AI in Precision Oncology
Volume 00, Number 00, 2024
© Mary Ann Liebert, Inc.
DOI: 10.1089/aipo.2024.0015

AI IN PRECISION ONCOLOGY

Open camera or QR reader and
scan code to access this article
and other resources online.

SPECIAL ISSUE: THE FUTURE OF CANCER SCREENING

An Ensemble AI Model for RET Alteration Detection Using H&E Images as a Putative Screening Tool for More Efficient Genomic Alteration Detection

Krishna Bairavi,[1,†] Bobak Kechavarzi,[2,3,†] Arkan Abadi,[2] Xinying Mu,[1] Kelly M. Credille,[2] Narasimha Marella,[2] Nino Sireci,[2] Michael D. Mathews,[2] Ross Grinvalds,[2] Robert Klopfleisch,[4] Thomas Colarusso,[1] Oscar Puig,[2,5] Thomas W. Chittenden,[6,7] Andrea De Souza,[8,*] and Alan Jerusalmi[1,*]

Abstract

RET-activating gene alterations are present in 1%–2% of non-small cell lung cancers. Therapeutics that specifically and effectively target these *RET* alterations have recently been approved. Broad-based genomic testing, inclusive of *RET* fusions, is recommended by National Comprehensive Cancer Network (NCCN) and ASCO/AMP/CAP guidelines for patients with advanced non-small cell lung cancer, but screening patients for such rare biomarkers in drug development can be impractical and costly. Here, we develop and validate a deep neural network pipeline to detect *RET* alterations from readily available hematoxylin and eosin (H&E)-stained images. As the pipeline is intended for prescreening and sample prioritization for genomic testing for *RET* fusions during drug development, 100% sensitivity was a primary objective to ensure that no *RET* fusion-positive samples were missed. In total, 523 images were used for model development and partitioned for training (70%), validation (15%), and testing (15%). The approach resulted in 100% sensitivity and 72.4% specificity, corresponding to an area under receiver operating characteristic curve (AUROC) of 0.86 on the test set. An additional dataset of 121 images was used for an independent blind assessment. The overall sensitivity of the model on the second independent dataset was 100% with a 63.3% specificity and an AUROC of 0.82. All 20 *RET* fusion-positive cases in this dataset were correctly detected with no false negative cases and 36 false positive cases in the blind dataset. These findings suggest deep learning can be used as a complementary method to prescreen H&E-stained images and enhance the rate of *RET* alteration positivity in subsequent genomic testing.

Keywords: NSCLC, RET alteration, H&E, machine learning, deep learning, artificial intelligence

[1]BioAI Health, Goffstown, New Hampshire, USA.
[2]Eli Lilly, Indianapolis, Indiana, USA.
[3]Currently Cleveland Clinic, Cleveland, Ohio, USA.
[4]Institute of Veterinary Pathology, Freie Universitaet, Berlin, Germany.
[5]Currently Nucleai, Chicago, Illinois, USA.
[6]Digital Environment Research Institute, Queen Mary University of London, London, UK.
[7]BullFrog AI, Gaithersburg, MD, USA.
[8]Consulting for IonQ and Qubit Pharmaceuticals, USA.
[†]These authors contributed equally to this work.

*Address correspondence to: Alan Jerusalmi, PhD, BioAI Health, 105 Maple Avenue, Goffstown, NH 03045, USA, E-mail: ajerusalmi@bioaihealth.com; Andrea De Souza, MA, MBA, Formerly Eli Lilly, Indianapolis, Indiana, USA. Formerly Eli Lilly, currently consulting for IonQ and Qubit Pharmaceuticals, E-mail: andrea.desouza@sloan.mit.edu

Introduction

Lung cancer is one of the deadliest cancers with over 75% of patients losing their battle within 5 years of diagnosis.[1] In 2020, there were 2.21 M new cases of lung cancer which represented 11.4% of all new cancers and was second only to breast cancer. However, the mortality rate of breast cancer was 6.9% with 685 K patients compared to 18% or 1.80 M patients for lung cancer.[2] More people succumb to lung cancer than to colon, breast, and prostate cancers combined.[3] Genomic alterations in several genes including *KRAS*, *EGFR*, and *ALK* can drive the development and progression of non-small cell lung cancer (NSCLC). These genomic alterations occur with a prevalence of 15%–25%, 5%–15%, 3%–7%, respectively.[4–6] Specific genomic drivers can influence therapeutic decision-making by predicting treatment response. For example, genomic alterations in *EGFR* are strongly predictive of favorable responses to *EGFR*-targeted therapies in NSCLC.[7] More recently, *RET* has emerged as an oncogenic driver in NSCLC.[8–11]

RET is a glycoprotein receptor with tyrosine kinase activity whose activation via autophosphorylation triggers downstream cell proliferation and survival pathways including RAS-MAPK, PI3K-AKT, JAK-STAT, PLC-gamma, and PKC.[8,10,12,13] The *RET* gene, located on the long arm of chromosome10 (10q11.21), is subject to gain-of-function gene-fusions through rearrangements that result in constitutive receptor activation.[8–10,14,15] *RET* gene fusions are detected in 5%–10% of papillary thyroid cancer (PTC) and 1%–2% of NSCLC among others.[8–11,13,16] *RET* fusions do not tend to co-occur with other major NSCLC driver alterations (e.g., *KRAS* or *EGFR* mutations, *ALK* or *ROS1* rearrangements) and are associated with low tumor mutation burden and decreased expression of PD-L1.[17,18] In 2020, two selective *RET* inhibitors selpercatinib and pralsetinib, received FDA approval for metastatic *RET* fusion-positive NSCLC, advanced or metastatic *RET* fusion-positive thyroid cancer requiring systemic therapy that are radioactive iodine-refractory, and locally advanced or metastatic *RET* fusion-positive solid tumors.[16,19–21] A recent study of selpercatinib in patients with *RET* fusion-positive tumors (open label) showed a median PFS was 22.0 months in treatment-naïve patients, 35% of whom were alive and progression-free at the data cutoff (median follow-up of 21.9 months).[22]

Genomic testing allows for tailored therapeutics based on a patient's specific tumor molecular profile.[23,24] For these reasons, National Comprehensive Cancer Network (NCCN) and ASCO/CAP/AMP guidelines recommend broad-based genomic testing of all patients with metastatic NSCLC to identify actionable mutations. However, during clinical development, the cost of genomic profiling of NSCLC presents a significant barrier to patient screening and enrollment, particularly when the frequency of targetable genomic modifications is low (e.g., *RET* fusions 1%–2%) or in earlier stage disease where routine NGS testing is uncommon.[23,24] Recent advancements in artificial intelligence to rapidly detect genomic alterations from histological images are a viable prescreening tool to overcome these barriers. Investigators leveraged readily available patient datapoints such as hematoxylin and eosin (H&E)-stained slides from biopsies to detect genomic alterations and build molecular profiles in breast, liver, and colorectal cancers.[25–28] In NSCLC, a deep learning approach was implemented to detect genomic alterations in common oncogenic drivers such as *KRAS*, *EGFR*, *STK11*, *FAT1*, *SETBP1*, and *TP53* with area under receiver operating characteristic curve (AUROC) ranging from 0.73 to 0.86.[29] Mutations in *SPOP*, an established tumor suppressor gene, can be predicted from H&E samples in prostate cancers using pretrained ResNET models.[30–32] These results support the deployment of lower cost algorithmic approaches to prescreen and prioritize patients for more expensive genomic testing in a clinical development setting.

