Artificial intelligence (AI) foundation model as a predictor of efficacy of next-generation checkpoint inhibition with botensilimab (BOT) + balstilimab (BAL) in solid tumors using pretreatment H&E images

Ryan Dalton[1,*], Eshed Margalit[1,*], Keith Mitchell[1,*], Michela Meister[1,*], Daniel Millman[1], Maede Zolanvari[1], Lucas Cavalcante[1], Joy S. Tea[1], Dexter Antonio[1], Maxime Dhainaut[1], Chloe Delepine[2], Dulce Ovando[1], Francis Fernandez[1], Aaron Salm[1], Angela V. Hafner[2], Joseph E. Grossman[2], Dhan Chand[2], Ronald W. Alfa[1], Daniel Bear[1], Emily Corse[1], Lacey Padrón[1]

1. Noetik-AI, South San Francisco, CA. 2. Agenus, Lexington, MA. * These authors contributed equally to this work.

Introduction

- Next-generation immune checkpoint inhibition with botensilimab (BOT, an Fc-enhanced anti-CTLA-4 antibody that enhances T-cell priming, depletes intratumoral regulatory T cells, and activates antigen-presenting cells)[1] and balstilimab (BAL, an anti-PD-1 antibody pharmacologically comparable to approved PD-1 inhibitors) has demonstrated meaningful clinical activity across immunologically "cold" and ICI-refractory tumors, including microsatellite-stable metastatic colorectal cancer (MSS mCRC), sarcoma, and ovarian cancer.[2–4]
- Response to BOT+BAL remains difficult to predict with traditional biomarkers like PD-L1 or TMB, or with histology, particularly in immunologically cold tumors.
- The underlying biology of response likely exists in the spatial and molecular makeup of the TME, but interrogating the TME at high resolution is not feasible in clinical practice.
- Foundation AI models have the potential to discover the biology of response by learning the relationships between H&E and higher-resolution or higher-dimensional data.
- TARIO-2 models predict spatial transcriptomic data and perform simulations of the TME using H&E images alone. These models may therefore be able to use pre-treatment H&E to identify patients likely to respond to treatment.
- In this work we use TARIO-2 models to predict response to BOT+BAL using only H&E images.

Methods

BOT + BAL Clinical Trial Design and Sample Selection

C-800-01 Phase 1b Trial (NCT03860272)

Key Eligibility:

- Advanced solid tumors refractory to standard treatments
- Prior ICI therapy allowed

Botensilimab *Q6W* — *1 mg/kg, 2 mg/kg, or 150 mg*

\+

Balstilimab *Q2W* — *3 mg/kg or 450 mg*

Treatment

***N*=113**

Efficacy-evaluable[a] BOT+BAL–treated patients with H&E images.

Evaluated tumor cohorts selected based on:

MSS mCRC (no active liver metastases [NLM])	Ovarian Cancer	Sarcomas
N=53	*N=27*	*N=33*
Sufficient tumor content for model prediction[b]	Sufficient tumor content for model prediction[b]	Sufficient tumor content for model prediction[b]
BOR: N=49; OS: N=47	*BOR: N=23; OS: N=24*	*BOR: N=31; OS: N=31*

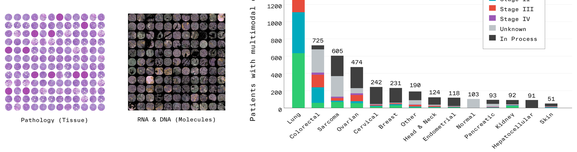
Data cut-off date: December 13, 2025.
[a]Efficacy-evaluable patients received ≥1 post-baseline 6-week imaging scan.
[b]Differences in quality control resulted in 2 additional patients in MSS mCRC NLM BOR studies and 1 additional patient in ovarian cancer OS studies.

References:

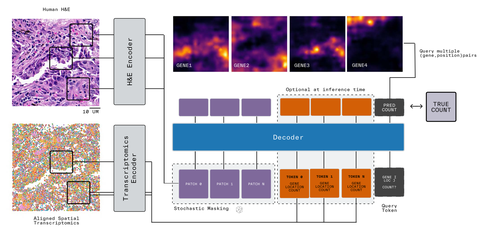
1. Chand D, et al. *Cancer Discov*. 2024;14(12):2407–2429.
2. Bullock AJ, et al. *Nat Med*. 2024;30(9):2558–2567.
3. Wilky BA, et al. *J Clin Oncol*. 2025;43(11):1358–1368.
4. Porter R, et al. *J Immunother Cancer*. 2025;13(12):e013222.

