P2-06-21: Aligning AI Recurrence Predictions with Real-world Recurrence Rates Through Survival Calibration

Ken Zeng[1], Jan Witowski[1], Joseph Cappadona[1], Bartosz Machura[1], Carlos Fernandez-Granda[2,3], Krzysztof J. Geras[1,4]
[1]Ataraxis AI, [2]NYU Center for Data Science, [3]NYU Courant Institute of Mathematical Sciences, [4]NYU Langone Health

SAN ANTONIO BREAST CANCER SYMPOSIUM

Introduction

Prognostic scores, including scores generated by genomic assays, are typically evaluated with respect to their discrimination, i.e., in whether they can distinguish between high- and low-risk patients. However, these scores are uncalibrated, meaning they do not directly correspond to the probability of recurrence. Therefore, they lack a straightforward interpretation. As a result, while such tools may be prognostic, translating their outputs into clinical decisions can be challenging. Calibration offers a solution by aligning predicted risks with observed event rates to produce absolute recurrence risk estimates. However, calibration within survival analysis remains challenging due to censoring. In this study, we demonstrate the critical role of calibration in prognostic model development which ensures that scores are aligned with observed event rates.

Methods

- We calibrated a multi-modal artificial intelligence-based model developed to predict risk of breast cancer recurrence.
- The AI model was previously developed using data from 5,351 early stage breast cancer (BC) patients from 13 international cohorts[1].
- We evaluated the AI model calibration on a hold-out set of 3,457 patients from 5 independent cohorts (separate from the 13 training cohorts).
- To calibrate the AI model, we used spline-based methods.
- During evaluation, we stratified evaluated patients into quartiles based on predicted risk.
- Within each quartile, we compared the median AI score with the observed recurrence rate estimated via Kaplan-Meier estimator.

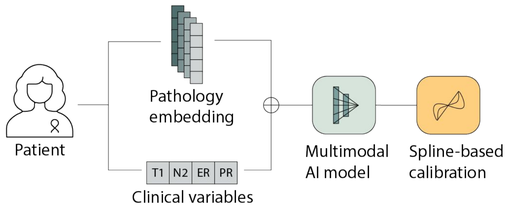


Figure 1: Our AI model integrates patient clinical features with pathology-derived embeddings to predict patient outcomes. We then apply calibration to ensure the predicted probabilities align with outcomes.

Results

Model predictions align with real world recurrence rates

- 3,457 patients were stratified into 4 percentile bins according to their predicted risk scores.
- For each bin, the predicted risk was compared with the corresponding observed recurrence rate estimated using the Kaplan-Meier (KM) estimator.
- Lin's concordance correlation coefficient (CCC) was used to assess how closely the predicted values aligned with the observed outcomes.

Percentile	Observed recurrence rate	Uncalibrated	Calibrated
0-25th	2.5% [1.5%-4.1%]	17.2% [15.1%-18.4%]	3.7% [2.1%-4.8%]
25-50th	4.9% [3.5%-6.8%]	19.9% [18.4%-22.0%]	6.3% [4.8%-8.6%]
50-75th	11.5% [9.3%-14.3%]	24.7% [22.0%-29.0%]	11.9% [8.6%-17.4%]
75-100th	19.5% [16.2%-23.4%]	34.5% [29.0%-77.9%]	25.5% [17.4%-81.2%]

Table 1: AI model scores before and after calibration compared to observed 5-year recurrence rates. For each risk group, we present the KM estimates (observed recurrence rate) with 95% CIs and summary statistics (median, min and max scores) of the uncalibrated and calibrated risk scores.

Calibration of ODX RS based on SWOG 8814 results

- We evaluated Oncotype DX recurrence score (ODX RS) calibration by comparing ODX score against recurrence estimates, as reported in the SWOG 8814 ODX translational study[2].
- First, patients with ODX RS in our cohort were divided into quartiles based on their score.
- The median ODX RS from each quartile was then mapped to the corresponding 5-year disease-free survival (DFS) estimate reported in the SWOG 8814 translational study for ODX RS.
- The analysis was performed separately for different treatment regimens (CAF+Tamoxifen vs. Tamoxifen alone) and nodal status (1–3 vs. ≥4 positive nodes).

