analyses indicate that small models developed on a relatively limited data set are sufficient to transform histopathologic changes into exposure scores and latent features in order to investigate the association between these changes and smoking exposure. Other benefits of having a small model compared to existing large models include a lower risk for overfitting on limited datasets due to the low parameter capacity[59] and a smaller representation feature dimension (64 globally pooled features for our model vs. 1,280 globally pooled features for EfficientNet B0[60]), lower memory

occupation (3MB for our model vs. 38 MB for EfficientNet B0[60]), and a faster image processing (i.e., at least the half of the time of large state-of-art models). Furthermore, our analysis of exposure scores for cases originating from different centers did not show any biases despite each center's having its own slide processing and staining protocol. Moreover, we were able to identify common patterns in multi-site cases using the latent features generated by the model and a well-accepted data visualization algorithm. Accordingly, applying additional stain color normalization is not necessary and our strategy for image augmentation is sufficient to achieve a generalizable result.

We disregarded the classification prediction for former smokers and only considered the prediction score for active smokers for the association analyses. The primary reason for including the class for former smokers during model development is to regularize the model prediction. However, considering the former smoker status for external validation is not appropriate given the difficulty in identifying the regeneration level of the tissues resulting from stopping smoke exposure. In contrast, we can assume that the baseline environmental exposure (including passive smoking) affects both smokers and non-smokers equally, while smokers have direct exposure to tobacco.

The association between histology images and smoke exposure status at 20x magnification is random, whereas the association at 10x magnification is significant and not random. This finding indicates that histologic patterns (i.e., tissue architecture) at 10x

magnification are predictive of smoke exposure, supporting the key message of the current work regarding the potential association between active smoke exposure and histological images.

In the current study, we further investigated the local features using 3D t-SNE and visualized the 3D feature spaces. We identified global feature subspaces that revealed a high concentration for patches dominated by either active smokers or non-smokers. These findings further indicate reorganization of histologic features associated with these subspaces. Our pathologists investigated and confirmed distinguishable histologic patterns between the subspaces we identified using 3D t-SNE visualization. We emphasize that the 3D visualization of feature spaces provides significantly more information than the 2D visualization of the feature spaces.

Although the current study provides robust results based on a prospectively randomized trial cohort, a potential limitation of the study is the absence of information regarding additional exposures to carcinogens (e.g., occupational, or environmental) leading to BC. Additionally, the PLCO trial cohort may potentially be subject to volunteer bias and differential dropout bias [61]. Nevertheless, the current study reveals the potential of AI in exploring latent patterns from histology images associated with the smoker status, thereby helping to examine hypothesis as in our case for bladder cancers.

According to the evaluation of the MIL approach, we did not find performance benefits compared to the non-MIL approach for our research question, indicating the need to customize the model design according to the study question and the absence of a universal solution. Furthermore, the prediction of the MIL approach for smoke status is random based on the 95% Confidence Interval for AUROC, which was between 0.437 and 0.685. In the in-training validation set, the resection sample from this patient exhibited sparse epithelial components, measuring approximately 0.5 cm in diameter. However, we prioritized the selection of a model that could provide more accurate predictions for patches derived from cases of individuals who had never smoked. It's important to note that the in-training validation set represents a quasi-out-of-distribution compared to the training set. Our rationale for focusing on never smoker cases was based on the assumption that the primary determinants of histology appearance are primarily related to common tumor biology, while smoke exposure acts as an additional factor driving histological alterations.

Conclusion

Histopathologic findings reflected in latent features and exposure scores are associated with the smoking exposure status in bladder cancers. The results of the current

study are useful in advancing our understanding about the impact of smoking on bladder cancer pathology.

Supplementary figures and videos

Supplementary figure 1: The flowchart with the case inclusion or exclusion for the current study cohort.

Supplementary figure 2: (A) a whole slide image (WSI) with bladder tissue samples; (1) the figure shows the distinctive coloration of hematoxylin-stained nuclei and contrasting appearance of urothelium from less cellular stromal tissue. (2) we applied a color mask to the WSI thumbnails, whose pixels have hue, saturation, and value (HSV) colors located in the color range between (H:140, S:20, V:50) and (H:150, S:255, V:255); (3) Background noise was then filtered from the WSI using the erosion function (kernel size: 2 x 2). Masked areas were expanded, and missing portions were filled with the dilation function (kernel size: 5 x 5); (4) masked areas were split into 10% overlapping squares and subsequently rescaled and remapped to the original WSI size. The masked areas were tiled on WSI at 10x magnification into small patches (512x512 pixels, one pixel = 1 μm) for processing due to memory constraints. (B) illustrates the previous steps on another whole slide image.