Building a Multi-Modal, Pan-Cancer Dataset to Train Foundation Models

Noetik's model training dataset consists of in-house generated Tissue Microarrays (TMAs, patent pending) profiled with 3 spatial technologies: H&E (left), Spatial Transcriptomics (center), and multiplex immunofluorescence (not shown and not used for this study). All data modalities are aligned with sub-cellular resolution. Right: Noetik's underlying cancer dataset, summarized by cancer type and stage. This dataset currently includes >5000 patients, each with all 3 spatial data types across 10+ TMA cores, including >725 CRC, >605 Sarcoma, & >474 Ovarian patients. Foundation models were not exposed to the prediction cohort in this study.

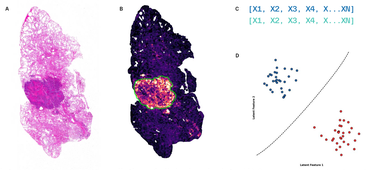


Training AI Models to Infer CosMx From H&E



Foundation AI models trained on multimodal data infer spatial transcriptomics from pretreatment H&E. TARIO-2 is a foundation model that is trained on aligned tiles of human H&E images and spatial RNA transcripts. Each modality is encoded into a sequence of tokens with a modality-specific encoder; then the sequences from both modalities are concatenated. During training, a transformer-based decoder is used to predict later tokens in the sequence from earlier tokens. Training does not include clinical samples. At inference time, H&E from pretreatment samples is fed to the model, which predicts dense spatial patterns of gene expression for each gene in the transcriptomics panel.

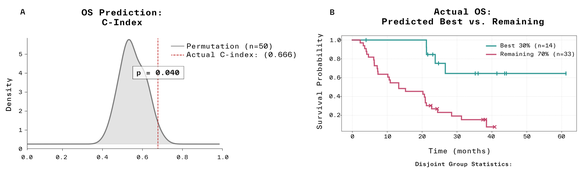
Predicting Clinical Endpoints from H&E Alone



Synthetic spatial transcriptomics data enables prediction of clinical endpoints. (A) An example pre-treatment H&E image. Only patients from the efficacy-evaluable group (defined by having at least one on-treatment scan) in the background study, and for whom a baseline FFPE tissue sample was available at the time of the analysis, were included. H&E images were subject to quality control. The analysis population herein includes one image per patient, totaling 53 MSS mCRC NLM, 27 ovarian cancer, and 33 sarcoma patients. (B) Inferred spatial transcriptomics on an MSS mCRC NLM image, with inferred *CD27* shown as an example (not indicative of relationship to response). (C) Tumor, nontumor, or other inferred regions are aggregated into feature vectors. Aggregations used expert pathology annotations. (D) Classifier models were trained for binary classification tasks, with responders defined as PR and CR. For all prediction tasks, models were evaluated using 50 repeats of MCCV (BOR) or LOGO (OS), with accuracy calculated on out-of-fold data at the patient level. Any hyperparameter tuning was done within training folds. For BOR, data are summarized as median AUROC and for OS prediction data are summarized as median C-index. For both AUROC and C-index, a value of 0.5 denotes a chance classifier and a value of 1.0 denotes a perfect classifier. All endpoints were predicted independently, with no data leakage between endpoints.

Results

Predicting OS in the MSS mCRC NLM Cohort

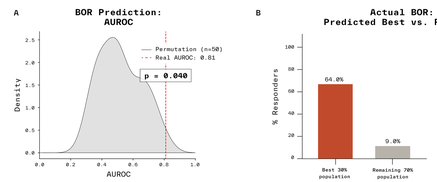


OS prediction: MSS mCRC NLM. (A) For OS prediction, 47 WSIs were used for feature generation, followed by Gradient Boosting model training and Cox proportional hazard loss for censored OS data. Data shown are the median C-index (0.666) from LOGO cross-validation. Significance was tested by shuffling labels to create permutation distributions (gray). The C-index had a p-value of 0.04 relative to the permutation distribution. (B) For all patients used in OS prediction tasks, a median risk score was calculated across out-of-fold predictions. Based on these scores, patients were partitioned into two groups: best 30% (blue) and remaining 70% (red). These groups were pre-specified and were not model-determined. Actual OS was then compared for these partitions. The best 30% population has a hazard ratio of 0.28 as compared to the remaining patients, and median OS was not reached. The remaining population has a median OS of 13.3 months.