In this work, we designed and developed a novel workflow comprised of deep learning models and algorithms for processing and classification of H&E-stained histopathology images in a NSCLC cohort to classify a small subpopulation of patients with *RET* fusions. This unique approach to image normalization and processing diminishes the impact of biases from digitization artifacts, tissue preparation, and additional confounders. In the datasets evaluated, the predictive model achieved 100% sensitivity while maintaining greater than 50% specificity, prioritizing at-need patients while reducing costs for comprehensive screening of patients for clinical trial sponsors.

Methods

Dataset

In total, 621 digital H&E NSCLC images were selected for use in this project from two clinical trials: Libretto-001 (NCT03157128) and Libretto-431 (NCT04194944).[19,33] Trial participants included men and women with both primary and metastatic NSCLC disease. Current or former smokers were not excluded and additional detail on patient demographics, prior treatments, and disease stage can be found in earlier publications.[19,33] Consent documents were carefully reviewed to ensure that data utilization was consistent with the terms therein. Stained slides were initially prepared and digitized across four contributing data sites and underwent manual quality control for inclusion. Formalin-fixed core needle and tumor resection samples of NSCLC were routinely processed to paraffin block, sectioned at 5-micron thickness, stained with H&E-stained slides, and scanned on the Leica AT Turbo or the 3D Histech Pannoramic P1000 scanning system with magnification at 20× or 40×. Scanned images were stored in SVS or MRXS format, respectively, and uploaded to a secure online storage location. Images were initially manually inspected by a trained pathologist for histology and scanning quality. Images of thick sections, insufficient tumor, all necrosis, or extensive out of focus areas were excluded. The *RET* fusion status of these samples was determined in a certified laboratory with the use of NGS, fluorescence in situ hybridization, or polymerase-chain-reaction assay.

All 266 images from Libretto-001 were exclusively *RET* fusion-positive as they were collected from enrolled participants only, whereas the 355 images from Libretto-431 included both *RET* fusion-positive and -negative samples as they were collected during screening for the trial. The images were also collected from primary and metastatic samples (Supplementary Table S1). The trial image dataset was partitioned into two datasets, a dataset of 500 samples which was initially used for model development and a second dataset of 121 images that was held back for blind evaluation of model performance. Several factors were considered when partitioning the trial images: all 266 Libretto-001 were SVS format, *RET* fusion-positive, and allocated exclusively to the model development dataset, whereas 355 images from Libretto-431 were split across the model development or blind evaluation dataset balancing for source (primary, metastatic, or unknown), *RET* fusion status, and data site. As a result, the 500-image model development dataset was sourced from 215 metastatic, 236 primary, and 49 unknown samples; contained 284 *RET* fusion-positive and 216 *RET* fusion-negative samples; and included 111 MRXS and 389 SVS image types. Please note that the MRXS images were only used for tumor model development, whereas SVS images were used for both tumor model and RET model development. The 121-image blind dataset was sourced from 47 metastatic, 54 primary, and 20 unknown samples; contained 20 *RET* fusion-positive and 101 *RET* fusion-negative samples; and included-121 SVS image types. The trial image datasets are detailed in Supplementary Table S1. The model development dataset was supplemented with an additional 23 images from the NCI Genomic Commons from the Cancer Genome Atlas-Lung Adenocarcinoma (TCGA-LUAD) cohort of images.[31,32] This brought the total images in the model development dataset to 523. During model development, this dataset was further divided into a training set (70%), validation set (15%), and test set (15%) as described in Figure 1.

Pathology image annotations and manual segmentation review

Further histopathologic annotations were performed by a board-certified expert pathologist who also conducted a secondary general quality check of the cases. This included confirmation of appropriate staining, scanning quality, and confirmation of the presence of lung tissue and metastasis. No images were excluded after the secondary quality assessment.

Pathologist image annotations were performed in three stages. The first stage included the annotation of regions in the TCGA-LUAD cohort to build the underlying tumor segmentation models for tumor and nontumor regions. The second manual annotation stage was incorporating those from the same regions in the trial set of images of SVS format. A final annotation round was necessary to adapt the segmentation model to the MRXS image file format. The quality of the tumor segmentation model was visually assessed by our pathology team. Supplementary Figure S1 contains examples of images that were annotated and used for development of tumor segmentation model. Annotations included the following labels: "tumor," "nontumor,"

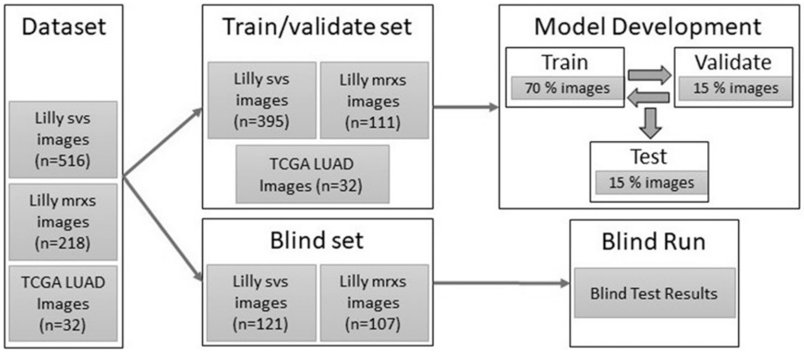


Fig. 1. Data utilization and split. In total, 728 digital H&E NSCLC images were used for this project. The initial model development dataset consisted of 500 images which were used for model development and an additional 121 images were held back for a blind assessment. Bio-AI Health further divided the model development dataset set of images used for model development into a training set (70%), validation set (15%), and test set (15%). H&E, hematoxylin and eosin; NSCLC, non-small cell lung cancer.

