Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Journal Club : Nov 3, 2023

Empirical evaluation of language modeling to ascertain cancer outcomes from clinical text reports

From BMC Bioinformatics (2023) 24:328

Ekapob Sangariyavanich, M.D.
Presenter

Wisdom of the Land

Topic

- Background
- Method
- Results
- Discussion
- Conclusion

Background

Background

- Precision oncology aims to personalize cancer treatment based on individual patient and tumor characteristics.
- This strategy relies on connecting molecular tumor data with patient outcomes.
- However, the clinical outcomes for precision cancer research remains a significant challenge, as essential information is typically documented in free-text reports by radiologists and oncologists during routine care.

- Retrieving these outcomes from electronic health records (EHR) has traditionally been a labor-intensive manual process which it was often carried out by individual research groups without consistent data standards, resulting in datasets of uncertain applicability.
- Our research group developed the ‘PRISSMM’ framework for EHR review.
- PRISSMM is a structured rubric for manual annotation of each pathology, radiology/imaging, and medical oncologist report to ascertain cancer features and outcomes; each imaging report is reviewed in its own right to determine whether it describes cancer response, progression, or neither.

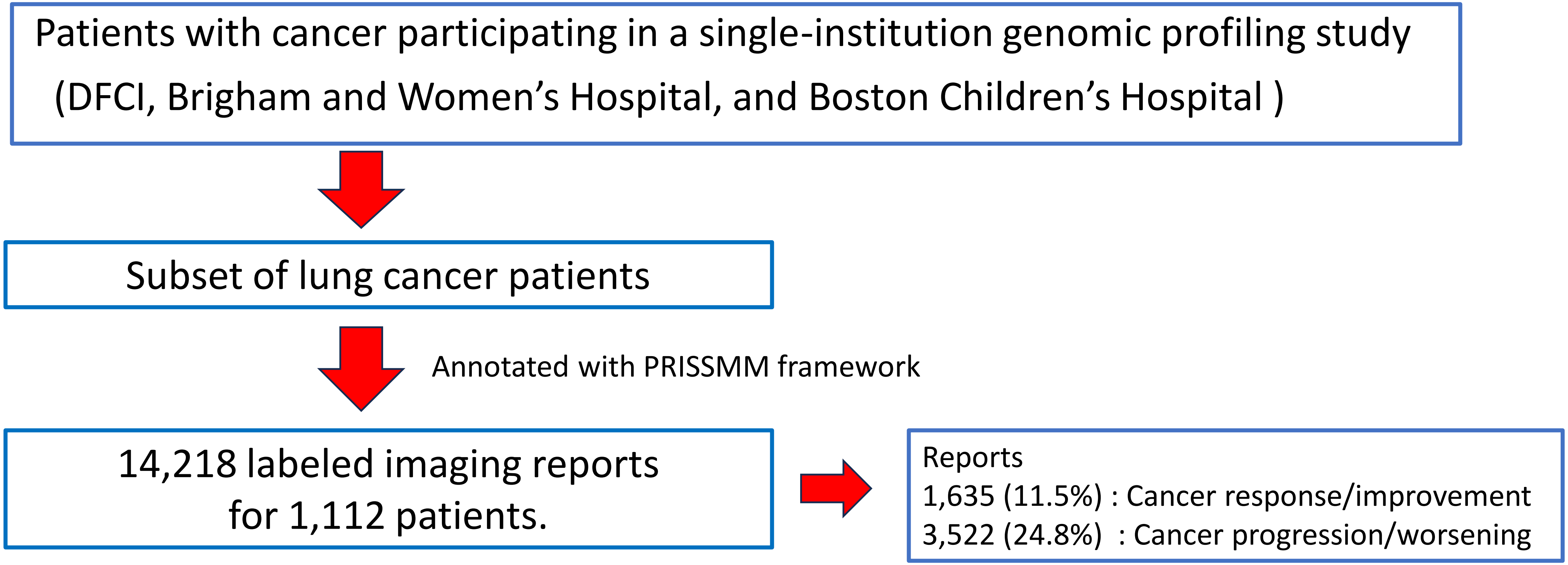
- Applying NLP to clinical documents can expedite outcome determination but traditionally demands a substantial amount of manually annotated data for model training.
- Modern NLP techniques, including semi-supervised learning and Language Model Fine-Tuning method show promise in reducing the need for extensive data labeling.
- The Transformer architecture and derivatives such as BERT have enabled the creation of powerful, large language models for clinical text processing.

- Transformer-based models have been employed in radiology reports and have shown superior performance compared to simpler methods for various general medical annotation tasks.
- Transformer-based models also support the paradigm of zero-shot learning.
- For question-answering tasks, the Text-to-Text Transfer Transformer with instruction fine-tuning has delivered impressive results, even with reasonably sized models.

- Previous research has not produced strong results on biomedical NLP tasks when using general large language models (LLMs).
- It is uncertain whether these models are practically useful for determining cancer outcomes in clinical research settings when dealing with a limited amount of labeled electronic health record (EHR) text data.
- In this study, we evaluated the performance of various NLP architectures at capturing cancer response and progression from imaging reports for a cohort of patients with lung cancer.

Methods

Cohort



Patients with cancer participating in a single-institution genomic profiling study (DFCI, Brigham and Women's Hospital, and Boston Children's Hospital)

Subset of lung cancer patients

Annotated with PRISSMM framework

14,218 labeled imaging reports for 1,112 patients.

