Multi-modal Large Language Models for Metastatic Breast Cancer Prognosis

Raphael Pelossof[1], Mark Carty[1], Talal Ahmed[1], Stanislas Lauly[1], Alberto Purpura[1], Erik Mueller[1], Justin Guinney[1]

[1]Tempus AI, Inc., Chicago, IL

Abstract Presentation #5006

INTRODUCTION

Inputs into cancer prognostic models are primarily structured data such as demographic and clinico-pathological features, and lack richer context often found in unstructured clinical notes. We hypothesized that creating a temporal clinical patient note from structured data that preserves longitudinal and clinical contextual information, and coupling it with a large language model (LLM) trained to prognosticate overall survival (OS) from time of metastatic diagnosis, may improve model accuracy with an interpretable embedding space.

METHODS

To facilitate the interaction of an LLM with structured dates, we developed the Patient Chronological Note (PCN). PCNs convert temporal structured data to a textual representation. Structured data include clinical information about demographics, diagnoses, treatments and outcomes with their timing. The goal of the PCN is to contain enough information for a physician to estimate the patient state at any point in time with respect to the NCCN Guidelines.

DistilBERT, a large language model, was pre-trained using PCNs from breast cancer patients (N=580,000), allowing the LLM to learn a representation of the patient journey. The resulting embeddings were fed into a fully-connected 2-layer network that was fine-tuned using Cox survival loss. Fine tuning was performed on PCNs derived from mBC patients (N=28,500), where the model was trained to predict OS from the time of first metastatic diagnosis. A held-out validation dataset of mBC patients (n=28,800) was used to validate survival prediction accuracy.

SUMMARY

- LLM-Cox model learning from patient chronological notes can improve clinical-molecular prognostication over a linear model.
- The internal embedding representation of the LLM was interpretable, and yielded distinct clinical-molecular subtypes that also showed distinct levels of prognostic risk.
- These groups provide an opportunity to use LLM to redefine traditional clinical risk groups.

RESULTS

Overview of the Cohort

Characteristic	Train [PT] (n=608,578)	Test [PT] (n=609,277)	Train [FT] (n=28,536)	Test [FT] (n=28,888)
Age at diag. (yrs)				
Median (Q1, Q3)	61 (51, 70)	61 (51, 70)	59 (49, 68)	59 (50, 68)
Gender, n				
Female	590,890	591,306	27,624	27,980
Male	5,441	5,471	430	403
Unknown	12,427	12,500	482	505
Race, n				
White	393,065	392,605	17,389	17,459
Black or African American	59,526	60,131	3,212	3,371
Asian	19,481	19,129	794	820
Other Race	34,356	34,078	1,716	1,695
Unknown	102,330	103,334	5,425	5,543
Ethnicity, n				
Hispanic or Latino	28,354	27,629	1,444	1,447
Not Hispanic or Latino	354,874	355,273	14,626	14,723
Unknown	225,530	226,375	12,466	12,718
Stage, n				
0	20,832	21,007	189	184
I	124,874	125,024	2,195	2,168
II	66,004	66,074	3,505	3,600
III	26,400	26,106	2,867	2,882
IV	27,849	28,040	19,403	19,675
Unknown	342,799	343,026	377	379
Grade, n				
I	36,877	36,818	1,658	1,598
II	69,191	68,942	6,672	6,744
III	50,503	50,473	7,367	7,511
Unknown	452,187	453,044	12,839	13,035

Patient Chronological Note

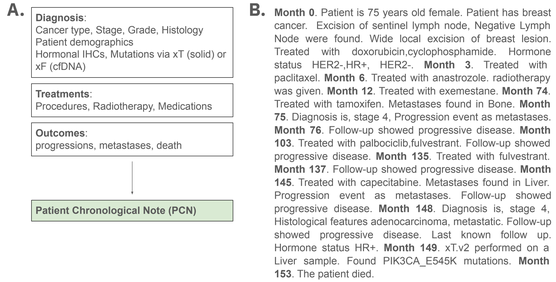

Figure 1. A. Structured data elements that are converted to a Patient Chronological Note (PCN). **B.** PCNs combine both clinical and molecular information. Molecular information includes IHC results and reported DNA variants found by solid tumor (Tempus xT) or blood (Tempus xF) next-generation sequencing.