MSS mCRC NLM	Best 30% by Predicted OS Risk *(N = 14)*	Remaining 70% by Predicted OS Risk *(N = 33)*
OS hazard ratio	0.18***	5.63
Median OS (months)	NR***	13.3
PFS hazard ratio	0.71	1.18
Median PFS (months)	5.5	4.1
Median Age, years (range)	60 (38-71)	57 (36-82)
Female, *N* (%)	*7* (50%)	*15* (45%)
ECOG of 1 at baseline, *N* (%)	*7* (50%)	*22* (67%)
Median prior lines of therapy (range)	2 (1-7)*	4 (1-10)
Prior PD-(L)1, *N* (%)	*5* (36%); *N=14*	8 (24%); *N=33*
PD-L1 TPS – median (range); mean; *N*	0 (0-20); mean=2.0; *N=13*	0 (0-10); mean=0.5; *N=32*
PD-L1 CPS – median (range); mean; *N*	5 (0-90); mean=13.1; *N=13*	3 (0-20); mean=5.7; *N=32*

Demographics & outcomes among model-selected cohort. Patients were stratified into groups based on their median predicted OS risk. Demographic and disease characteristics were then compiled for these groups. Differences in ECOG and prior lines of therapy between groups were not significant, but may point towards prognostic OS prediction. Details on statistical methods in following section.

Predicting Best Overall Response for the MSS mCRC NLM Cohort

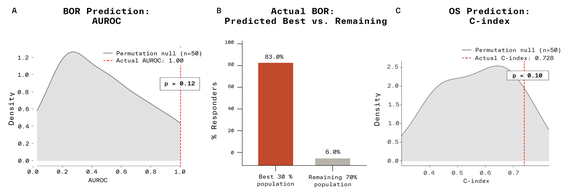


Improved outcomes among model-selected cohort. (A) For BOR prediction, 49 H&E WSIs were used for feature generation, with modeling assessed as described in the background sections. Classifiers reached a median AUROC of 0.81 and p-value = 0.04 (B) Using median predicted responses on out-of-fold data, patients were stratified into best 30% and remaining 70% groups. Within the best 30% group the actual response rate was 64%, compared to a 9% actual response rate among the remaining 70% group.

MSS CRC, no liver mets	Best 30% by Predicted Response *(N = 14)*	Remaining 70% by Predicted Response *(N = 35)*
Responders, *N* (%)	9 (64%)***	3 (9%)
PFS hazard ratio	0.46*	2.16
Median PFS (months)	9.8*	4.1
Median Age, years (range)	60 (48-82)	57 (36-81)
Female, *N* (%)	*5* (36%)	*18* (51%)
ECOG of 1 at baseline, *N* (%)	*7* (50%)	*24* (69%)
Median prior lines of therapy (range)	4 (1-10)	4 (1-9)
Prior PD-(L)1, *N* (%)	2 (14%); *N=14*	*11* (31%); N=35
PD-L1 TPS – median (range); mean; *N*	0 (0-20); mean=2.8; *N=13*	0 (0-5); mean=0.2; *N=33*
PD-L1 CPS – median (range); mean; *N*	5 (0-90); mean=12.8; *N=13*	3 (0-20); mean=5.6; *N=33*

Demographics & outcomes among model-selected cohort. Patients were stratified into groups based on their median predicted BOR. Demographic and disease characteristics were then compiled for these groups. Categorical variables were compared with two-sided Fisher's exact tests (with missing values dropped per group), continuous variables with two-sided Mann-Whitney U tests, and time-to-event endpoints (OS, PFS) with two-sided log-rank tests between the best 30% and remaining 70% groups, with displayed hazard ratios derived from Cox models comparing each group against the other cohort (best vs. remaining, remaining vs. best)
(* p<0.05, ** p<0.01, *** p<0.001).

Predicting BOR and OS for the Ovarian Cancer Cohort

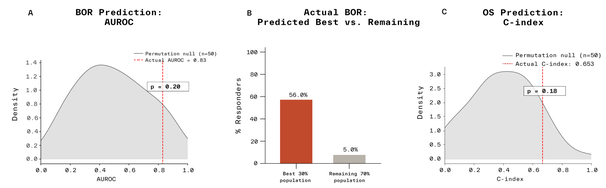


Best overall response & OS prediction: ovarian cancer: (A) For BOR prediction, 23 H&E WSIs were used for feature generation and model prediction. Median AUROC = 1 and p-value = 0.12. (B) Patients were stratified into groups based on their median predicted response using held-out data, yielding a best 30% population with an actual response rate of 83%, compared with a 6% response rate for the remaining 70% population. (C) For OS prediction, 24 H&E WSIs were used. C-index = 0.728 and p-value = 0.1.