Reports
1,635 (11.5%) : Cancer response/improvement
3,522 (24.8%) : Cancer progression/worsening

PRISSMM framework

- A structured framework for curation of clinical outcomes among patients with solid tumors using medical records data.
- If the imaging report indicated any cancer :

  1) responding/improving

  2) progressing/worsening

  3) stable (neither improving nor worsening)

  4) mixed (with some areas improving and some worsening)

  5) indeterminate (if assigning a category was not possible due to radiologist uncertainty or other factors)
- If the imaging report not indicated any cancer : no cancer

- For NLP model training

  1) responding/improving

  2) progressing/worsening

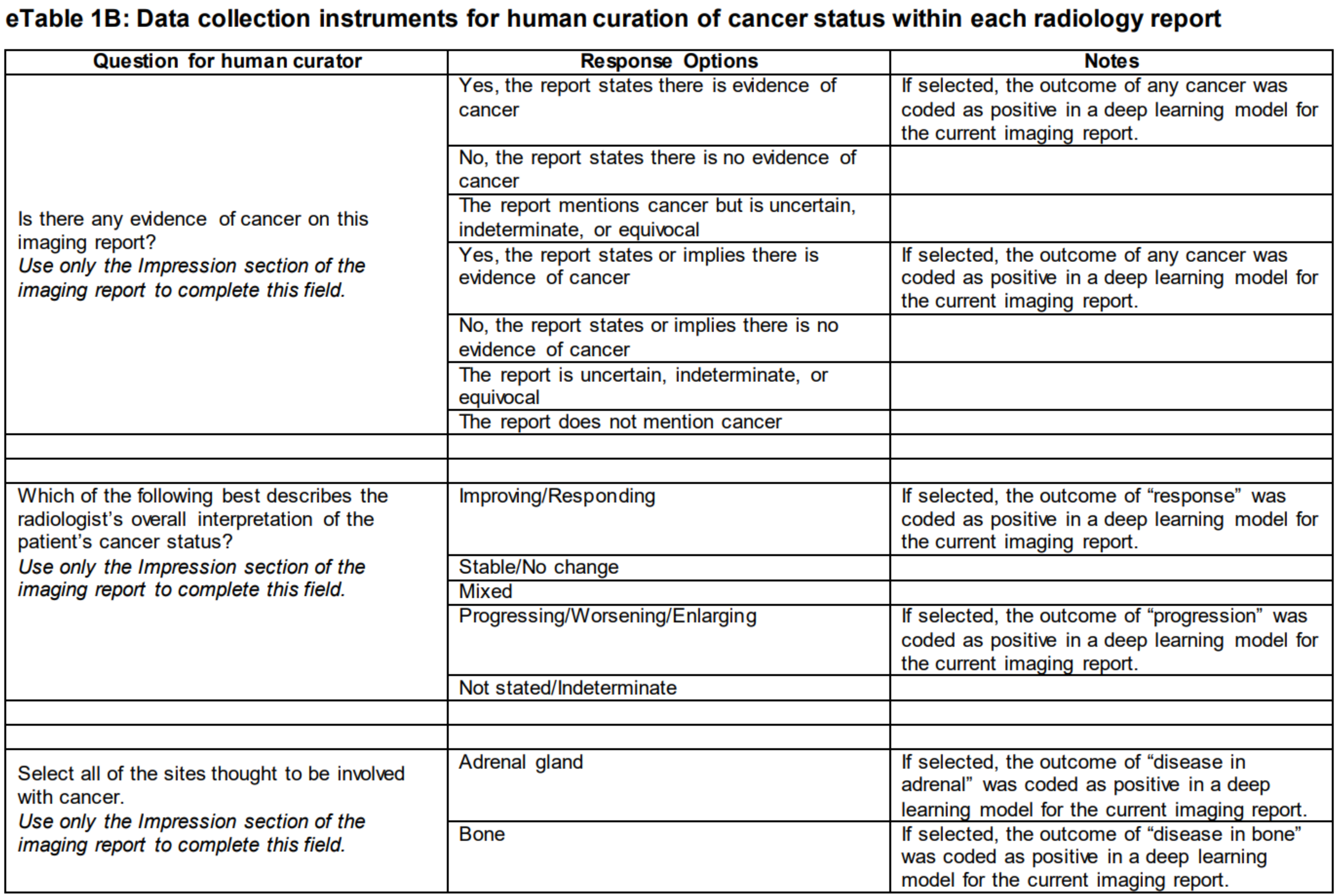
  3) stable (neither improving nor worsening)

  4) mixed (with some areas improving and some worsening)

  5) indeterminate (if assigning a category was not possible due to radiologist uncertainty or other factors)

  6) no cancer

Coded as
“neither improving
nor worsening”



eTable 1B: Data collection instruments for human curation of cancer status within each radiology report

