Evaluating Pre-Trained Foundation Models for Glioblastoma Overall Survival Prediction

Meixu Chen, PhD[1], Kai Wang, PhD[2], and Jing Wang, PhD[1]
[1]MAIA Lab, Department of Radiation Oncology, University of Texas Southwestern Medical Center
[2]Department of Radiation Oncology, University of Colorado School of Medicine

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Accurate glioblastoma (GBM) survival prediction from pre-treatment MRI is challenging due to
- limited labeled datasets
- Substantial inter-institu. variability

We investigated whether frozen, pre-trained medical imaging foundation models (FMs) provide transferable representations for overall survival prediction that generalize across institutions without retraining.

Models were developed on **UPENN-GBM** (n=611) and evaluated on three independent external cohorts:
- UCSF-PDGM (n=400)
- CFB-GBM (n=201)
- BraTS-TCGA-GBM (n=97).

Four pre-trained foundation models:
- BrainSegFounder
- VoCo
- BrainIAC
- SAM-Med3D

For feature extraction from T1, T1ce, T2, and T2-FLAIR MRI, which were combined with clinical variables in a DeepSurv model.

The foundation model ensemble achieved a C-index of 0.684 on the UPENN-GBM developing cohort and demonstrated consistent external performance with C-indices of 0.656 (UCSF-PDGM), 0.641 (BraTS-TCGA-GBM), and 0.608 (CFB-GBM). While imaging features did not significantly outperform a clinical-only model, they matched handcrafted radiomics on most cohorts without requiring tumor segmentation. Combining foundation model, radiomics, and clinical features further improved performance across all datasets.

CONTACT

Jing Wang, PhD, DABR
UT Southwestern Medical Center
2280 Inwood Rd, Dallas, TX 75235, USA
Jing.Wang@UTSouthwestern.edu
214-648-1795

UT Southwestern Medical Center

INTRODUCTION

Glioblastoma (**GBM**)
- The **most common, aggressive** primary malignant brain tumor
- Median overall survival of only 14–16 months despite standard treatment
- Accurate pre-treatment survival prediction is important for risk stratification, treatment planning, and clinical trial enrollment
- Survival prediction remains challenging because of the biological heterogeneity and the limited availability of large datasets
- **MRIs** provides complementary information about tumor morphology and biology and has been investigated for survival prediction
- Deep learning models **overfit** small datasets and generalize poorly across institutions because of variations in imaging protocols, scanners, and patient populations
- Recent **imaging foundation models** have demonstrated strong transferability across medical imaging tasks, but survival prediction is not usually on the scope

This work focuses on evaluating the ability of **pre-trained 3D imaging foundation models** to provide robust and transferable representations for GBM **overall survival prediction** across multiple institutions.

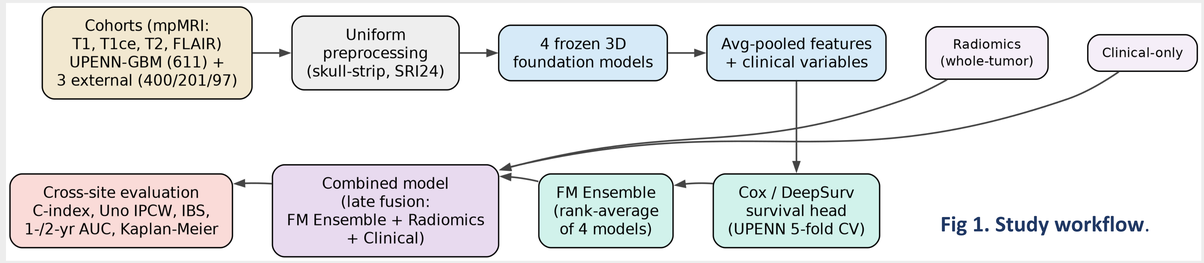


Fig 1. Study workflow.

METHODS AND MATERIALS

In this study we evaluate four state-of-the-art frozen 3D medical-imaging foundation models shown in **Table 1**, for GBM OS prediction from mpMRI.