Model and Performance

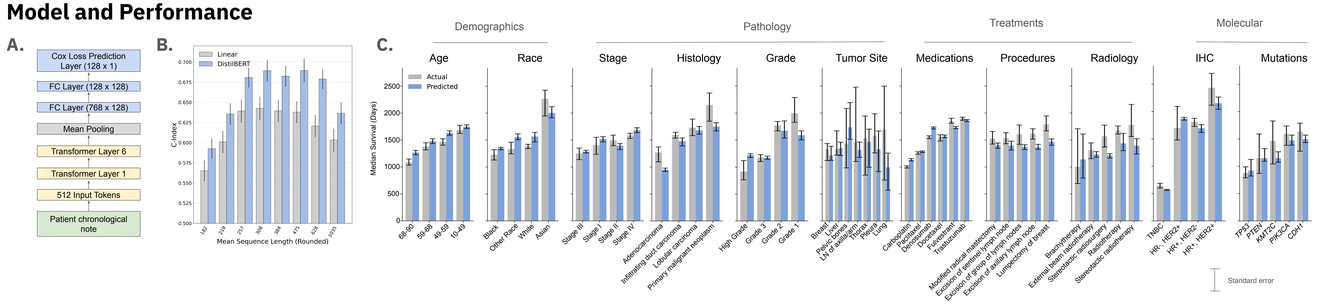

Figure 2. A. The LLM model on test cohort. **B.** Prediction performance for DistilBERT-cox 0.66 (concordance index), outperforming a linear cox model that achieved 0.62. Variable performance is observed as a function of the PCN length. **C.** Comparing actual RWD risk to risk predicted by the model. Actual RWD risk is the KM estimate for median overall survival for each one of the sub-cohorts listed.