Supplementary video 1: this video provides a 3D navigation through the feature space labeled by the tobacco exposure status (yellow: active smoker, blue: non-smoker).

References

1.   Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin. 2021;71(1):7-33. Epub 2021/01/13. doi: 10.3322/caac.21654. PubMed PMID: 33433946.
2.   Freedman ND, Silverman DT, Hollenbeck AR, Schatzkin A, Abnet CC. Association between smoking and risk of bladder cancer among men and women. JAMA. 2011;306(7):737-45. Epub 2011/08/19. doi: 10.1001/jama.2011.1142. PubMed PMID: 21846855; PubMed Central PMCID: PMCPMC3441175.
3.   Collaborators GBDT. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet. 2021;397(10292):2337-60. Epub 2021/05/31. doi: 10.1016/S0140-6736(21)01169-7. PubMed PMID: 34051883; PubMed Central PMCID: PMCPMC8223261 Boehringer Ingelheim/Lilly alliance, Bristol Myers Squibb/Pfizer alliance, and Novo Nordisk outside of the submitted work. RA reports consultancy and speakers' fees from UCB, Sandoz, AbbVie, Zentiva, Teva, Laropharm, CEGEDIM, Angelini, Biessen Pharma, Hofigal, AstraZeneca, and Stada outside of the submitted work. BeA reports personal fees from Australian Institute of Sports; non-financial support from Zydus Cadila; and grants and non-financial support from Natural Remedies outside of the submitted work. FG was employed by Public Health England during the conduct of the study, which does not necessarily endorse this study. GJH reports personal fees from the American Heart Association outside of the submitted work. SMSI reports grants from National Heart Foundation of Australia and from the Australian National Health and Medical Research Council (NHMRC) outside of the submitted work. SVK reports grants from Chief Scientist Office and UK Medical Research Council during the conduct of the study. KK reports non-financial support from UGC Centre of Advanced Study (CAS II), Department of Anthropology, Panjab University, Chandigarh, India, outside of the submitted work. StL reports personal fees from Akcea Therapeutics, Amedes, AMGEN, Berlin-Chemie, Boehringer Ingelheim Pharma, Daiichi Sankyo, Lilly, MSD Sharp & Dohme, Novo Nordisk, Sanofi-Aventis, Synlab, Unilever, and Upfield, and non-financial support from Preventicus outside of the submitted work. WM is Program Analyst in Population and

4.   Vineis P, Pirastu R. Aromatic amines and cancer. Cancer Causes Control. 1997;8(3):346-55. Epub 1997/05/01. doi: 10.1023/a:1018453104303. PubMed PMID: 9498898.
5.   Droller MJ. Alterations of the p53 gene in occupational bladder cancer in workers exposed to aromatic amines. J Urol. 1998;160(2):618. Epub 1998/07/29. doi: 10.1016/s0022-5347(01)62973-2. PubMed PMID: 9679939.
6.   Auerbach O, Garfinkel L. Histologic changes in the urinary bladder in relation to cigarette smoking and use of artificial sweeteners. Cancer. 1989;64(5):983-7. Epub 1989/09/01. doi: 10.1002/1097-0142(19890901)64:5<983::aid-cncr2820640502>3.0.co;2-9. PubMed PMID: 2758391.

7.   Friedell GH, Burney SW, Bell JR, Soto E. Pathology as related to tryptophan metabolite excretion, occupational history, and smoking habits in patients with bladder cancer. J Natl Cancer Inst. 1969;43(1):303-6. Epub 1969/07/01. PubMed PMID: 5796395.

8.   Barbosa ALA, Vermeulen S, Aben KK, Grotenhuis AJ, Vrieling A, Kiemeney LA. Smoking intensity and bladder cancer aggressiveness at diagnosis. PLoS One. 2018;13(3):e0194039. Epub 2018/03/24. doi: 10.1371/journal.pone.0194039. PubMed PMID: 29570711; PubMed Central PMCID: PMCPMC5865728.