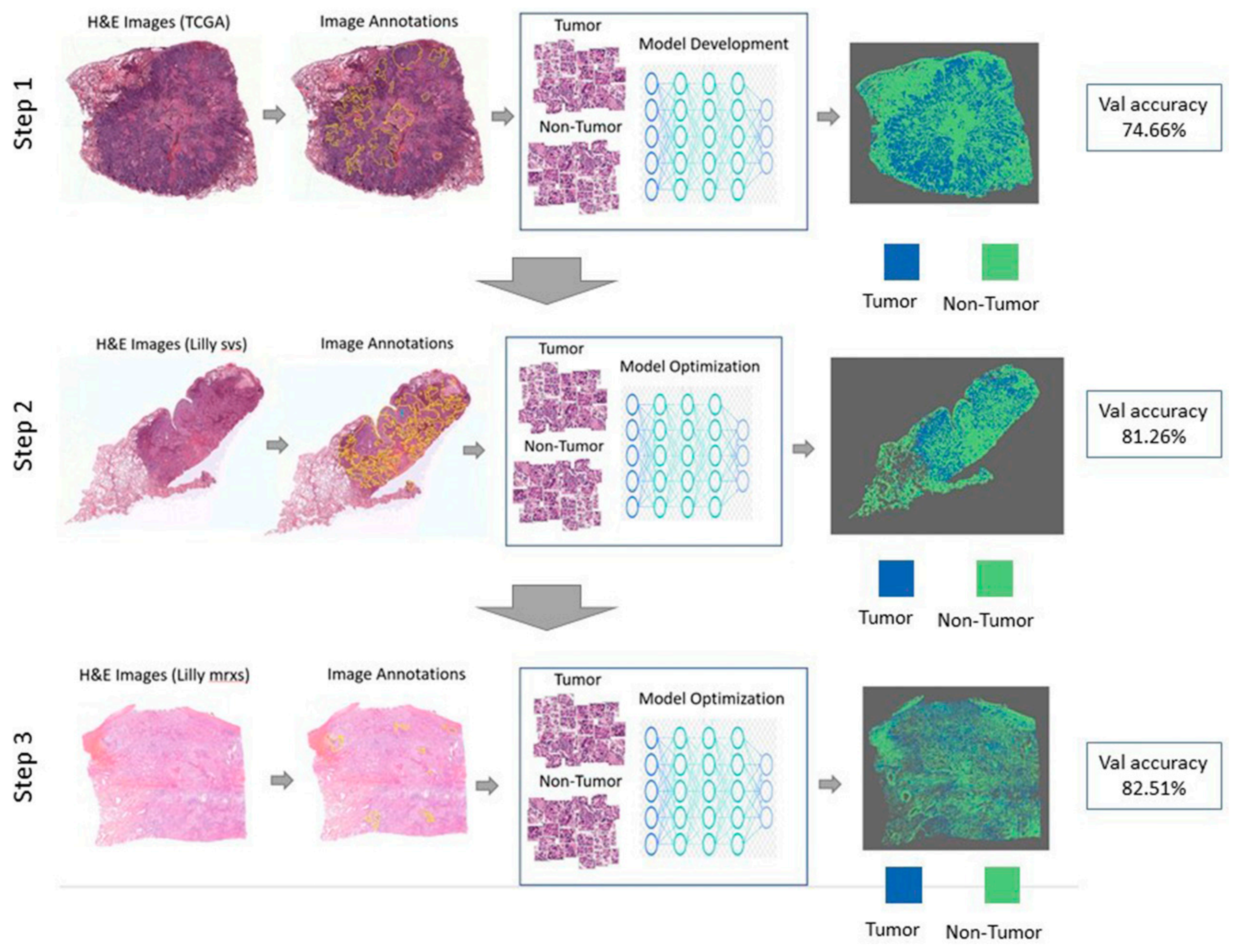


Fig. 2. Tumor Segmentation model workflow. Development of a NSCLC tumor segmentation model was constructed in three steps. Initial steps used TCGA-LUAD images and corresponding annotations and yielded a validation accuracy of 74.66%. Step 2 added annotations from clinical trial Leica based images which resulted in a validation accuracy of 81.26%. Step 3 added 3D Histech images and annotations and saw a final validation accuracy on model of 82.51%. NSCLC, non-small cell lung cancer; TCGA-LUAD, Cancer Genome Atlas-Lung Adenocarcinoma.

"ignore," and were incorporated into PREDICT-X platform and used during model development.

Tumor recognition and segmentation

The tumor segmentation model used for selection of tumor regions in images was developed using a three-step approach (Fig. 2). Step 1 involved the optimization of a model previously developed for tumor segmentation using TCGA-LUAD data only. Briefly, a VGG19 architecture was used to generate a prediction label on tiles sized 256 × 256 with a binary classification of tumor vs nontumor.[34] This deep learning model includes 147 million parameters and 26 layers. The last layer was a softmax, where the output of the model for one tile was a vector with two probability scores corresponding to each of the classes. The class with the highest score was used as the predicted label for the tile. In step 2, additional annotated regions from trial cohort (SVS format only) images were used to optimize the network for the metastatic biopsy tissue. Finally, step 3

Table 1. Detailed Metrics of Image Tiles Used for Optimization and Development of the Tumor Segmentation Model

	Region	*Train*	*Validation*	*Test*
Step 1 (TCGA-LUAD only)	Tumor	10,806	5,510	5,003
	Nontumor	10,748	5,533	4,975
Step 2 (Trial SVS)	Tumor	41,651	20,687	23,056
	Nontumor	4,204	2,894	2,804
Step 3 (Trial MRXS)	Tumor	6,937	2,000	2,000
	Nontumor	6,851	2,000	2,000

TCGA-LUAD, Cancer Genome Atlas-Lung Adenocarcinoma.

involved addition of annotated patches from the MRXS scanner to account for the scan-specific differences in the file format. Table 1 summarizes the number of tiles used in each phase of development during each step.

All steps involved a similar methodology for model optimization and development except for image sources. Briefly, images were partitioned into tiles and used for the optimization of a previously developed convolutional neural network (CNN) model on PREDICT-X platform for NSCLC tumor detection. Pathology-labelled tile regions were split into a training set, a validation set, and a test set (70% train, 15% validation, and 15% test). Background tiles were easily selected and excluded based on a color intensity threshold 220. The predicted probabilities of image tiles were summarized into a heatmap of tumor probability, where each pixel in the heatmap corresponded to an image tile in the original pathology image. The results were also visually inspected and evaluated by an anatomical pathologist. Once an accuracy of >80% was achieved and a satisfactory performance report was given by a certified pathology visual inspection, tiles were saved and used for subsequent development.

Tile image QC and normalization

The PREDICT-X platform contains a previously developed QC model that comprises of a modified version of ResNet, which excludes unwanted tiles based on specific content.[35] We further optimized this model against the clinical trial dataset to detect and exclude image artifacts that can have a negative impact on the *RET* fusion status prediction model. This optimization included tile sets labelled as TAR (tiles that contain anthracotic pigmentation) and RBC (tiles that contain numerous red blood cells) to detect these problematic tiles for exclusion from downstream steps. Figure 3 shows examples of tiles that were used for further training the QC model and Table 2 describes the amount of data used for training and optimization of model. Each image had between 0% and 5% of total tiles excluded from the QC model which had an overall performance accuracy of 0.95.

To overcome the variability in staining and image processing from multiple data sites and scanners, the Reinhard Color Normalization approach was applied.[36] The technique was initially used in traditional computer vision problems and was adapted for color correction and normalization in H&E-stained digital histopathology slides.[37,38] The approach uses a reference images color profile to transform a source image. As a middle ground, the source image, which is comprised of RGB channels, is converted to $l\alpha\beta$ color space proposed by Ruderman et al., using set of linear transformations.[39] This aids in matching the mean and standard deviation of the two images (source and reference image) in the $l\alpha\beta$ color space. Below is a set of equations which converts the

