Laboratory Development Test Validation of Deep Learning Histogenomic Models to Predict MSI Status in Multiple Cancer Types

Jacob W. H. Gordon[1], Qiyuan Hu[1], Kunal Nagpal[1], Rohan P. Joshi[1], Yoni Muller[1], Alvin Ihsani[1], Nike Beaubier[1]
[1]Tempus AI, Inc., Chicago, IL, USA

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

INTRODUCTION

Microsatellite instability-high (MSI-H) is a tumor-agnostic biomarker for immune checkpoint inhibitor therapy. Previous studies have shown that AI-based imaging predictors can infer MSI status from hematoxylin and eosin (H&E) whole-slide images (WSIs). We have developed AI models that predict MSI status in prostate, colorectal, and endometrial cancer and performed laboratory development test (LDT) validations following the CAP/CLIA standards for these models.

DESIGN

- H&E-stained WSIs of biopsies and surgical resections containing prostate cancer, colorectal cancer, and endometrial cancer were split into:
  - Model development (prostate: n=4252, MSI-H 2.5%, colorectal: n=10445, MSI-H 6.7%, endometrial: n=2354, MSI-H 21%); and
  - Validation sets enriched for MSI-H (prostate: n=198, MSI-H 31%, colorectal: n=234, MSI-H 41%, endometrial: n=150, MSI-H 35%)
- Attention-based multiple instance learning models were trained to predict MSI status for each cancer type
  - Pathologists annotate tumor regions, which are used as input to the models. (See Figure 1)
- In LDT validation, each model was evaluated for its analytical accuracy, analytical precision, analytical sensitivity, and analytical specificity with predefined acceptance criteria
  - The analytical accuracy study evaluated the model performance for predicting MSI status
  - The analytical precision study validated reproducibility and repeatability using inter-scanner and intra-scanner rescans of the same slides
  - For analytical sensitivity and specificity, we established a limit of detection (LoD) on the tumor area and a limit on the amount of blurring and color distortion in the images each model could tolerate

Figure 1. WSI Tumor Region Annotation

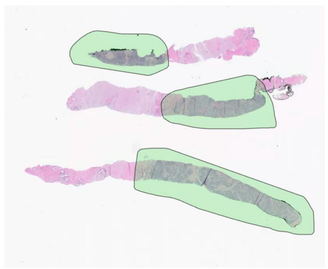

Figure 1. Example of pathologist in the loop tumor region annotations

SUMMARY

- We trained H&E Deep Learning models to predict MSI status in prostate, colorectal, and endometrial cancer
- We validated our AI models using a standardized procedure
- All three of our models passed our validation
- The prostate model has been deployed internally; more algorithm deployments will follow in the future

RESULTS

Figure 2. Analytical Accuracy

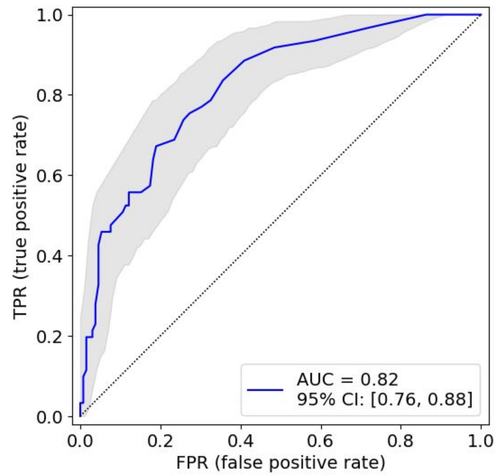

Figure 2. Representative example ROC from our prostate cancer model. All models were found to be significantly predictive, with AUCs ranging from 0.82 to 0.90.