The Model Embedding Space is Prognostic and Interpetable

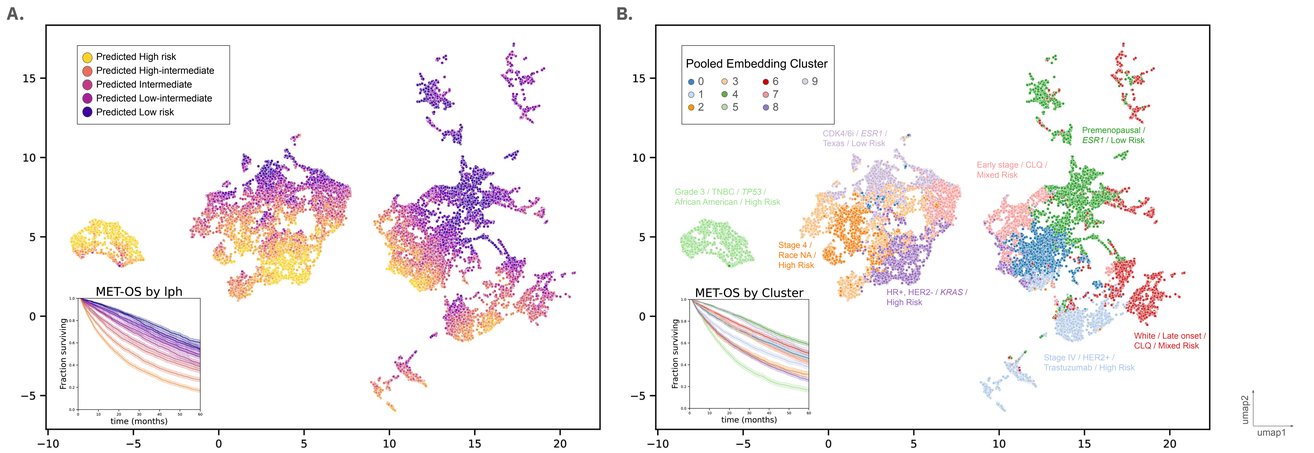

Figure 3. A. The mean-pooling embedding mapping with UMAP onto 2-dimensions. The embedded layer is prognostic, and shows smooth risk gradients. Insert shows survival KM-curves for 5 risk quintiles of test data split according to risk score. The embedding identifies three main sub-cohorts. The left-most is predicted to be high risk, and is mainly TNBC. The other two are further partitioned in figure 3B. **B.** Clustering the LLM embeddings revealed 10 distinct patient groups enriched for key mBC traits, including a high-risk cluster enriched for triple-negative status, *TP53* mutations, and african american race, a low-risk cluster enriched for *ESR1* mutations and CDK46 treatment, and a cluster enriched for low-risk early-onset patients. The different clusters showed different prognostic risk levels.

ACKNOWLEDGMENTS

We thank Matthew Kase for poster development.

Correspondence: Rafi.Pelossof@tempus.com

Guidelines and Deadlines

Important Deadlines

➢ **March 12, 2024**: Presentations due for SciComm review
➢ **March 26, 2025:** Presentation sent to Legal/Leadership for review
  ○ **Poster Printing by Genigraphics (pick up at conference):**
    ● **Posters should be printed at 100% (60 in x 40 in)**
    ● **Plan Ahead** order deadline is **1:00 PM CT on Friday, April 4, 2025;**
    ● **RUSH order** deadline is **1:00 PM CT on Wednesday, April 16, 2025** (rush fee applies)
    ● Upload for printing here: https://www.genigraphics.com/aacr
    ● Tempus presenters may submit for reimbursement through SAP Concur
➢ **April 7, 2025**: **e-poster upload deadline**
➢ **April 25-30, 2025**: Annual Meeting, Chicago, IL

For E-posters ONLY

1. **Make a copy of your google slide (File > make a copy)**
2. **Delete any extra slides (including this one)**
3. **Resize by going to File > Page setup > Resize to 30 x 20**
4. **Download as a pdf: File > download > pdf**

Data Visualization Guidelines

Tempus Color palettes

Qualitative

SciComm preferred palette
This is a minor update to the default palette recommended by graphic design (see below) chosen to minimize the grouping of similar colors.

['#5993F7', '#D97C4F', '#62B882', '#CC78A7', '#774D9A', '#515CBE', '#E9C74E', '#B8E382', '#A54A72', '#C8B1F6']

Graphic design recommendation
For cases where data are paired or grouped in a logical way, we recommend using this ordering (or any re-ordering) that results in the clearest presentation of the data

['#5993F7', '#515CBE', '#D97C4F', '#E9C74E', '#774D9A', '#C8B1F6', '#A54A72', '#CC78A7', '#62B882', '#B8E382']

Graphic design variant
In the event that a slightly lighter look is preferred, this palette (or a logical re-ordering of colors to fit the application) is acceptable

['#86B2FF', '#738AFF', '#F99B6D', '#FCE285', '#AD6CE4', '#CCB2FF', '#E777A8', '#FFC0E3', '#89D3A5', '#D1ECAF']

Continuous

(Note: while these palettes are meant to be used in continuous applications, they are ultimately constructed from discrete color palettes with code examples showing how to properly extrapolate and create a continuous palette for applications such as heatmaps. However, these palettes may also be used in their discrete form [depending on the application], much the same as the qualitative palettes listed above.)

Sequential

['#29293C', '#384162', '#485889', '#5770AF', '#6687D6', '#779BEB', '#8BACED', '#9EBDF0', '#B3CCF3', '#C7DDF6']

Diverging

['#384162', '#475889', '#5770AF', '#6687D6', '#8BACED', '#FFFFFF', '#DDC2CD', '#C99EAD', '#B47A8F', '#9F5773', '#893157']

1. SciComms Data Visualization Best Practices
2. Figure Sizing and Exporting