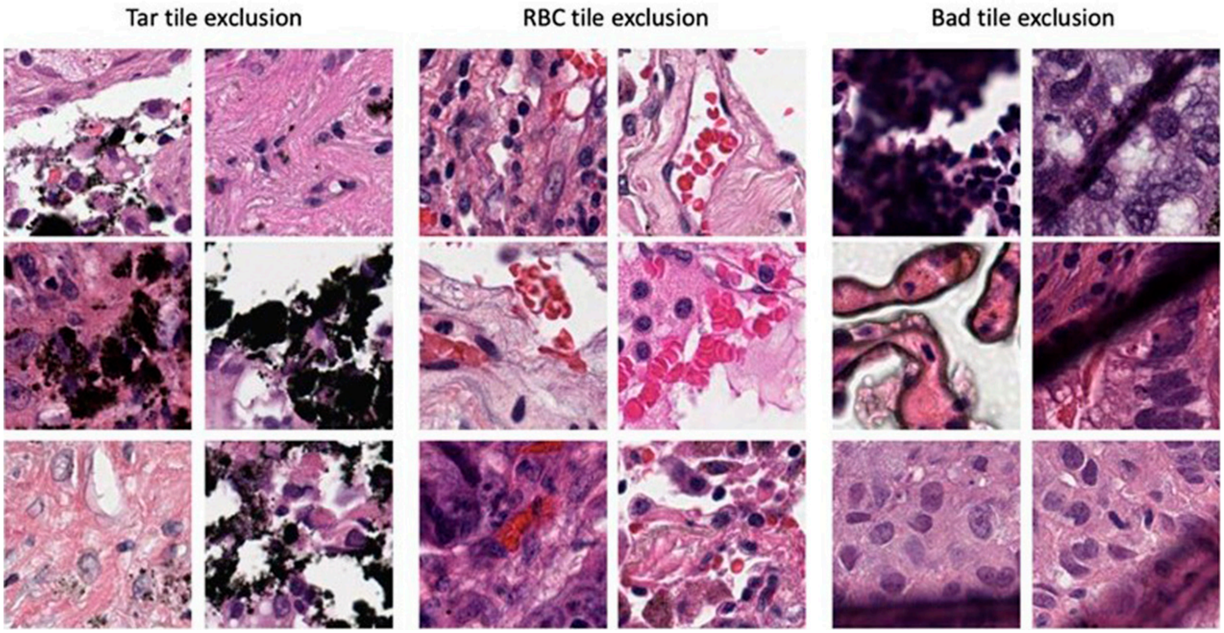


Fig. 3. Pathology Annotations—Representative examples of images that were annotated and used for development of tumor segmentation model. Annotations included the following labels: “tumor,” “nontumor,” “ignore,” and were incorporated into Bio-AI Health platform and used during model development.

Table 2. Detailed Metrics of Image Tiles Used for Optimization and Development of the Predict-X QC Model

Tile label	*Train*	*Validation*	*Test*
Good	3,731	799	799
TAR	1,175	252	252
RBC	3,097	663	663

RBC, tiles that contain numerous red blood cells; TAR, tiles that contain anthracotic pigmentation.

original RGB channel source image into $l\alpha\beta$ color space:

$$l_{mapped} = \frac{l_{original} - \bar{l}_{original}}{\hat{l}_{original}} \hat{l}_{target} + \bar{l}_{target} \tag{1}$$

$$\alpha_{mapped} = \frac{\alpha_{original} - \bar{\alpha}_{original}}{\hat{\alpha}_{original}} \hat{\alpha}_{target} + \bar{\alpha}_{target} \tag{2}$$

$$\beta_{mapped} = \frac{\beta_{original} - \bar{\beta}_{original}}{\hat{\beta}_{original}} \hat{\beta}_{target} + \bar{\beta}_{target} \tag{3}$$

Supplementary Figure S2 contains representative examples of a tile image before normalization and after normalization along with the reference tile image used for the transformation.

RET model

The PREDICT-X platform inspired by Neural Architectural Search (NAS) was used to build a deep learning probability model to determine *RET* fusion status in digital H&E slides.[40] NAS does a search across different architectures on a small subsample of data and chooses the right one for model building on the large sample set. The PREDICT-X platform was used as a pretrained model to further optimize for *RET* fusion status prediction. Because of data being generated from multiple sources and imbalances between *RET* fusion-positive and *RET* fusion-negative cases, we used an ensemble classification approach with different combinations of cases as well as different CNN models to find the optimal model for classifying the *RET* fusion status. K-fold cross validation was applied. In addition, image augmentation techniques were applied to up-sample the data. This included rescaling, horizontal, and vertical flipping, zooming in and out of random images, varying brightness intensity, and shifting image width and height. In total, over 30 models were developed. These models were developed and tested using different convolutional neural network architectures as well as different optimization parameters. The composition of the training and validation data was determined after an extensive analysis of the available data and interpretation of interim models. In the end, two separate models were retained using two separate deep neural networks and an ensemble approach generated the final prediction. We used a model ensemble strategy where we trained on more positive samples from multiple locations for the first model and for the second model, we trained it on more negative samples. We used NASnet architecture for both and later assigned percentage weightage to get the right prediction.[40] Each CNN model was trained and developed on image tiles independently of each other and combined to generate the final prediction on a patient level. The prediction is generated by aggregating tiles within a patient image using a positivity threshold of 0.4 to classify each case as *RET* fusion-positive or -negative. Prediction results from the blind test set are described in detail for the ensemble model. The final threshold was determined after extensive analysis of all cases in model development dataset to eliminate all false negatives at the expense of false positives. Model performance metrics such as AUROC, which plots the relationship between true positive rate and false positive rate across different predictive thresholds, were used to determine the model quality. Because of lack of balance in the MRXS training set, and the inherit differences noted on the images when compared to SVS images, the MRXS dataset was only used for development of a tumor model. Development of the *RET* predictive model was limited to the SVS images only. There were not enough MRXS balanced images to develop a MRXS-specific model as only five images in the set were positive.

Results

The BioAI PREDICT-X platform (a secure statistical machine learning and data management environment that houses a proprietary high-performance computing workflows that automate multimodal data processing to develop, deploy, and optimize the latest AI strategies) was chosen for this study after outperforming two other proprietary software platforms that failed to positively detect cases in a blinded preassessment. Briefly, a dataset of 500 NSCLC H&E-stained images produced from samples collected from the Libretto-001 (NCT03157128) and Libretto-431 (NCT04194944) clinical trials was assembled for model development in this study.[19,33] This dataset was supplemented with 23 images from TCGA-LUAD dataset available through NCI Genomic Commons at https://gdc.cancer.gov/. Trial images included primary, metastatic, and unknown tissue biopsies from four different data sites and two image types (See Supplementary Table S1 and Methods).[19,33] The images were generated using Leica Aperio Scanscope AT whole slide scanner platform, which generates a virtual slide file with a ".svs" extension (SVS, $n = 389$) as well as the 3D Histech whole slide scanner which generates a virtual slide file with a ".mrxs" extension (MRXS, 111). These images were produced at four different anonymized locations. The distribution of images across *RET* fusion status, data sites, and image types is shown in Supplementary Figure S3. All TCGA-LUAD images were scanned on the Leica Aperio Scanscope platform (i.e., SVS) and contained a mix of *RET* fusion-positive and negative samples. In addition, a separate set of 121 images from the Libretto-431 trial was held out of the model development dataset and used only as a blind evaluation dataset. This hold-out dataset included primary (54), metastatic (47), and unknown (20) tissue biopsies in all 121 (SVS) image types.

The overall PREDICT-X workflow used to develop the *RET* classifier is illustrated in Figure 4A. First, images were assessed with an automated tumor segmentation model to select for tumor positive regions within tissue. Segmented tumor tissues were tiled with a deep learning model to specifically eliminate tiles containing artifacts that negatively impacted model development (see Methods). Next, tumor tiles underwent a color-normalization process before the final step of *RET* classification. Multiple models were developed and optimized for each of these major steps and iterated on annotations, as described in more detail below. Figure 4B summarizes this workflow.