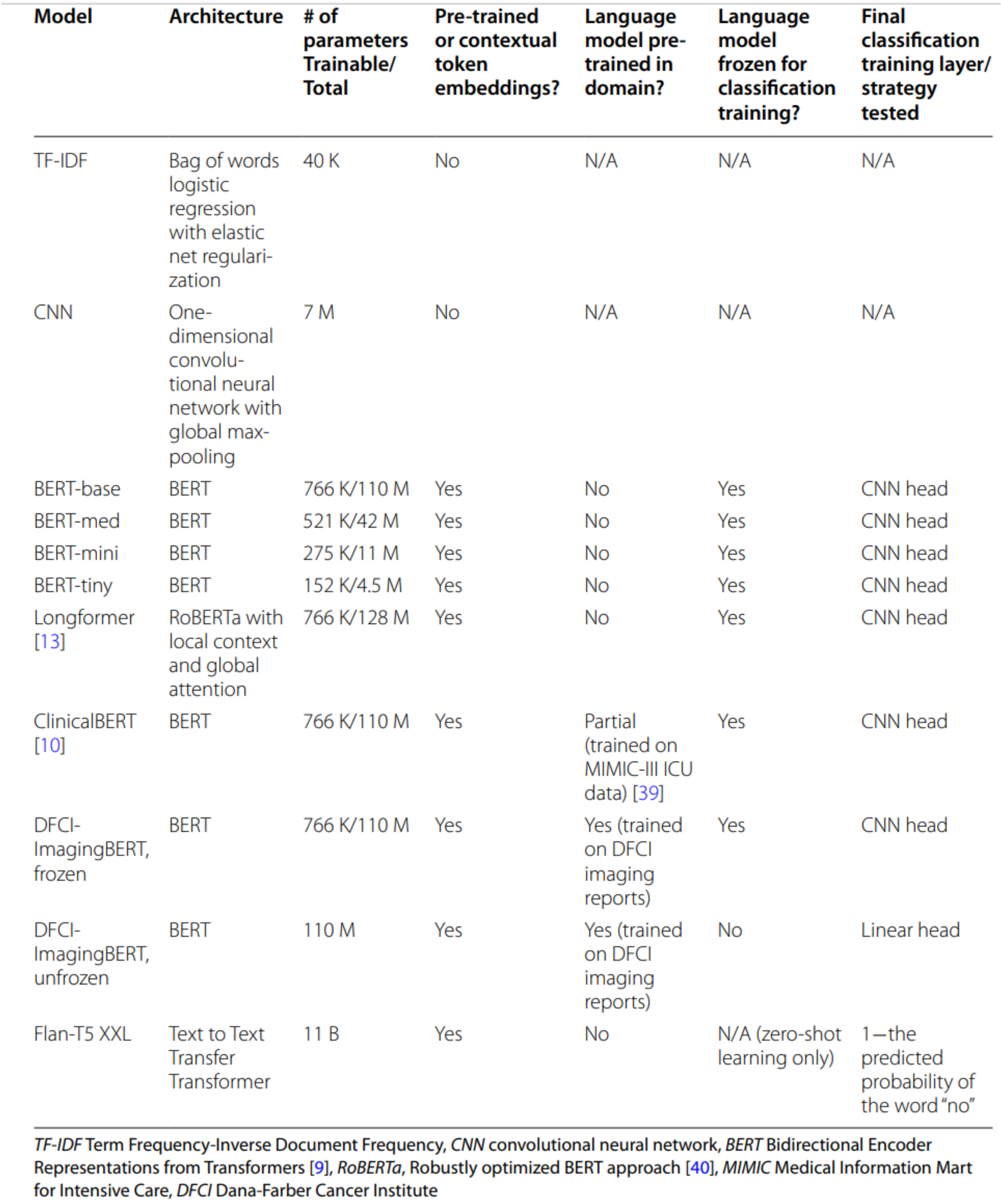
Question for human curator	Response Options	Notes
Is there any evidence of cancer on this imaging report? *Use only the Impression section of the imaging report to complete this field.*	Yes, the report states there is evidence of cancer	If selected, the outcome of any cancer was coded as positive in a deep learning model for the current imaging report.
	No, the report states there is no evidence of cancer	
	The report mentions cancer but is uncertain, indeterminate, or equivocal	
	Yes, the report states or implies there is evidence of cancer	If selected, the outcome of any cancer was coded as positive in a deep learning model for the current imaging report.
	No, the report states or implies there is no evidence of cancer	
	The report is uncertain, indeterminate, or equivocal	
	The report does not mention cancer	
Which of the following best describes the radiologist's overall interpretation of the patient's cancer status? *Use only the Impression section of the imaging report to complete this field.*	Improving/Responding	If selected, the outcome of "response" was coded as positive in a deep learning model for the current imaging report.
	Stable/No change	
	Mixed	
	Progressing/Worsening/Enlarging	If selected, the outcome of "progression" was coded as positive in a deep learning model for the current imaging report.
	Not stated/Indeterminate	
Select all of the sites thought to be involved with cancer. *Use only the Impression section of the imaging report to complete this field.*	Adrenal gland	If selected, the outcome of "disease in adrenal" was coded as positive in a deep learning model for the current imaging report.
	Bone	If selected, the outcome of "disease in bone" was coded as positive in a deep learning model for the current imaging report.

From: Kehl KL, Elmarakeby H, Nishino M, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol.* 2019;5(10):1421–1429. doi:10.1001/jamaoncol.2019.1800

Models

1) Logistic regression with TF-IDF (term frequency-inverse document frequency) vectorization ----Baseline model
2) One-dimension convolutional neural networks (CNN)
3) Transformer-based networks
   - BERT-base, BERT-med, BERT-mini, BERT-tiny
   - Longformer
   - Clinical BERT
   - DFCI-imaging BERT ---- Author’s model
   - Flan-T5 XXL

