Validation of H&E-based Laboratory Development Test Model Predicting Molecular Alterations in Non-Small Cell Lung Cancer

Kshitij Ingale[1], Qiyuan Hu[1], Sun Hae Hong[1], Rohan P. Joshi[1], Jacob W. H. Gordon[1], Yoni Muller[2], Ben Terdich[1], Jason Blue-Smith[1], Ryan Jones[1], Nike Beaubier[1], Chithra Sangli[1], Riccardo Miotto[1]
[1]Tempus AI, Inc., Chicago, IL, [2]Work done while at Tempus AI, Inc., Chicago, IL

Disclosure Statement: All authors are current or former employees of Tempus AI, Inc. All current employees hold stock in the company.

INTRODUCTION

Molecular alterations detected by next generation sequencing (NGS) are important for identifying patients who could benefit from targeted therapies. However, NGS is not always performed due to cost, tissue availability and long turnaround times. Digital pathology models using routinely generated hematoxylin and eosin (H&E) stained images offer a promising solution to prioritize patients for NGS testing. In this study, we developed and validated an H&E-based model to predict clinically-actionable pathogenic and/or likely pathogenic alterations in *ALK, BRAF, EGFR, ERBB2, MET, RET,* and *ROS1* in non-small cell lung cancer (NSCLC).

METHODS

The model was trained using a cohort of >40,000 images (40,990 patients) and later validated on >500 images (510 patients) along with their corresponding labels generated from Tempus xT, an NGS-based DNA assay. A multi-task model was trained to predict the binary labels using an attention based multiple instance learning model with H-optimus-0 foundation model serving as a featurization model. Simulation studies were performed to determine limits for tissue area and artifact content by progressively incorporating more tissue patches and simulated blurry and color perturbed patches during model inference. Model robustness to images obtained from different scanners and multiple scans obtained from the same scanner was also evaluated. Operating point based metrics were calculated on the validation cohort using thresholds determined on a tuning set with similar data characteristics. Model stability against technical variations, including scanner types, artifacts, and tissue size was evaluated using Root mean squared error (RMSE), Lin's Concordance Correlation Coefficient (CCC) and Pearson's correlation (r). Additionally, area under Receiver Operating Characteristic curve (ROC-AUC) was used to evaluate predictive accuracy

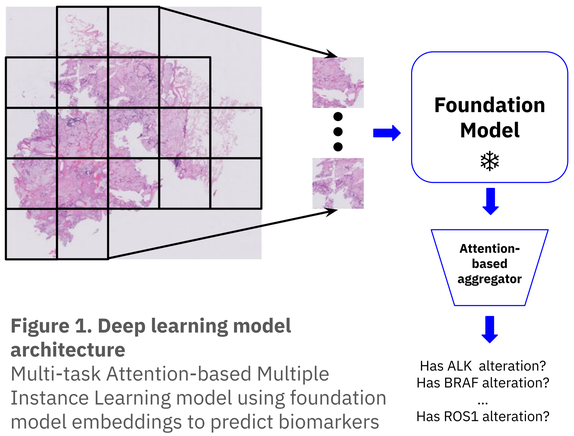

Figure 1. Deep learning model architecture
Multi-task Attention-based Multiple Instance Learning model using foundation model embeddings to predict biomarkers

SUMMARY

- Our H&E image based deep learning model can predict actionable mutations in NSCLC, demonstrating potential for patient stratification
- Patients with high likelihood of biomarker can be prioritized for subsequent confirmatory testing
- Validation studies also indicate that model would be robust to real-world deployment scenarios