H&E-stained images, stored as SVS and MRXS, showed clear differences in size, pixel resolutions, and spectral properties (Fig. 5). Although pixel size and other visually detectable differences between the image formats can adversely influence model performance, color normalization was used to correct for spectral differences. The distribution of *RET* fusion-positive and *RET* fusion-negative images (ground truth) in the model development dataset was 72%/28% for the SVS images and 5%/95% for the *MRXS* images, respectively. Subsequently, MRXS images were only used during the tumor model development phase of the project as the balance of positive and negative *RET* fusion samples does not allow for predictive model development (Supplementary Fig. S3).

From the TCGA-LUAD cohort, in total 1,606 annotated areas covering 24,001,164 μm^2 of tissue were assessed, including 806 tumor annotations (12,570,264 μm^2) and 800 nontumor annotated regions (11,530,900 μm^2). Training images from the trial dataset yielded an additional 5,781,162,847 μm^2 of annotated regions. Further insights during model development showed that additional nontumor area was necessary to increase the recall/precision for tumor detection. Therefore, additional 108 annotations covering 155,766,722 μm^2 were performed. These annotations were used for the identification and selection of tumor tiles for subsequent *RET* model development.

Figure 2 summarizes the three-step development of the NSCLC tumor segmentation model. The goal of segmentation was to annotate tiles from the H&E images, which initially yielded a validation accuracy of 74.66%, 81.26% following step 2, and finally reaching 82.51% after step 3. In addition, only individual tiles that were selected as tumor tiles with probability score of >0.9 were used for *RET* model development and analysis. Supplementary Figure S4 shows representative heatmap images with

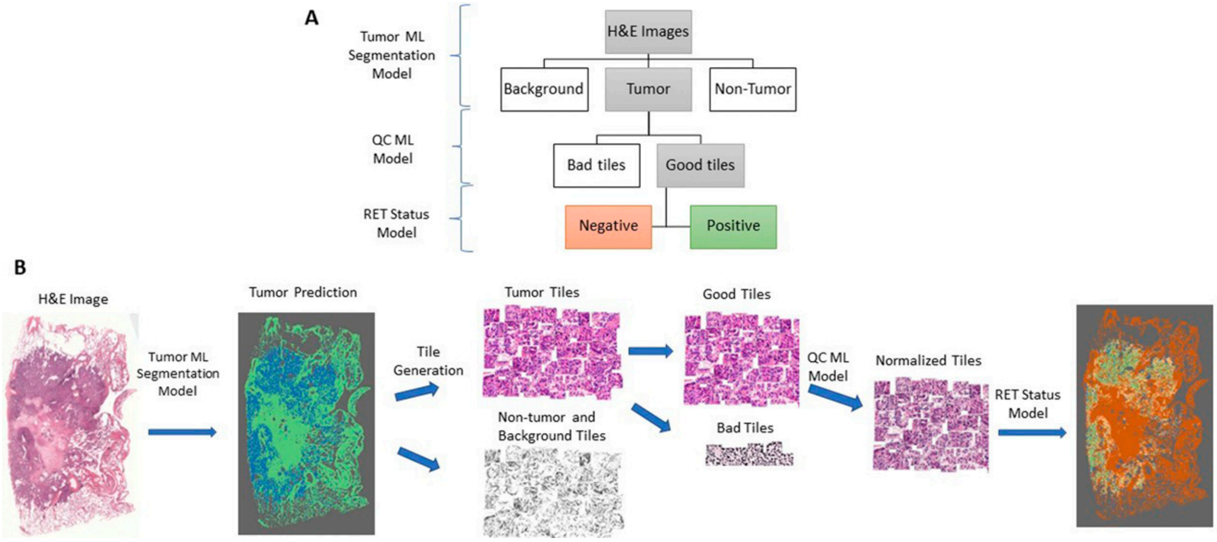
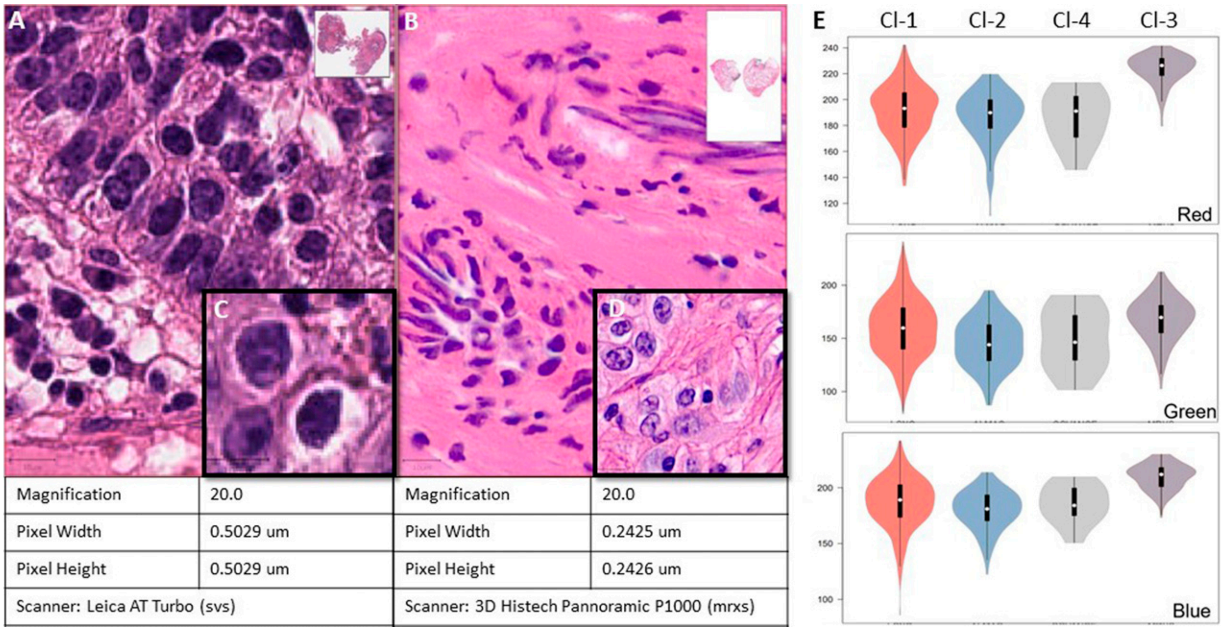


Fig. 4. **(A)** Overall workflow used to develop the *RET* fusion status classifier. Briefly, images were segmented into their tumor components and corresponding tiles are generated. The tumor tiles were then cleaned of any artifacts that had a negative impact on model development and color normalized before selection of *RET* fusion status. **(B)** Summarizes the analysis performed demonstrating a single image workflow for demonstration purposes.