9.   Pietzak EJ, Mucksavage P, Guzzo TJ, Malkowicz SB. Heavy Cigarette Smoking and Aggressive Bladder Cancer at Initial Presentation. Urology. 2015;86(5):968-72. Epub 20150717. doi: 10.1016/j.urology.2015.05.040. PubMed PMID: 26190088.

10.   Jiang X, Castelao JE, Yuan JM, Stern MC, Conti DV, Cortessis VK, et al. Cigarette smoking and subtypes of bladder cancer. Int J Cancer. 2012;130(4):896-901. Epub 2011/03/18. doi: 10.1002/ijc.26068. PubMed PMID: 21412765; PubMed Central PMCID: PMCPMC3210924.

11.   Grotenhuis AJ, Ebben CW, Aben KK, Witjes JA, Vrieling A, Vermeulen SH, et al. The effect of smoking and timing of smoking cessation on clinical outcome in non-muscle-invasive bladder cancer. Urol Oncol. 2015;33(2):65 e9-17. Epub 2014/07/16. doi: 10.1016/j.urolonc.2014.06.002. PubMed PMID: 25023787.

12.   Szegedy C, Ioffe S, Vanhoucke V, Alemi AA, editors. Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence; 2017.

13.   LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44. Epub 2015/05/29. doi: 10.1038/nature14539. PubMed PMID: 26017442.

14.   Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med. 2019;16(1):e1002730. Epub 2019/01/25. doi: 10.1371/journal.pmed.1002730. PubMed PMID: 30677016; PubMed Central PMCID: PMCPMC6345440.

15.   Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep. 2018;8(1):3395. Epub 2018/02/23. doi: 10.1038/s41598-018-21758-3. PubMed PMID: 29467373; PubMed Central PMCID: PMCPMC5821847.

16.   He Y, Zhao H, Wong STC. Deep learning powers cancer diagnosis in digital pathology. Comput Med Imaging Graph. 2021;88:101820. Epub 2021/01/17. doi: 10.1016/j.compmedimag.2020.101820. PubMed PMID: 33453648; PubMed Central PMCID: PMCPMC7902448.

17.   Sakellaropoulos T, Vougas K, Narang S, Koinis F, Kotsinas A, Polyzos A, et al. A Deep Learning Framework for Predicting Response to Therapy in Cancer. Cell Rep. 2019;29(11):3367-73 e4. doi: 10.1016/j.celrep.2019.11.017. PubMed PMID: 31825821.

18.   Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer cell. 2020;38(5):672-84. e6.

19.   Hasson MA, Fagerstrom RM, Kahane DC, Walsh JH, Myers MH, Caughman C, et al. Design and evolution of the data management systems in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials. 2000;21(6 Suppl):329S-48S. Epub 2001/02/24. doi: 10.1016/s0197-2456(00)00100-8. PubMed PMID: 11189686.

20.   Prorok PC, Andriole GL, Bresalier RS, Buys SS, Chia D, Crawford ED, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials. 2000;21(6 Suppl):273S-309S. Epub 2001/02/24. doi: 10.1016/s0197-2456(00)00098-2. PubMed PMID: 11189684.

21.   Bradski G, Kaehler A. Learning OpenCV: Computer vision with the OpenCV library: " O'Reilly Media, Inc."; 2008.

22.   Elsken T, Metzen JH, Hutter F. Neural architecture search: A survey. The Journal of Machine Learning Research. 2019;20(1):1997-2017.

23.   Eminaga O, Abbas M, Kunder C, Loening AM, Shen J, Brooks JD, et al. Plexus Convolutional Neural Network (PlexusNet): A novel neural network architecture for histologic image analysis. arXiv preprint arXiv:190809067. 2019.

24.   Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

25.   Ilse M, Tomczak J, Welling M, editors. Attention-based deep multiple instance learning. International conference on machine learning; 2018: PMLR.

26.   Zhao L, Bao J, Qiao X, Jin P, Ji Y, Li Z, et al. Predicting clinically significant prostate cancer with a deep learning approach: a multicentre retrospective study. Eur J Nucl Med Mol Imaging. 2023;50(3):727-41. Epub 20221121. doi: 10.1007/s00259-022-06036-9. PubMed PMID: 36409317; PubMed Central PMCID: PMCPMC9852176.