Models are developed on **UPENN-GBM** (n=611) with five-fold cross-validation and evaluated on **three strictly held-out external cohorts** spanning three countries (**Table 2**).

We benchmark against a Clinical-Only model, a convolutional FM baseline (UniBrain), and a radiomics model, and we test whether late fusion improves prognostication.

Predictions use a Cox proportional-hazards head, and we jointly assess discrimination (C-index, year-1/year-2 ROC-AUC) and calibration (Integrated Brier Score), which we show can diverge **under cross-institutional domain shift**.

Table 1. Evaluated 3D foundational Models.

Model	Architecture	Pre-training data	Pre-training objective	Feat. dim
BrainSegFounder	3D Swin Transformer	~41k brain MRI (healthy + disease)	Two-stage self-supervised	768
VoCo	3D Swin Transformer	~160k CT/MRI (mostly CT)	Volume contrastive	768
BrainIAC	3D ViT (MAE)	~48k multi-sequence brain MRI	Masked autoencoding	768
SAM-Med3D	3D ViT (SAM)	~22k volumes / 143k masks	Promptable segmentation	384
UniBrain	ResNet-34 (CNN)	Brain MRI	Knowledge-enhanced pre-training	768
ResNet-34 (control)	ResNet-34 (CNN)	none (random init)	none (no pretraining)	768

Table 2. Datasets and patient characteristics.

Characteristic	UPENN-GBM ($n = 611$, train)	UCSF-PDGM ($n = 400$)	CFB-GBM ($n = 201$)	BraTS-TCGA-GBM ($n = 97$)
Country	USA	USA	France	USA (multi)
Age, median (range)	64 (19–88)	61 (17–94)	70 (25–85)	60 (19–85)
Male, n (%)	367 (60)	241 (60)	124 (62)	61 (63)
MGMT methylated	18%	70%	N/A	N/A
IDH mutated	2.6%	5.0%	N/A	N/A
OS, median (IQR)	403 (182–684)	369 (173–662)	336 (161–504)	384 (209–634)
Censoring	3.8%	42.2%	0%	8.2%

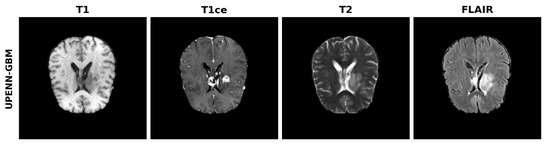


Fig 2. Sample image for UPENN-GBM.

RESULTS

Table 3. Per-model performance, C-index and 1-year/2-year AUC. UPENN-GBM is within-site 5-fold cv; the other cohorts are cross-site.

Model	UPENN-GBM (n=611)		UCSF-PDGM (n=400)		CFB-GBM (n=201)		BraTS-TCGA-GBM (n=97)	
	C-index	AUC (1y/2y)	C-index	AUC (1y/2y)	C-index	AUC (1y/2y)	C-index	AUC (1y/2y)
FM Ensemble	**0.684**	**0.758 / 0.752**	0.656	0.723 / 0.717	**0.608**	**0.679 / 0.652**	**0.641**	**0.686 / 0.703**
BrainIAC	0.650	0.719 / 0.703	0.652	0.723 / 0.709	0.595	0.666 / 0.649	0.616	0.662 / 0.659
SAM-Med3D	0.647	0.692 / 0.711	0.645	0.709 / 0.711	0.606	0.664 / 0.629	0.641	0.687 / 0.699
VoCo	0.635	0.693 / 0.699	0.640	0.697 / 0.699	0.597	0.668 / 0.644	0.639	0.685 / 0.685
BrainSegFounder	0.603	0.628 / 0.664	0.657	0.723 / 0.708	0.590	0.649 / 0.628	0.626	0.646 / 0.708
UniBrain	0.645	0.693 / 0.693	0.659	0.727 / 0.713	0.600	0.654 / 0.641	0.640	0.692 / 0.698
ResNet-34 (no pretrain)‡	0.630	0.672 / 0.688	0.627	0.681 / 0.661	0.568	0.628 / 0.603	0.613	0.646 / 0.631
Clinical-Only	0.650	0.709 / 0.708	0.659	0.727 / 0.717	0.606	0.666 / 0.643	0.634	0.687 / 0.693
FM+Rad+Clin (combined)§	**0.690**	**0.769 / 0.764**	**0.663**	**0.728 / 0.732**	**0.617**	**0.673 / 0.640**	**0.683**	**0.763 / 0.781**

Fig 3. Calibration (Brier score over time) for the Cox FM Ensemble (solid) and the combined FM+radiomics+clinical model (dashed).