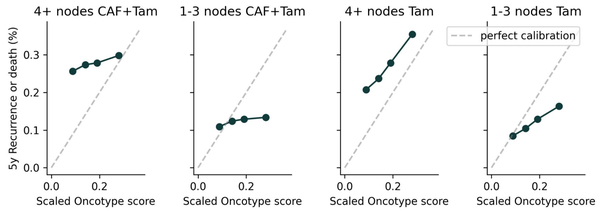


Figure 2: Scaled ODX RS compared with SWOG DFS outcomes. ODX RS calibration is mixed according to results for DFS presented in the SWOG 8814 study, with CCC ranging from 0.13 (for patients with ≥4 nodes, CMF+Tam) to 0.47 (1-3 nodes, Tam only). ODX RS values were scaled to a 0-1 range for consistency.

Comparison between calibrated and uncalibrated scores

- The uncalibrated AI model is poorly calibrated, with a CCC of 0.29 (95% CI: [-0.14, 0.63], p=0.19).
- Spline methods substantially enhanced calibration to a CCC of 0.92 (95% CI: [0.53, 0.99], p=0.001).
- Among patients with ODX RS, the AI model remained calibrated, with a CCC of 0.90 (95% CI: [0.19, 0.99], p=0.02).

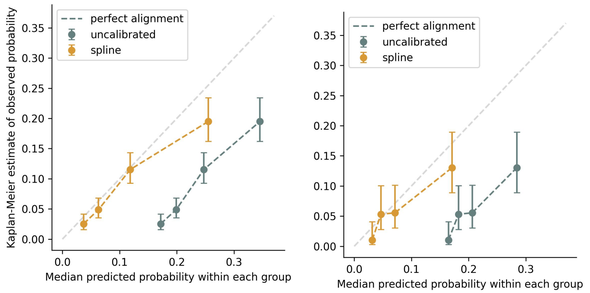
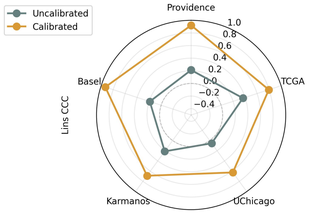


Figure 3: Predicted and observed recurrence probability before and after calibration in all patients (left) and patients with ODX RS (right). Observed outcomes are real-world recurrence outcomes at 5 years. For each risk group, the mean and 95% CI of the Kaplan-Meier estimated 5-year recurrence rates are reported.

Performance in individual evaluation cohorts



	n
Providence	1,687
TCGA	930
UChicago	410
Karmanos	163
Basel	267

Figure 4: Lin's CCCs before and after spline calibration across individual datasets. Calibration substantially improved concordance in all datasets. Table on the right displays the number of patients in each individual cohort.

Conclusion

- The results based on the SWOG study indicate that the ODX RS demonstrates poor calibration.
- Calibration enables our model to produce risk estimates that more accurately align with true recurrence probabilities.
- A well calibrated scores is more interpretable, especially when weighing the cost and benefits of different lines of treatment.
- In future work, we aim to further enhance calibration methodologies and extend their application to models that predict additional clinical outcomes, such as chemotherapy benefit and other treatment effects.

References

1. Jan Witowski, Ken Zeng, Joseph Cappadona, Jailan Elayoubi, Elena Diana Chiru, Nancy Chan, Young-Joon Kang et al. "Multi-modal AI for comprehensive breast cancer prognostication." *arXiv preprint arXiv:2410.21256* (2024).
2. Albain, K. S., Barlow, W. E., Shak, S., Hortobagyi, G. N., Livingston, R. B., Yeh, I. T., ... & Hayes, D. F. (2009). Prognostic and predictive value of the 21-gene recurrence score assay in a randomized trial of chemotherapy for postmenopausal, node-positive, estrogen receptor-positive breast cancer. *Lancet Oncology*, *11*(1), 55.

More information

This presentation is the intellectual property of the author/presenter. Contact them at krzysztof.geras@ataraxis.ai for permission to reprint and/or distribute.