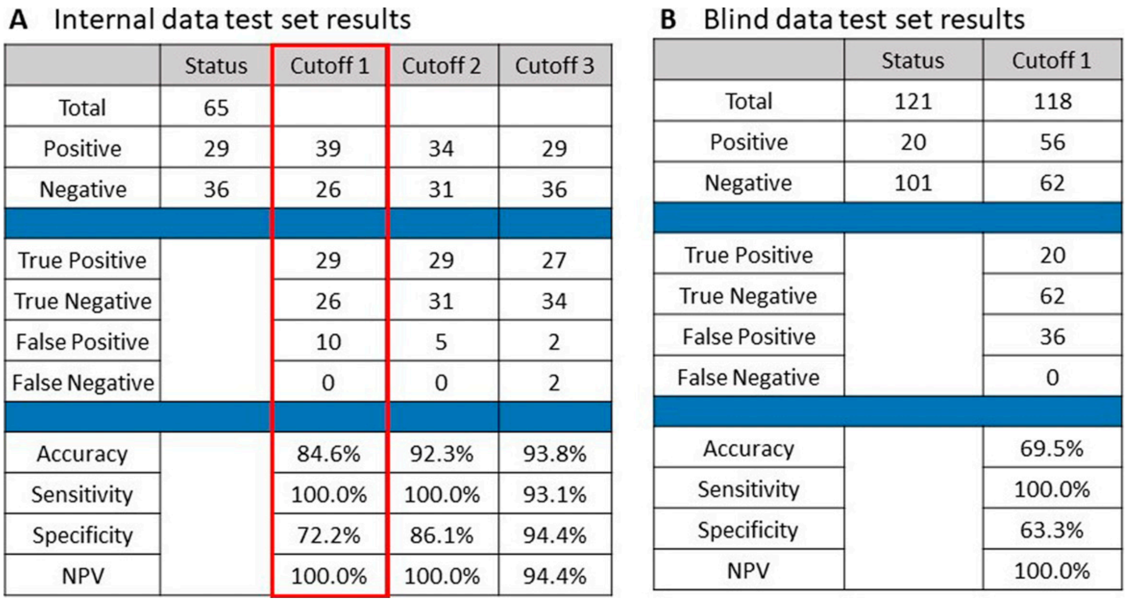


Magnification	20.0	Magnification	20.0
Pixel Width	0.5029 um	Pixel Width	0.2425 um
Pixel Height	0.5029 um	Pixel Height	0.2426 um
Scanner: Leica AT Turbo (svs)		Scanner: 3D Histech Pannoramic P1000 (mrxs)	

Fig. 5. Observed differences between Leica and 3D Histech scanned images. Aside from significant spectral variances between the formats as demonstrated on **(A)** Leica and **(B)** 3D Histech images. In addition, there is a measurable pixel resolution difference between **(C)** Leica and **(D)** 3D Histech formats. **(E)** Plots highlight the spectral differences observed between the separate data sites (Cl-1, Cl-2, Cl-3, Cl-4) where images were obtained. Plot shows distribution of red, green, and blue channels for each tissue in development set. Note largest contrast between data sites which used MRXS format.

corresponding tumor tiles. A high-resolution automated QC tool was developed, which identifies artifacts and abnormalities in tumor tiles that compromise model training and performance. Once tumor tiles were selected by the segmentation model, they were further filtered using the QC tool (see Methods). This resulted in the removal of approximately 5% of tiles from each case owing to abnormality detection. Supplementary Figure S5 shows the results and performance metrics of the QC tool. A high degree of spectral variability was observed on images owing to variability across the data sites from which the images were sourced. To correct for this variation, we normalized all tumor tiles using the Reinhard Color Normalization approach.[36] Representative images before and after normalization are shown in Supplementary Figure S2.

Because of observed differences in RGB channels between MRXS- and SVS-based images (Fig. 5) and challenges combining these image types, *RET* fusion predictive model development was focused on the SVS set of images only. Furthermore, the limited number and diversity of *RET* fusion-positive images in the MRXS format (only five and all images were from data site 3) meant that developing and validating a *RET* fusion status detection model specific to MRXS was not possible. Therefore, the MRXS images, while useful for developing the tumor segmentation model, were not considered further for the *RET*-alteration prediction model. The PREDICT-X platform used a subsequent ensemble classification approach combining 2 different CNN models. The first model was trained to specifically detect the positive *RET* fusion signal whereas the second model was trained to detect the negative cases. The ensemble approach was powered to achieve a 100% sensitivity to ensure that no positive cases were missed. A conservative classification threshold of 0.4 was used to classify positive cases, resulting in a higher number of false positives while ensuring no false negatives were observed. The overall results of the initial test set are presented in Figure 6A. This strategy resulted in a measured sensitivity of 100%, a specificity of 72.4%, an AUROC of 0.86 (Supplementary Fig. S6A), and a corresponding balanced accuracy of 72.7%. These results allowed us to end development and deploy the finalized pipeline on the blind evaluation dataset. The blind dataset consisted of 121 SVS images (101 *RET* fusion-negative, 20 *RET* fusion-positive). Although three cases were excluded from analysis as not enough tumor tiles were detected by the model, similar performance was achieved on the SVS images from blind dataset (Fig. 6B). The overall sensitivity of the SVS images from the blind dataset was 100% with a specificity of 63.3%, and an AUROC of 0.82 (Supplementary Figure S6



A Internal data test set results

	Status	Cutoff 1	Cutoff 2	Cutoff 3
Total	65			
Positive	29	39	34	29
Negative	36	26	31	36
True Positive		29	29	27
True Negative		26	31	34
False Positive		10	5	2
False Negative		0	0	2
Accuracy		84.6%	92.3%	93.8%
Sensitivity		100.0%	100.0%	93.1%
Specificity		72.2%	86.1%	94.4%
NPV		100.0%	100.0%	94.4%

B Blind data test set results

	Status	Cutoff 1
Total	121	118
Positive	20	56
Negative	101	62
True Positive		20
True Negative		62
False Positive		36
False Negative		0
Accuracy		69.5%
Sensitivity		100.0%
Specificity		63.3%
NPV		100.0%

Fig. 6. **(A)** The overall results of the internal test set where approach resulted in a measured sensitivity of 100%, specificity of 72.2%, and overall accuracy of 84.6%. **(B)** Performance metrics for the held back blind data set. It is important to note that three cases were excluded from analysis as not enough tumor tiles were detected by the model. Measured sensitivity remained at 100%, specificity was 63.3%, and overall accuracy of 69.5%.

images tested, all positive cases were correctly detected with no false negatives and 36 false positives). Furthermore, model performance was comparable between primary and metastatic lesions.

Discussion

Advances in screening and targeted treatment approaches ranging from low dose CT, genomic testing (*KRAS*, *ALK*, *EGFR*), and novel therapeutics will be critical to altering the course for lung cancer. Genomic screening can be costly and may pose a challenge to trial sponsors developing targeted therapies for relatively small populations. Here we propose leveraging deep learning methodologies to triage samples for drug development efforts using readily available data for more extensive screening with no further sample reduction. This can enable drug development for rare alterations, lowering the cost and sample quantity barriers to entry.

In this study, we developed a novel workflow for the processing and classification of H&E-stained histopathology images in a NSCLC cohort to predict a subset of patients with *RET* alterations. This unique computational approach to image normalization and processing diminishes the impact of biases from digitization artifacts, tissue preparation, and additional confounders. As the objective was to use this AI as a prescreening tool, before genomic testing for trial enrollment, sensitivity was prioritized, to ensure that no *RET* fusion-positive cases were missed. The ensemble classification model that was ultimately selected did meet this objective, achieving 100% sensitivity while maintaining >60% specificity. Considering the rate of *RET* fusion-positivity in NSLC is 1%–2%, prescreening and prioritizing patients to achieve >60% *RET* fusion-positivity will ultimately decrease the cost for genomic testing in future drug development efforts (e.g., for every 100 patients sequenced currently only 1–2 return *RET* fusion-positive status, but with AI prioritization this number would increase to >60 while ensuring that no RET fusion-positive patients are not sequenced due to 100% sensitivity).