Model characteristics in this study



Model	Architecture	# of parameters Trainable/Total	Pre-trained or contextual token embeddings?	Language model pre-trained in domain?	Language model frozen for classification training?	Final classification training layer/strategy tested
TF-IDF	Bag of words logistic regression with elastic net regularization	40 K	No	N/A	N/A	N/A
CNN	One-dimensional convolutional neural network with global max-pooling	7 M	No	N/A	N/A	N/A
BERT-base	BERT	766 K/110 M	Yes	No	Yes	CNN head
BERT-med	BERT	521 K/42 M	Yes	No	Yes	CNN head
BERT-mini	BERT	275 K/11 M	Yes	No	Yes	CNN head
BERT-tiny	BERT	152 K/4.5 M	Yes	No	Yes	CNN head
Longformer [13]	RoBERTa with local context and global attention	766 K/128 M	Yes	No	Yes	CNN head
ClinicalBERT [10]	BERT	766 K/110 M	Yes	Partial (trained on MIMIC-III ICU data) [39]	Yes	CNN head
DFCI-ImagingBERT, frozen	BERT	766 K/110 M	Yes	Yes (trained on DFCI imaging reports)	Yes	CNN head
DFCI-ImagingBERT, unfrozen	BERT	110 M	Yes	Yes (trained on DFCI imaging reports)	No	Linear head
Flan-T5 XXL	Text to Text Transfer Transformer	11 B	Yes	No	N/A (zero-shot learning only)	1—the predicted probability of the word "no"

TF-IDF Term Frequency-Inverse Document Frequency, *CNN* convolutional neural network, *BERT* Bidirectional Encoder Representations from Transformers [9], *RoBERTa*, Robustly optimized BERT approach [40], *MIMIC* Medical Information Mart for Intensive Care, *DFCI* Dana-Farber Cancer Institute

Model training

- The full text of radiology reports (the findings concatenated to the impression) was used for each model
- For classification models, separate binary prediction models were trained to identify response/improvement and progression/worsening.
- For BERT model domain adaptation on imaging reports from our institution (DFCI-ImagingBERT)
  - The base model was BERT-base
  - Pre-training was performed over 10 epochs, which took 10.5 days on a single machine equipped with an NVIDIA Tesla T4 GPU (16 GB GDDR6).

- We trained each model using fixed samples of reports from the training set:

  10, 30, 50, 70, 100, 200, 300, 500, 700, or 884 patients.
- For BERT-based models, experiments were also conducted to examine :

  - the effect of various classification head architectures, including linear, convolutional, and recurrent neural network architectures

  - the impact of freezing the weights of the underlying language model when fine-tuning for classification.
- Sequence length

  - BERT-based models : 512 tokens of each report were used.

  - The Longformer model : 1024 tokens were used.

  - The simple CNN model : 1000 tokens was used.

  Reports shorter than the maximum length were padded to the maximum length.

- Zero-shot learning using the T5 encoder-decoder model (the Flan-T5-XXL model) :

Template	Improvement task question	Progression task question
“question: {question} context: {imaging report}”	"Is there improvement/response/shrinking of cancer (yes/no)?"	"Is there worsening of cancer (yes/no)?."

- Imaging report: full input text without truncation.
- To determine the classification output

  - extracted the first token logits of the generated output text.

  - the probability of the "yes" class was computed using softmax.

Model evaluation

- Classification performance for outcomes:

  (1) cancer response

  (2) cancer progression was evaluated using

- Metrics:

  - The area under the receiver operating characteristic curve (AUROC)

  - The area under the precision-recall curve (AUPRC).