27.   Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555-70. Epub 20210301. doi: 10.1038/s41551-020-00682-w. PubMed PMID: 33649564; PubMed Central PMCID: PMCPMC8711640.

28.   Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. Journal of statistical software. 2017;82(1):1-26.

29.   Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. Distill. 2016;1(10):e2.

30.   Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. BMC medicine. 2010;8(1):1-7.

31.   di Stephano J. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. Functional Ecology. 2003;17(5):707-9.

32.   Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology. 2016;31:337-50.

33.   Lipsey MW. Design sensitivity: Statistical power for experimental research: sage; 1990.

34.   Myors B, Murphy KR, Wolach A. Statistical power analysis: A simple and general model for traditional and modern hypothesis tests: Routledge; 2014.

35.   Chollet F. Deep learning with Python: Simon and Schuster; 2017.

36.   Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn: Machine learning without learning the machinery. GetMobile: Mobile Computing and Communications. 2015;19(1):29-33.

37.   Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:14065823. 2014.

38.   Yamamoto Y, Kiyohara C, Suetsugu-Ogata S, Hamada N, Nakanishi Y. Biological interaction of cigarette smoking on the association between genetic polymorphisms involved in inflammation and the risk of lung cancer: A case-control study in Japan. Oncol Lett. 2017;13(5):3873-81. Epub 20170317. doi: 10.3892/ol.2017.5867. PubMed PMID: 28529598; PubMed Central PMCID: PMCPMC5431513.

39.   Shiels MS, Katki HA, Freedman ND, Purdue MP, Wentzensen N, Trabert B, et al. Cigarette smoking and variations in systemic immune and inflammation markers. J Natl Cancer Inst. 2014;106(11). Epub 20141001. doi: 10.1093/jnci/dju294. PubMed PMID: 25274579; PubMed Central PMCID: PMCPMC4200029.

40.   Lee J, Taneja V, Vassallo R. Cigarette smoking and inflammation: cellular and molecular mechanisms. J Dent Res. 2012;91(2):142-9. Epub 20110829. doi: 10.1177/0022034511421200. PubMed PMID: 21876032; PubMed Central PMCID: PMCPMC3261116.

41.   Letasiova S, Medve'ova A, Sovcikova A, Dusinska M, Volkovova K, Mosoiu C, et al. Bladder cancer, a review of the environmental risk factors. Environ Health. 2012;11 Suppl 1:S11. Epub 20120628. doi: 10.1186/1476-069X-11-S1-S11. PubMed PMID: 22759493; PubMed Central PMCID: PMCPMC3388449.

42.   Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002;21(48):7435-51. Epub 2002/10/16. doi: 10.1038/sj.onc.1205803. PubMed PMID: 12379884.

43.   Lee H-W, Park S-H, Weng M-w, Wang H-T, Huang WC, Lepor H, et al. E-cigarette smoke damages DNA and reduces repair activity in mouse lung, heart, and bladder as well as in human lung and bladder cells. Proceedings of the National Academy of Sciences. 2018;115(7):E1560-E9.

44.   Barbosa AL, Vermeulen SH, Aben KK, Grotenhuis AJ, Vrieling A, Kiemeney LA. Smoking intensity and bladder cancer aggressiveness at diagnosis. PLoS One. 2018;13(3):e0194039.

45.   Yu D, Geng H, Liu Z, Zhao L, Liang Z, Zhang Z, et al. Cigarette smoke induced urocystic epithelial mesenchymal transition via MAPK pathways. Oncotarget. 2017;8(5):8791-800. doi: 10.18632/oncotarget.14456. PubMed PMID: 28060741; PubMed Central PMCID: PMCPMC5352442.

46.   Liu YN, Guan Y, Shen J, Jia YL, Zhou JC, Sun Y, et al. Shp2 positively regulates cigarette smoke-induced epithelial mesenchymal transition by mediating MMP-9 production. Respir Res. 2020;21(1):161. Epub 20200626. doi: 10.1186/s12931-020-01426-9. PubMed PMID: 32586329; PubMed Central PMCID: PMCPMC7318404.