Figure 3. Analytical Precision

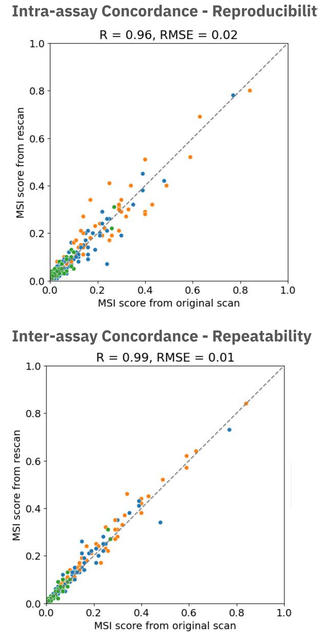

Figure 3. Representative example of intra/inter-assay concordance from our prostate cancer model. Scatter plot of prediction scores from images scanned by different scanner types. Pearson correlation coefficient (R) and root mean square error (RMSE) are labeled above each scatter plot. All models passed our criteria for R to exceed 80% at sensitivities of 70% and 90%.

Figure 4. Analytical Sensitivity

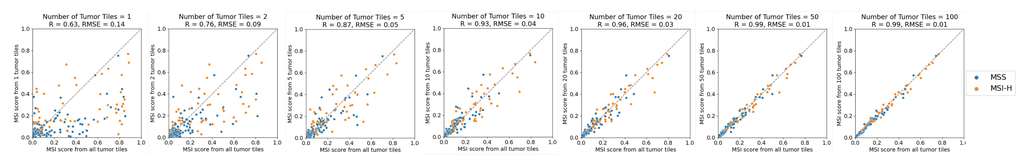

Figure 4. Representative example of setting a LoD from our prostate cancer model. Scatter plots of prediction scores from various levels of tumor areas against prediction scores from the entire tumor region. Pearson correlation coefficient (R) and root mean square error (RMSE) are labeled above each plot. The LoD is set to the smallest number of tiles which can satisfy our acceptance criteria.

Table 1. Full LDT Results for each Model

Test	Acceptance Criteria	Prostate	Colorectal	Endometrial
Analytical Sensitivity	Limit of detection such that RMSE ≤ 0.1 and R ≥ 0.8 between predictions on subsampled and original tumor regions	LoD = 0.069 mm^2	LoD = 0.14 mm^2	LoD = 0.14 mm^2
Analytical Specificity	Maximum artifact % such that RMSE ≤ 0.1 and R ≥ 0.8 between predictions on perturbed and original slides	Maximum artifact % = 20%	Maximum artifact % = 20%	Maximum artifact % = 20%
Analytical Accuracy	Significantly predictive AUC	AUC [95% CI] = 0.82 [0.76, 0.88]	AUC [95% CI] = 0.90 [0.86, 0.94]	AUC [95% CI] = 0.87 [0.82, 0.93]
Analytical Precision - Reproducibility	Inter-scanner prediction concordance ≥ 0.8 at target sensitivities	Concordance at 70% sensitivity = 94%	Concordance at 70% sensitivity = 92%	Concordance at 70% sensitivity = 81%
		Concordance at 90% sensitivity = 87%	Concordance at 90% sensitivity = 88%	Concordance at 90% sensitivity = 91%
Analytical Precision - Repeatability	Intra-scanner prediction concordance ≥ 0.8 at target sensitivities	Concordance at 70% sensitivity = 95%	Concordance at 70% sensitivity = 86%	Concordance at 70% sensitivity = 93%
		Concordance at 90% sensitivity = 96%	Concordance at 90% sensitivity = 98%	Concordance at 90% sensitivity = 97%

Table 1. Full results of the LDT validation for all four studies on the three models, along with descriptions of the acceptance criteria associated with each study

- For analytical sensitivity the Limit of Detection (LoD), was determined by testing drift in model prediction, measured by Pearson correlation coefficient (R) and root mean square error (RMSE), while using only 1, 2, 5, 10, 20, 50, 100, or all available tiles within the slide containing tumor
  - The smallest number of tiles which can satisfy the acceptance criteria was then converted into an area for the LoD
- For analytical specificity, the maximum allowable artifact percentage was determined by simulating model predictions with 5%, 10% 15%, and 20% of tiles containing color or blur augmentation
  - The largest percentage which satisfies the acceptance criteria is the maximum artifact percent reported in the table
- For analytical accuracy, the area under the receiver operating characteristic curve (AUC) and its 95% confidence interval (CI) are reported
- For analytical precision, the inter/intra-scanner concordance is reported at target sensitivities of 70% and 90%

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

Correspondence: jacob.gordon@tempus.com