  - Accuracy

  - Precision

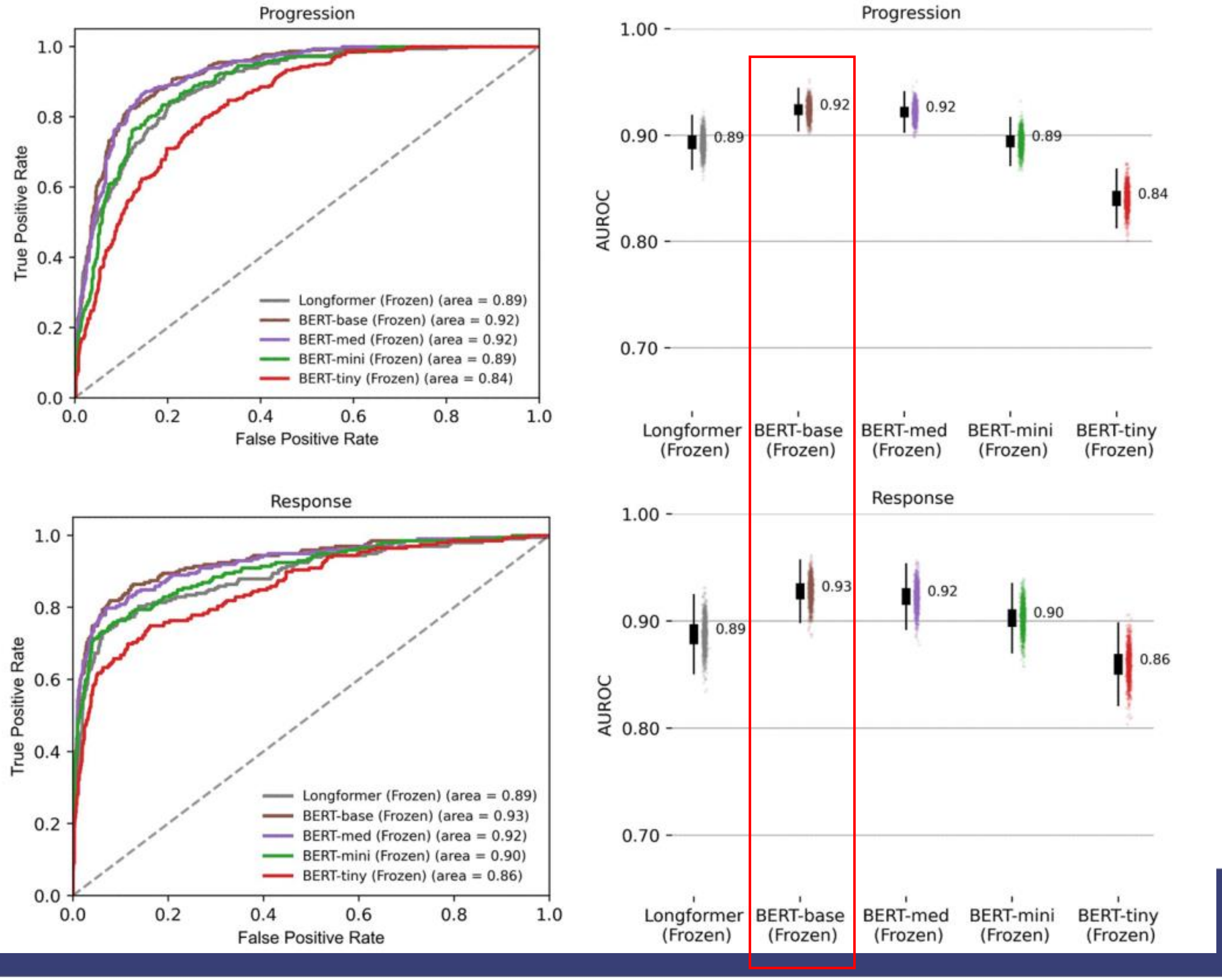
  - Recall

  - Matthew correlation coefficient (MCC)