47.   Gu W, Wang L, Deng G, Gu X, Tang Z, Li S, et al. Knockdown of long noncoding RNA MIAT attenuates cigarette smoke-induced airway remodeling by downregulating miR-29c-3p-HIF3A axis. Toxicol Lett. 2022;357:11-9. Epub 20211222. doi: 10.1016/j.toxlet.2021.12.014. PubMed PMID: 34953943.

48.   Wu PH, Gilkes DM, Phillip JM, Narkar A, Cheng TW, Marchand J, et al. Single-cell morphology encodes metastatic potential. Sci Adv. 2020;6(4):eaaw6938. Epub 2020/02/06. doi: 10.1126/sciadv.aaw6938. PubMed PMID: 32010778; PubMed Central PMCID: PMCPMC6976289.

49.   Blanpain C, Mohrin M, Sotiropoulou PA, Passegue E. DNA-damage response in tissue-specific and cancer stem cells. Cell Stem Cell. 2011;8(1):16-29. Epub 2011/01/08. doi: 10.1016/j.stem.2010.12.012. PubMed PMID: 21211780.

50.   Li X, Li J, Wu P, Zhou L, Lu B, Ying K, et al. Smoker and non-smoker lung adenocarcinoma is characterized by distinct tumor immune microenvironments. Oncoimmunology. 2018;7(10):e1494677. Epub 2018/10/06. doi: 10.1080/2162402X.2018.1494677. PubMed PMID: 30288364; PubMed Central PMCID: PMCPMC6169585.

51.   Khan MS, Thornhill JA, Gaffney E, Loftus B, Butler MR. Keratinising squamous metaplasia of the bladder: natural history and rationalization of management based on review of 54 years experience. Eur Urol. 2002;42(5):469-74. doi: 10.1016/s0302-2838(02)00358-5. PubMed PMID: 12429156.

52.   Ahmad I, Barnetson RJ, Krishna NS. Keratinizing squamous metaplasia of the bladder: a review. Urol Int. 2008;81(3):247-51. Epub 20081016. doi: 10.1159/000151398. PubMed PMID: 18931537.

53.   Cosio M, Ghezzo H, Hogg JC, Corbin R, Loveland M, Dosman J, et al. The relations between structural changes in small airways and pulmonary-function tests. N Engl J Med. 1978;298(23):1277-81. doi: 10.1056/NEJM197806082982303. PubMed PMID: 651978.

54.   Jeremy George P, Banerjee AK, Read CA, O'Sullivan C, Falzon M, Pezzella F, et al. Surveillance for the detection of early lung cancer in patients with bronchial dysplasia.

Thorax. 2007;62(1):43-50. Epub 20060706. doi: 10.1136/thx.2005.052191. PubMed PMID: 16825337; PubMed Central PMCID: PMCPMC2111271.

55.   Rausch S, Lotan Y, Youssef RF. Squamous cell carcinogenesis and squamous cell carcinoma of the urinary bladder: a contemporary review with focus on nonbilharzial squamous cell carcinoma. Urol Oncol. 2014;32(1):32 e11-6. Epub 20130220. doi: 10.1016/j.urolonc.2012.11.020. PubMed PMID: 23433891.

56.   Kantor AF, Hartge P, Hoover RN, Fraumeni JF, Jr. Epidemiological characteristics of squamous cell carcinoma and adenocarcinoma of the bladder. Cancer Res. 1988;48(13):3853-5. PubMed PMID: 3378221.

57.   Fleshner N, Garland J, Moadel A, Herr H, Ostroff J, Trambert R, et al. Influence of smoking status on the disease-related outcomes of patients with tobacco-associated superficial transitional cell carcinoma of the bladder. Cancer. 1999;86(11):2337-45. PubMed PMID: 10590376.

58.   Sturgeon SR, Hartge P, Silverman DT, Kantor AF, Linehan WM, Lynch C, et al. Associations between bladder cancer risk factors and tumor stage and grade at diagnosis. Epidemiology. 1994;5(2):218-25. doi: 10.1097/00001648-199403000-00012. PubMed PMID: 8172997.

59.   Caruana R, Lawrence S, Giles L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in neural information processing systems. 2001:402-8.

60.   Tan M, Le Q, editors. Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning; 2019: PMLR.

61.   Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ. 2013;346:e8668. Epub 2013/01/23. doi: 10.1136/bmj.e8668. PubMed PMID: 23338004; PubMed Central PMCID: PMCPMC4688419.