It is important to note the focused use case of this technology in the context of drug development that targets rare genetic alterations. In clinical practice, pathologists and oncologists must test advanced lung cancer patients for a multitude of genomic and proteomic markers, that are relevant to their treatment decisions. *RET* fusion status alone, regardless of outcome, would not obviate the need for comprehensive screening in clinical practice as the patient may harbor other targetable biomarkers in this disease. Instead, the value of a prescreening predictor such as this is to enable screening for rare biomarkers in the drug development space, where genomic sequencing for the biomarker is not yet a clinical priority but is of utmost importance for drug development. This study focused strictly on the tumor region of tissue and future studies that include the tumor microenvironment could explore whether inflammation, infiltrative growth patterns, and tumor/stroma regions improve model performance.

One limitation of the approach is that the model could only be applied to SVS images. In this study, all MRXS images came from the same data site and there were only 5 *RET* fusion-positive cases out of the total 111 MRXS images in the dataset. Therefore, incorporating richer datasets with a greater number of *RET* fusion-positive samples from the MRXS image type are needed before AI models can be fully developed to detect *RET* fusion status. Another limitation of this study is that the dataset that was withheld from model development for blind testing was drawn from the same collection of images as the model development dataset because of the need to balance *RET* fusion status among datasets. Therefore, the performance assessment in the blind dataset is likely optimistic and the robustness of the model needs to be examined in new data from independent sources. Relatedly, the rate of *RET* fusion positivity in both the model development and blind assessment datasets was much higher than in the general population; this makes it difficult to assess how the model will perform in a real-world setting *a priori*.

Moreover, this work adds *RET* to the growing list of gene alterations that can be detected from H&E stained images, including *KRAS*, *EGFR*, *STK11*, *FAT1*, *SETBP1*, and *TP53* in NSCLC.[29] Detection of chromosomal rearrangements in H&E stained lung cancer has been less studied but our results indicate this is possible as well.[41] Future innovations using interpretable machine learning may lead to novel biological insights into the physical, cellular, and morphological changes that underlie the algorithmic detection of such alterations. It is important to consider that the datasets used in this model were limited and expanding the scale and diversity of data for training and validation will improve our understanding of the generalizability and limitations of this approach in the real-world. This work supports the feasibility of *RET* fusion screening with AI and provides proof-of-concept for how such models can be developed to detect rare genetic alterations. Using AI-based models during initial screening could speed up decisions for both patients and drug developers as well as lower testing costs and tissue use. However, this model was not tested or validated on out-of-distribution data and this is necessary to ensure robustness against myriad sources of bias such as diverse imaging platforms, sampling, and staining protocols that are introduced in the real world. Challenges remain when applying such models to new datasets and wide adoption in practice.

Acknowledgments

The results here are in part based upon data generated by the cancer genome atlas. Research Network: https://www.cancer.gov/tcga. The authors would like to thank Loura Berlen and Ania Szpurka for their assistance in reviewing trial consent documents.

Data Availability Statement

Eli Lilly and Company provides access to all individual data collected during the trial, after anonymization, with the exception of pharmacokinetic, genomic, or genetic data. Data are available to request 6 months after the indication studied has been approved in the United States and European Union and after primary publication acceptance, whichever is later. No expiration date of data requests is currently set and will be set once data are made available. Access is provided after a proposal has been approved by an independent review committee identified for this purpose and after receipt of a signed data sharing agreement. Data and documents, including the study protocol, statistical analysis plan, clinical study report, and blank or annotated case report forms, will be provided in a secure data sharing environment. For details on submitting a request, see the instructions provided at www.vivli.org.

Ethics Approval and Consent to Participate

The Libretto-001 (NCT03157128) and Libretto-431 (NCT04194944) trials were done in accordance with Good Clinical Practice guidelines, in line with principles of the Declaration of Helsinki, and all applicable country and local regulations. The protocol was approved by the institutional review board or independent ethics committee at each investigative site. All patients provided written informed consent.

Authors' Contributions

B.K., K.M.C., O.P., T.C., and A.J.: Performed study concept and design. M.D.M., R.G., and N.M.: Provided data acquisition, curation and review, as well as preparing visualizations. K.B., X.M., A.J., and R.K.: Performed annotation, developed the QC pipeline, trained, and tested the model. A.A., B.K., K.M.C., O.P., A.J., N.S., and A.D.S.: Performed critical analysis, writing, review and revision of the paper as well as technical, material, and data interpretation support. A.J., K.B., O.P., and B.K. participated in study design, data collection and analysis, decision to publish, and preparation of the article. All authors read and approved the final paper.

Author Disclosure Statement

A.J. has read the journal's policy, and the authors of this article have the following competing interests: This study was funded by Eli Lilly and Company. A.A., K.M.C., M.D.M., and R.G. are full-time employees and stockholders at Eli Lilly and Company. N.M. and N.S. are full-time employees and stockholders at Loxo@Lilly | Eli Lilly and Company. B.D.K. was a full-time employee and stockholder at Eli Lilly and Company, currently is an employee at the Cleveland Clinic. O.P., currently an employee of Nucleai, Inc, was a full-time employee of Eli Lilly and Company at the time the research was performed, and currently holds Eli Lilly and Company stock. A.D.S. was a full-time employee and stockholder at Eli Lilly and Company and currently is an employee at the Novo Nordisk Inc. K.B., R.K., T.C., T.W.C., and A.J. are employees of and hold equity in BioAI Health. X.M. was a full-time employee of BioAI and is currently at CVS Pharmacy. No other potential competing interests relevant to this article were reported.

Funding Information

This study was funded by Eli Lilly and Company. A.A., K.M.C., M.D.M., and R.G. are full-time employees at Eli Lilly and Company. N.M. and N.S. are full-time employees at Loxo@Lilly | Eli Lilly and Company. B.D.K., O.P., and A.D.S. were full-time employees at the time the research was performed. K.B., X.M., R.K., T.C., T.W.C., and A.J. are full-time employees of BioAI Health.

Supplementary Material

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Figure S6
Supplementary Table S1