  - F1 score

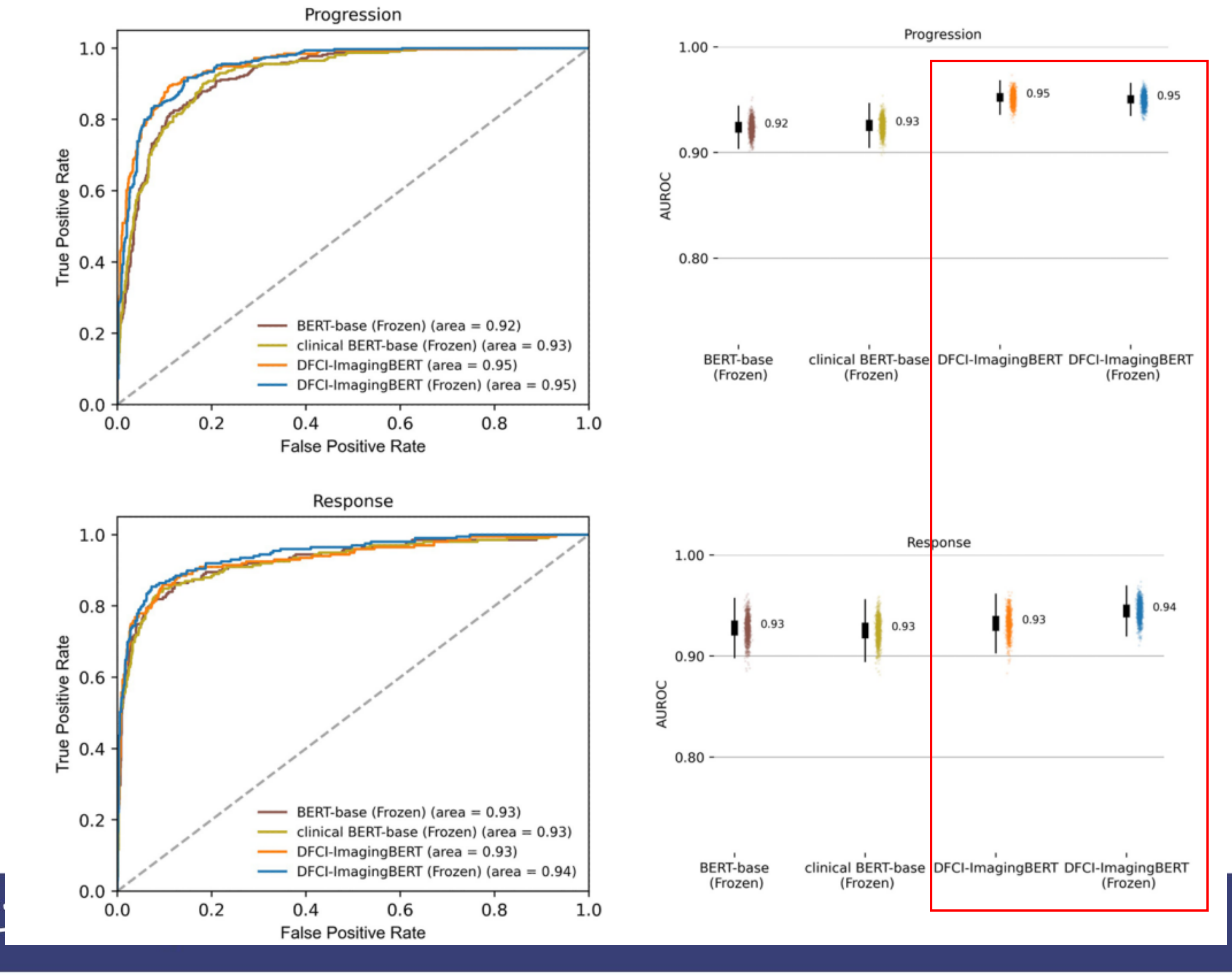
Results

1. Domain adaptation



1.1 No domain adaptation

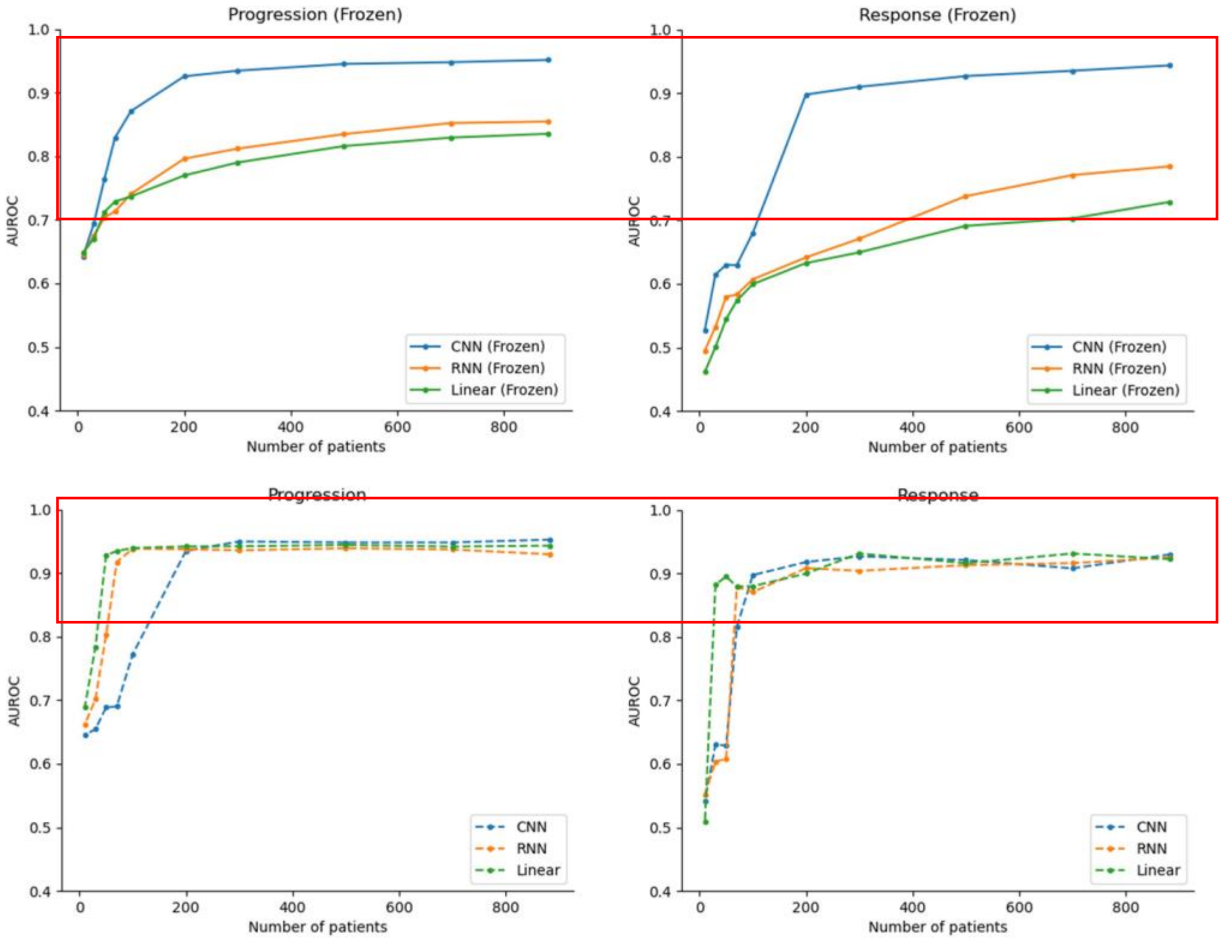
- BERT-base was outperformed



1.2 Domain adaptation

- DFCI-ImagingBERT (frozen and fine-tuned) were outperformed

2. Impact of classification layer



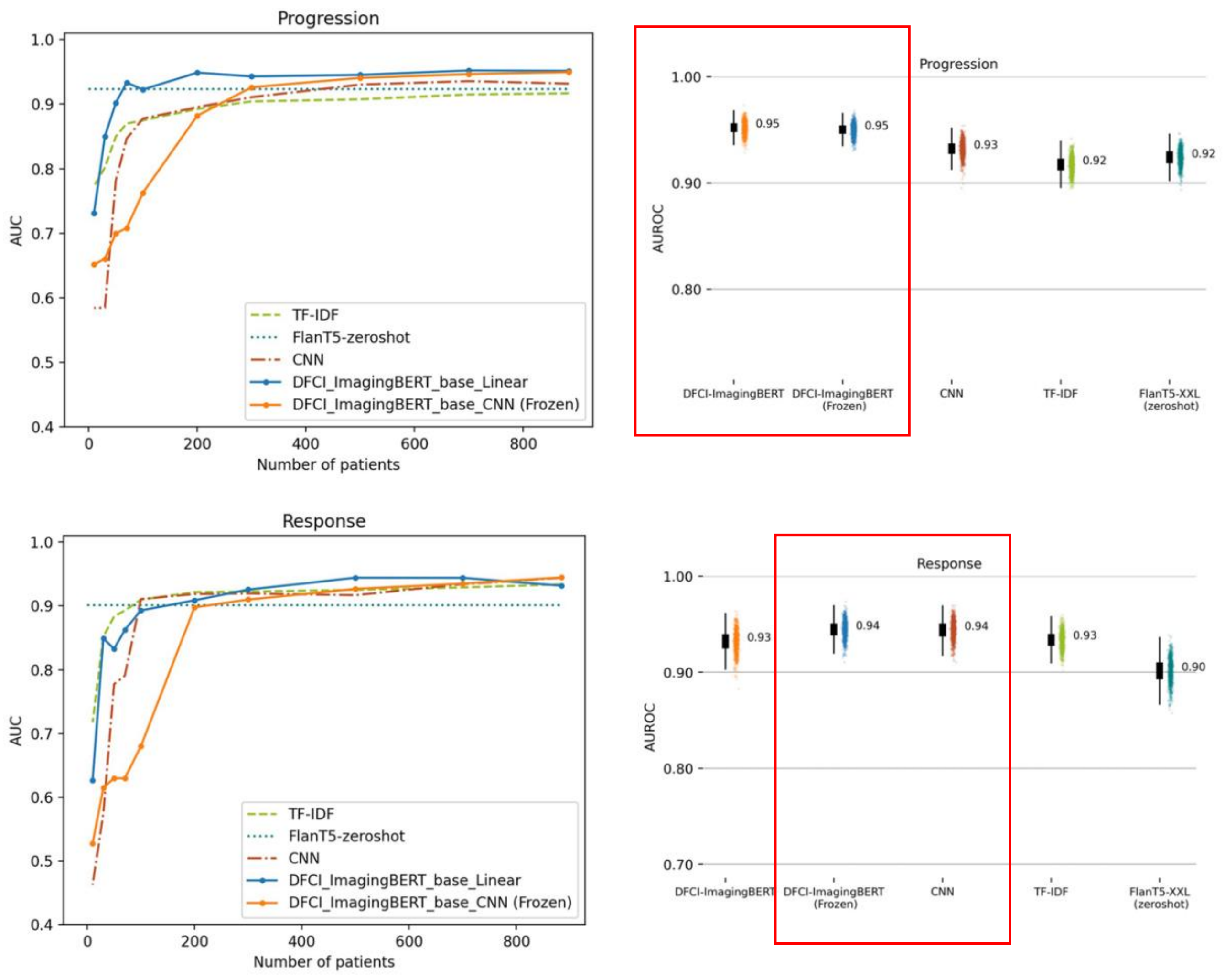