RESULTS

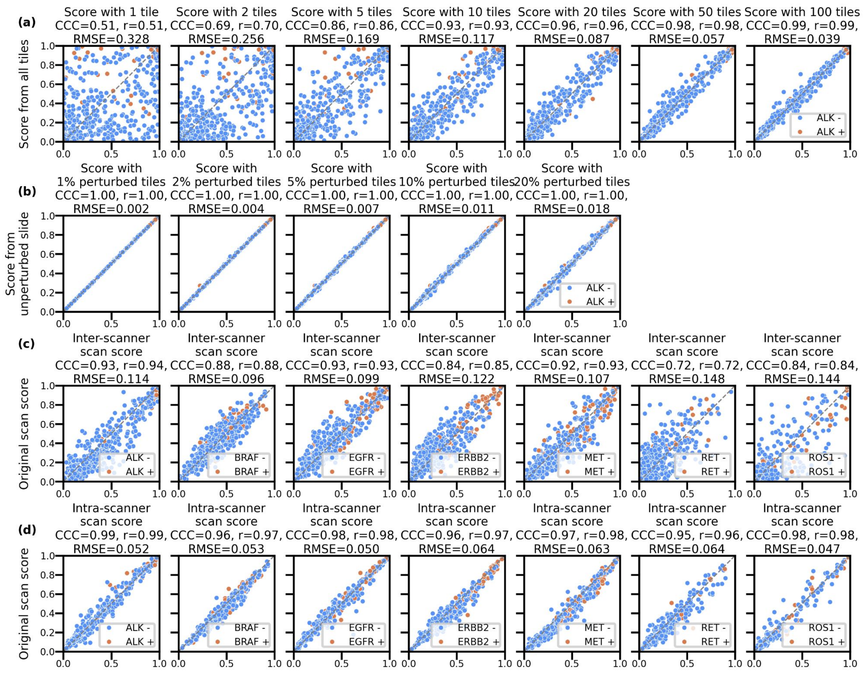

Figure 2. Studies evaluating model robustness as part of validation

(a) **Limit of Detection:** This analytical sensitivity study ran simulation by using 1, 2, 5, 10, 20, 50 and 100 tissue region tiles to calculate model score and compared with entire tissue region score. Acceptance criterion: RMSE < 0.1, allowed LoD: 1.08 mm^2

(b) **Artifact robustness:** This analytical specificity study simulated artifacts on 1%, 2%, 5%, 10% and 20% of tissue area and evaluated model robustness compared to unperturbed image. Acceptance criterion: RMSE < 0.1, allowed maximum artifact: 20% of area

(c) **Inter-scanner robustness:** This analytical precision study evaluated model reproducibility by comparing model scores on scans generated from different scanners. Acceptance criterion: RMSE < 0.1

(d) **Intra-scanner robustness:** This analytical precision study evaluated model repeatability by comparing model scores on scans generated from same scanner at different time points. Acceptance criterion: RMSE < 0.1, model is consistent across different scans

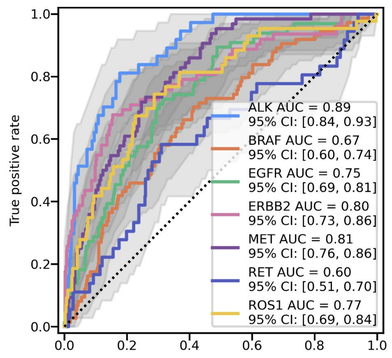

Figure 3. Receiver Operating Characteristic curves for all biomarkers
This analytical accuracy study evaluated model's ability to predict biomarker based on ROC-AUC. Model was significantly predictive of the biomarkers from digitized H&E image with AUCs ranging from 0.6 to 0.89 on validation test set.

Biomarker	Cohort Prevalence (N=510)*	Operating point	Sensitivity	Specificity
ALK	7%	95% Sensitivity	100.0%	42.6%
BRAF	14%	85% Sensitivity	78.4%	43.5%
EGFR	13%	85% Specificity	45.5%	83.6%
ERBB2	12%	90% Sensitivity	93.5%	27.4%
MET	12%	90% Sensitivity	85.9%	60.0%
RET	7%	95% Sensitivity	83.3%	13.1%
ROS1	8%	95% Sensitivity	93.0%	41.5%

Table 1. Validation set metrics
Metrics on validation set determined using threshold selected on tuning set. Model achieves desired sensitivity indicating potential for screening applications that could encourage confirmatory testing.
* Validation set was enriched for positive cases for rare biomarkers

ACKNOWLEDGMENTS
We thank Dana DeSantis from the Tempus Science Communications team for poster development.

Correspondence: kshitij.ingale@tempus.com