Ovarian	Best 30% by Predicted Response *(N = 6)*	Remaining 70% by Predicted Response *(N = 17)*
Responders, *N* (%)	*5* (83%)**	*1* (6%)
PFS hazard ratio	0.15**	6.79
Median PFS (months)	11.1**	1.4
Median Age, years (range)	66 (55-70)	59 (37-75)
ECOG of 1 at baseline, *N* (%)	*3* (50%)	*9* (53%)
Median prior lines of therapy (range)	2 (1-8)	4 (2-16)
Prior PD-(L)1, *N* (%)	1 (17%); *N=6*	1 (6%); *N=17*
PD-L1 TPS – median (range); mean; *N*	*1* (0-15); mean=3.0; *N=6*	*1* (0-30); mean=3.2; *N=17*
PD-L1 CPS – median (range); mean; *N*	7 (1-30); mean=10.7; *N=6*	*5* (0-65); mean=10.8; *N=17*
Histology		
HGS & Serous Adeno, *N* (%)	*5* (83%)	*14* (76%)
Clear Cell, *N* (%)	*1* (17%)	*3* (18%)

Demographics & outcomes among model-selected cohort. Patients were stratified into groups and statistical tests were performed as described for MSS CRC.

Predicting BOR and OS for the Sarcoma Cohort



Best overall response & OS prediction: sarcoma: (A) 31 H&E WSIs were used for feature generation and BOR prediction, reaching a median AUROC of 0.83 and p-value of 0.2. (B) Patients were stratified into groups based on their median predicted responses on out-of-fold data, yielding a best 30% population with an actual response rate of 56%, compared with a 5% response rate for the remaining 70% population. (C) For OS prediction, 31 H&E WSIs were used. C-index = 0.653 and p-value = 0.18.

Sarcoma	Best 30% by Predicted Response *(N = 9)*	Remaining 70% by Predicted Response *(N = 22)*
Responders, *N* (%)	*5* (56%)**	*1* (5%)
PFS hazard ratio	0.68	1.46
Median PFS (months)	5.6	1.4
Median Age, years (range)	64 (40-78)	62 (31-80)
Female, *N* (%)	*4* (44%)	*18* (82%)
ECOG performance status = 1 at baseline, *N* (%)	*5* (56%)	*12* (55%)
Median prior lines of therapy (range)	2 (1-6)	3 (1-10)
Prior PD-(L)1, *N* (%)	*1* (11%)	*1* (5%)
PD-L1 TPS – median (range); mean; *N*	2 (0-95); mean=18.6; *N=9*	0 (0-10); mean=2.2; *N=20*
PD-L1 CPS – median (range); mean; *N*	5 (0-98); mean=19.2; *N=9*	0 (0-20); mean=3.9; *N=20*
Histology		
Angiosarcoma, *N* (%)	*5* (55%)	*5* (23%)
LMS, *N* (%)	*2* (22%)	*10* (45%)
Other, *N* (%)	*2* (22%)	*7* (32%)

Demographics & outcomes among model-selected cohort. Patients were stratified into groups and statistical tests were performed as described for MSS CRC.

Benchmarking TARIO-2 Against the State of the Art

	BOR (AUROC)	OS (C-index)	PFS (C-index)
Noetik	0.83*	0.66*	0.54
Gigapath	0.66*	0.57	0.52
H-optimus-0	0.53	0.51	0.51

Prediction accuracies for Noetik models versus Gigapath, and H-Optimus-0. To benchmark TARIO-2 against external models, 49 H&E WSIs from the MSS mCRC NLM cohort were used. Each accuracy number shown is the best of all aggregation, feature selection, and modeling methods tried with that model. For each model, numerous approaches were tried, including literature recommendations for model utilization. Results that are statistically significant (p<0.05) by permutation testing are indicated with an asterisk.

Summary

- TARIO-2 applied AI-based spatial TME inference to routine pretreatment H&E images to predict clinical benefit from BOT + BAL.
- Using rigorous modeling approaches and permutation-based controls, TARIO-2 demonstrated statistically significant predictive performance for BOR and OS in MSS mCRC NLM, with supportive trends observed in sarcoma and ovarian cancer cohorts.
  - These findings demonstrate that inferred spatial transcriptomic data can predict clinical endpoints and support the potential utility of histology-based patient stratification approaches.
- TARIO-2 outperformed benchmark pathology foundation models for BOR and OS prediction in the MSS mCRC NLM cohort.
- These retrospective analyses support prospective validation of TARIO-2 as an H&E-based biomarker strategy for BOT + BAL, including evaluation in earlier-line MSS CRC settings, including 1L disease.
- Scaling TARIO-2 across H&E and clinical datasets may enable broader patient-to-therapy matching approaches in oncology.

Abstract 2535; Poster 325. Presented at the Annual American Society of Clinical Oncology Congress: May 29-June 2, 2026. Chicago, IL, US.