2.1 Frozen

The CNN head was associated with the best performance at the largest training set size.

2.2 Unfrozen

The linear head yielded AUROCs of 0.92 and 0.94, and the CNN was best with AUROCs of 0.93 and 0.95 for response and progression outcomes respectively.

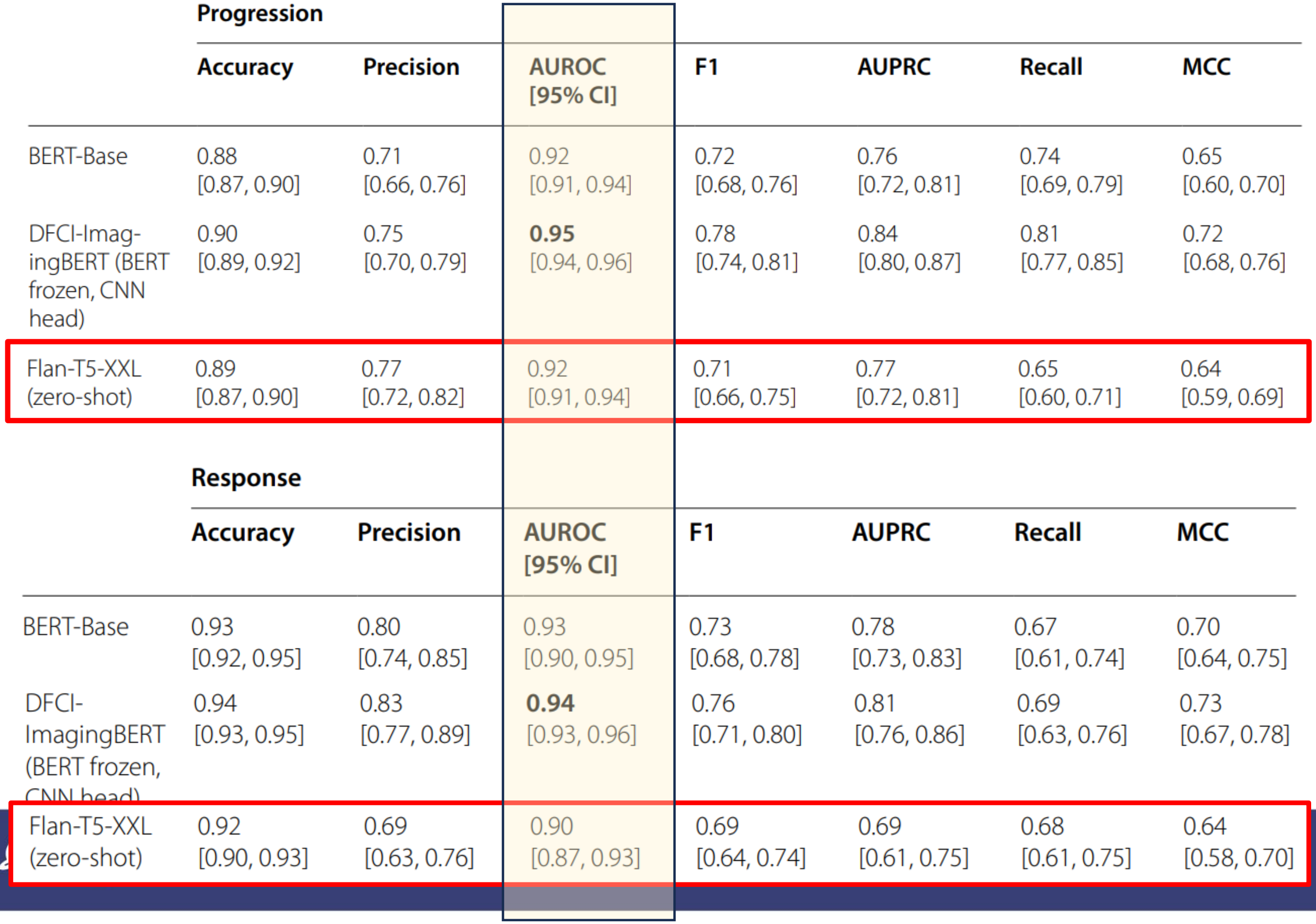
3. Compare DFCI-ImagingBERT to baseline model