References

1. Svoboda E. Artificial intelligence is improving the detection of lung cancer. Nature 2020;587(7834):S20–S22; doi: 10.1038/d41586-020-03157-9
2. Ferlay J, EMLFCMMLPMZASIBF. Global Cancer Observatory: Cancer Today. International Agency for Research on Cancer: Lyon, France; 2020.
3. Society AC. Key Statistics for Lung Cancer. American Cancer Society. Available from: https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html
4. Pao W, Girard N. New driver mutations in non-small-cell lung cancer. Lancet Oncol 2011;12(2):175–180; doi: 10.1016/S1470-2045(10)70087-5
5. Riely GJ, Kris MG, Rosenbaum D, et al. Frequency and Distinctive Spectrum of KRAS Mutations in Never Smokers with Lung Adenocarcinoma. Clin Cancer Res 2008;14(18):5731–5734; doi: 10.1158/1078-0432.CCR-08-0646
6. Desai A, Mohammed T, Rakshit S, et al. The landscape of ALK alterations in non-small cell lung cancer. 2021.
7. Fukuoka M, Wu Y, Thongprasert S, et al. Biomarker analyses from a phase III, randomized, open-label, first-line study of gefitinib (G) versus carboplatin/paclitaxel (C/P) in clinically selected patients (pts) with advanced non-small cell lung cancer (NSCLC) in Asia (IPASS). Jco 2009; 27(15_suppl):8006–8006; doi: 10.1200/jco.2009.27.15_suppl.8006
8. Kohno T, Ichikawa H, Totoki Y, et al. KIF5B-RET fusions in lung adenocarcinoma. Nat Med 2012;18(3):375–377; doi: 10.1038/nm.2644
9. Takeuchi K, Soda M, Togashi Y, et al. RET, ROS1 and ALK fusions in lung cancer. Nat Med 2012;18(3):378–381; doi: 10.1038/nm.2658
10. Lipson D, Capelletti M, Yelensky R, et al. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nat Med 2012;18(3):382–384; doi: 10.1038/nm.2673
11. Kohno T, Tabata J, Nakaoku T. REToma: A cancer subtype with a shared driver oncogene. Carcinogenesis 2020;41(2):123–129; doi: 10.1093/carcin/bgz184
12. Ju YS, Lee W-C, Shin J-Y, et al. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res 2012;22(3):436–445; doi: 10.1101/gr.133645.111
13. Stransky N, Cerami E, Schalm S, et al. The landscape of kinase fusions in cancer. Nat Commun 2014;5(1):4846; doi: 10.1038/ncomms5846
14. Romei C, Elisei R. RET/PTC Translocations and Clinico-Pathological Features in Human Papillary Thyroid Carcinoma. Front Endocrinol (Lausanne) 2012;3:54.
15. Mulligan LM. RET revisited: Expanding the oncogenic portfolio. Nat Rev Cancer 2014;14(3):173–186; doi: 10.1038/nrc3680
16. Subbiah V, Hu MI, Wirth LJ, et al. Pralsetinib for patients with advanced or metastatic RET-altered thyroid cancer (ARROW): A multi-cohort, open-label, registrational, phase 1/2 study. Lancet Diabetes Endocrinol 2021;9(8):491–501; doi: 10.1016/S2213-8587(21)00120-0
17. Kato S, Subbiah V, Marchlik E, et al. RET; Aberrations in Diverse Cancers: Next-Generation Sequencing of 4,871 Patients. Clin Cancer Res 2017; 23(8):1988–1997; doi: 10.1158/1078-0432.CCR-16-1679
18. Offin M, Guo R, Wu SL, et al. Immunophenotype and Response to Immunotherapy of RET-Rearranged Lung Cancers. JCO Precis Oncol 2019;3: 1–8; doi: 10.1200/PO.18.00386
19. Drilon A, Oxnard GR, Tan DSW, et al. Efficacy of Selpercatinib in RET Fusion–Positive Non–Small-Cell Lung Cancer. N Engl J Med 2020;383(9): 813–824; doi: 10.1056/NEJMoa2005653
20. Wirth LJ, Sherman E, Robinson B, et al. Efficacy of Selpercatinib in RET-Altered Thyroid Cancers. N Engl J Med 2020;383(9):825–835; doi: 10.1056/NEJMoa2005651
21. Subbiah V, Wolf J, Konda B, et al. Tumour-agnostic efficacy and safety of selpercatinib in patients with RET fusion-positive solid tumours other than lung or thyroid tumours (LIBRETTO-001): A phase 1/2, open-label, basket trial. Lancet Oncol 2022;23(10):1261–1273; doi: 10.1016/S1470-2045(22)00541-1
22. Drilon A, Subbiah V, Gautschi O, et al. Selpercatinib in Patients With RET Fusion–Positive Non–Small-Cell Lung Cancer: Updated Safety and Efficacy From the Registrational LIBRETTO-001 Phase I/II Trial. J Clin Oncol 2022;41(2):385–394; doi: 10.1200/jco.22.00393
23. Belli C, Penault-Llorca F, Ladanyi M, et al. ESMO recommendations on the standard methods to detect RET fusions and mutations in daily practice and clinical research. Ann Oncol 2021;32(3):337–350; doi: 10.1016/j.annonc.2020.11.021
24. McKenzie AJ, Dilks H, Jones SF, et al. Should next-generation sequencing tests be performed on all cancer patients? Expert Rev Mol Diagn 2019;19(2):89–93; doi: 10.1080/14737159.2019.1564043
25. Qu H, Zhou M, Yan Z, et al. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. NPJ Precis Oncol 2021;5(1):87–11; doi: 10.1038/s41698-021-00225-9
26. Liao H, Long Y, Han R, et al. Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. Clin Transl Med 2020;10(2):e102; doi: 10.1002/ctm2.102
27. Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: A retrospective study. Lancet Digit Health 2021;3(12): e763–e772; doi: 10.1016/S2589-7500(21)00180-1
28. Murchan P, Ó'Brien C, O'Connell S, et al. Deep Learning of Histopathological Features for the Prediction of Tumour Molecular Genetics. Diagnostics (Basel) 2021;11(8):1406; doi: 10.3390/diagnostics11081406
29. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24(10):1559–1567; doi: 10.1038/s41591-018-0177-5
30. Berger MF, Lawrence MS, Demichelis F, et al. The genomic complexity of primary human prostate cancer. Nature 2011;470(7333):214–220; doi: 10.1038/nature09744
31. Campbell JD, Alexandrov A, Kim J, et al. Cancer Genome Atlas Research Network. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet 2016;48(6): 607–616; doi: 10.1038/ng.3564

32. Collisson EA, Campbell JD, Brooks AN, et al. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511(7511):543–550; doi: 10.1038/nature13385
33. Zhou C, Solomon B, Loong HH, et al. LIBRETTO-431 Trial Investigators. First-Line Selpercatinib or Chemotherapy and Pembrolizumab in RET Fusion Positive NSCLC. N Engl J Med 2023;389(20):1839–1850; doi: 10.1056/NEJMoa2309457
34. Zisserman KSaA. Very Deep Convolutional Networks for Large-Scale Image Recognition. Arxiv; 2014, doi:10.48550/arxiv.1409.1556
35. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. IEEE; 2016.
36. Reinhard E, Adhikhmin M, Gooch B, et al. Color transfer between images. IEEE Comput Grap Appl 2001;21(4):34–41; doi: 10.1109/38.946629
37. Wang Y-Y, Chang S-C, Wu L-W, et al. A Color-Based Approach for Automated Segmentation in Tumor Tissue Classification. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:6577–6580.
38. Magee D, Treanor D, Crellin D, et al. Colour Normalisation in Digital Histopathology Images. Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy: OPTIMHisE 2009.
39. Ruderman DL, Cronin TW, Chiao C-C. Statistics of cone responses to natural images: Implications for visual coding. J Opt Soc Am A 1998;15(8): 2036–2045; doi: 10.1364/JOSAA.15.002036
40. Zoph B, Vasudevan V, Shlens J, et al. Learning Transferable Architectures for Scalable Image Recognition. Arxiv 2017; doi: 10.48550/arxiv.1707.07012
41. Beretta C, Ceola S, Pagni F, et al. The role of digital and integrative pathology for the detection of translocations: A narrative review. Precis Cancer Med 2022;5:16–16.