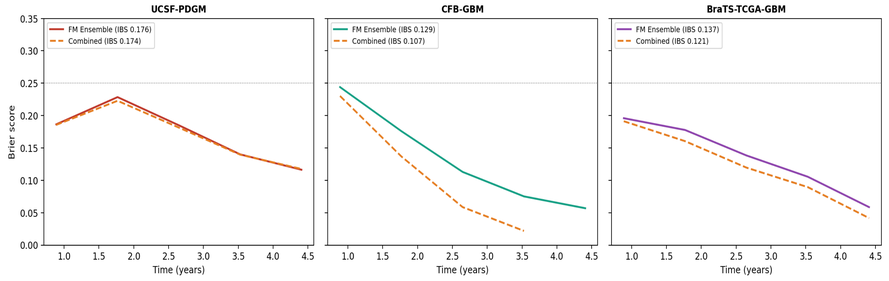
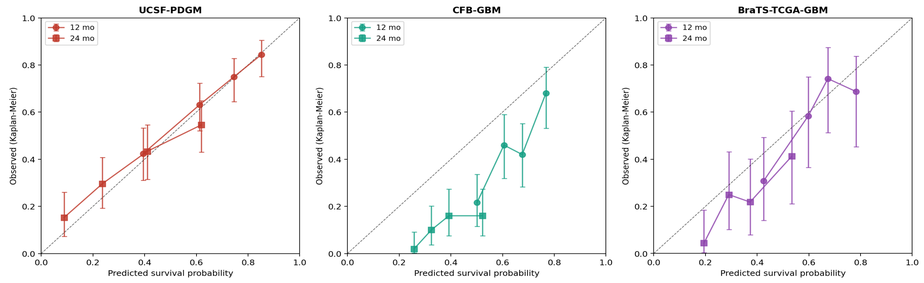


Fig 4. Calibration reliability diagrams for the Cox FM Ensemble (cross-site)



DISCUSSION

Frozen 3D medical imaging foundation models
- Demonstrated **robust** transferability for GBM survival prediction across multiple external cohorts without retraining.
- Foundation model ensemble **consistently provided the most stable performance** and maintained strong discrimination despite differences in acquisition protocols, institutions, and available clinical information, indicating that large-scale pre-training captures transferable prognostic imaging representations.
- Among the individual backbones, performance varied across cohorts, whereas ensembling multiple foundation models improved both robustness and calibration by leveraging complementary feature representations.
- The **ensemble also demonstrated stable calibration** across external datasets, supporting its potential for reliable individualized risk estimation under domain shift.

Although **imaging features did not significantly outperform routine clinical variables**, they achieved **comparable performance without requiring tumor segmentation** and remained applicable when molecular biomarkers were unavailable.

Several **limitations** should be acknowledged.
- The external cohorts lacked complete molecular information.
- The frozen feature representations remain difficult to interpret
- The current late-fusion strategy used a simple unweighted combination.
- Future work should investigate interpretable foundation model representations and learned multimodal fusion approaches.

Table 4. Single-view models vs. the combined late-fusion model. External rows are cross-site (UPenn→external); the UPENN-GBM row is within-site five-fold cross-validation.

Setting	FM Ensemble	Radiomics	Clinical	FM+Rad+Clin
UPENN-GBM	0.684	0.618	0.650	0.690
UCSF-PDGM	0.656	0.596	0.659	0.663
CFB-GBM	0.615	0.554	0.604	0.617
BraTS-TCGA-GBM	0.641	0.670	0.634	0.683