- For the progression/worsening outcome : DFCI-ImagingBERT yielded the best performance

- For the response/improvement outcome: DFCI-ImagingBERT with a frozen language model and CNN classification head yielded the best performance

4. Zero shot learning



Progression

	Accuracy	Precision	AUROC [95% CI]	F1	AUPRC	Recall	MCC
BERT-Base	0.88 [0.87, 0.90]	0.71 [0.66, 0.76]	0.92 [0.91, 0.94]	0.72 [0.68, 0.76]	0.76 [0.72, 0.81]	0.74 [0.69, 0.79]	0.65 [0.60, 0.70]
DFCI-Imaging BERT (BERT frozen, CNN head)	0.90 [0.89, 0.92]	0.75 [0.70, 0.79]	**0.95** [0.94, 0.96]	0.78 [0.74, 0.81]	0.84 [0.80, 0.87]	0.81 [0.77, 0.85]	0.72 [0.68, 0.76]
Flan-T5-XXL (zero-shot)	0.89 [0.87, 0.90]	0.77 [0.72, 0.82]	0.92 [0.91, 0.94]	0.71 [0.66, 0.75]	0.77 [0.72, 0.81]	0.65 [0.60, 0.71]	0.64 [0.59, 0.69]

Response

	Accuracy	Precision	AUROC [95% CI]	F1	AUPRC	Recall	MCC
BERT-Base	0.93 [0.92, 0.95]	0.80 [0.74, 0.85]	0.93 [0.90, 0.95]	0.73 [0.68, 0.78]	0.78 [0.73, 0.83]	0.67 [0.61, 0.74]	0.70 [0.64, 0.75]
DFCI-ImagingBERT (BERT frozen, CNN head)	0.94 [0.93, 0.95]	0.83 [0.77, 0.89]	**0.94** [0.93, 0.96]	0.76 [0.71, 0.80]	0.81 [0.76, 0.86]	0.69 [0.63, 0.76]	0.73 [0.67, 0.78]
Flan-T5-XXL (zero-shot)	0.92 [0.90, 0.93]	0.69 [0.63, 0.76]	0.90 [0.87, 0.93]	0.69 [0.64, 0.74]	0.69 [0.61, 0.75]	0.68 [0.61, 0.75]	0.64 [0.58, 0.70]

Flan-T5-XXL model achieved AUROC of 0.92 for the progression/worsening task and AUROC of 0.90 for the response/improvement task.

Discussion

- NLP has the potential to substantially accelerate precision oncology research by enabling observational clinical outcomes to be linked to molecular cancer data for downstream analysis,
- Transformer-based models have become standard for general NLP tasks given their potential to yield improved performance.
- We found that a BERT model with domain adaptation on text from our institution performed better than simpler TF-IDF and CNN models for text classification.

- But the simple models still yielded AUROCs > 0.9, complex models may not always be needed if training data are readily available.
- On the other hand, we found that the Flan-T5-XXL architecture with a small amount of prompt engineering yielded good zero-shot performance with no domain adaptation pretraining or fine-tuning on labeled data.
- Demonstrating the potential utility of large language models in this space when computational resources are readily available.

- Potential explanations for the similar performance observed between a transformer architecture and simpler models.

  1) The outcomes are distinctly keyword-sensitive.

  2) Clinical imaging reports are also substantially longer than typical sequence lengths for standard Transformer models (dilute the benefits derived from contextual token embedding)

- Strength

  1) Labels in this dataset have been shown to be clinically relevant and associated with overall survival.

  2) Our results provide practical guidance to researchers who may seek to gather just the necessary volume of labeled clinical data in order to train NLP models to perform cancer outcome extraction.

     (Model performance improves with greater training set size, but that the marginal improvement once the training set reached 300 patients (~3000 imaging reports) was relatively small.)

- Limitations

  1) Single-institution nature of the data

  2) Limited hyperparameter tuning

  3) No external generalizability evaluation

- NLP models trained on institutional protected health information may carry at least some risk of exposing that information to adversarial attacks, and further research into best practices for generalizable cross-institution NLP healthcare modeling is needed.

Conclusion

- BERT model with domain adaptation and supervised fine-tuning for classification yielded the best performance across tasks and metrics.
- Simpler models demonstrated good performance given large quantities of training data.
- Zero-shot learning based on modern large language models also demonstrated good performance on some metrics.

- If computational resources are plentiful but labeled training data are limited
  → large language models can be used for zero- or few-shot learning

- When computational resources are more limited but labeled training data are readily available → simple machine learning architectures

THE END