Memorial Sloan Kettering Cancer Center

A.I.-Assisted Clinical Data Curation to Determine Genomic Biomarkers of Cancer Metastasis

Anisha Luthra, Karl Pichotta, Brooke Mastrogiacomo, Samantha McCarthy, Steven Maron, Jianjiong Gao, Francisco Sanchez-Vega, Justin Jee, Christopher Fong, Nikolaus Schultz

Background

• While progression to metastatic disease is the main cause of cancer death, **little is known about the genomic mechanisms that drive metastasis**.

• Rapidly growing clinical genomic data sets have the potential to identify genomic biomarkers of cancer metastasis, however, **manual curation of clinical data is quickly emerging as a bottleneck**.

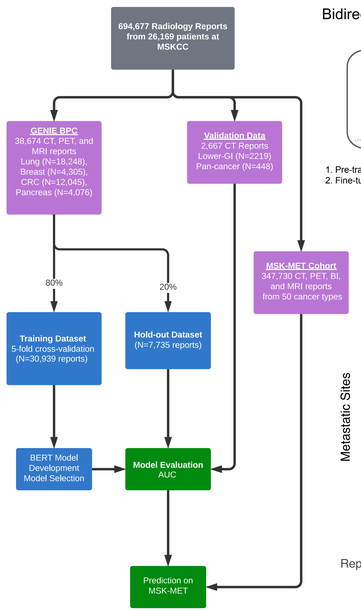
• To overcome this challenge, we have **developed a natural language processing (NLP) pipeline using ClinicalBERT**[1] to identify organs affected by metastasis from radiology reports of patients with cancer.

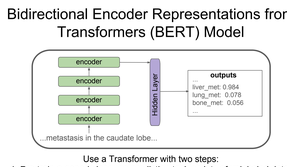
• MSK-MET[2] automates process of extracting information from **billing codes and sequenced metastases**; however, these sources have missing data. This is where radiology reports can supplement information.

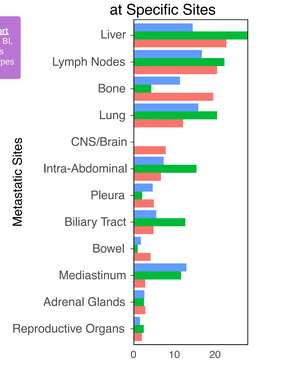
• To develop our NLP models, we leveraged the **AACR GENIE Biopharma Collaborative lung, colorectal, breast, and pancreatic cancer datasets** generated in part at Memorial Sloan Kettering Cancer Center (MSK), containing curated labels of **12 metastatic disease sites**.

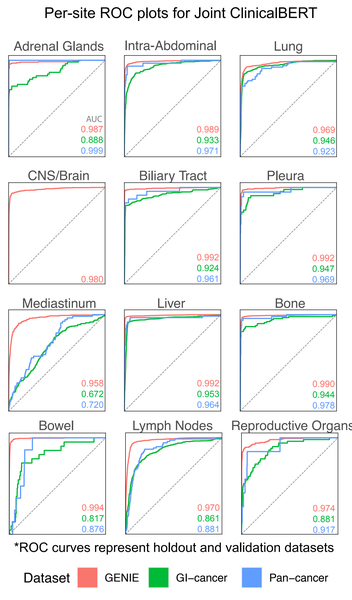
• We **validated our models on lower-GI and pan-cancer datasets**[3] consisting of CT reports curated at Memorial Sloan Kettering Cancer Center.

Methods

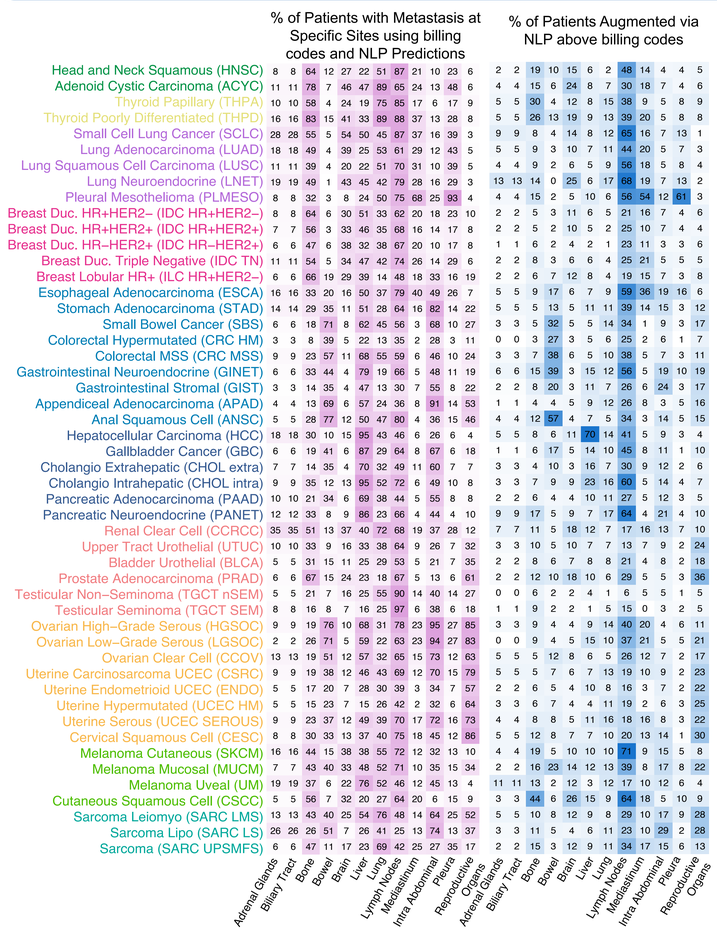


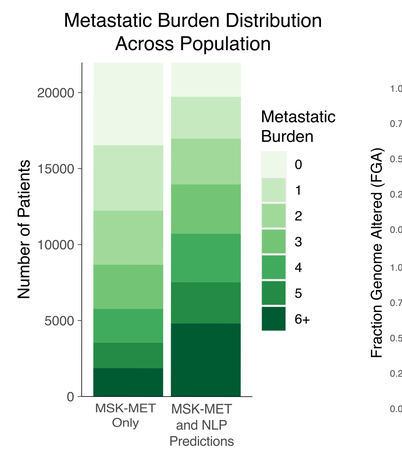


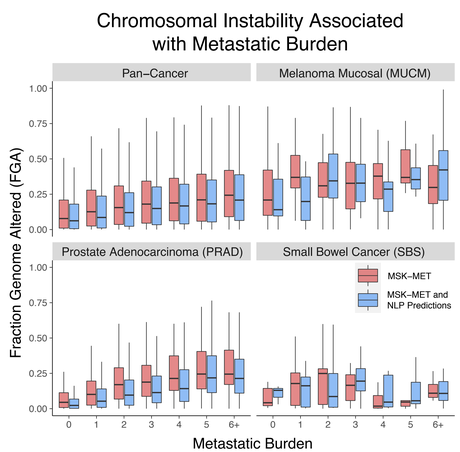


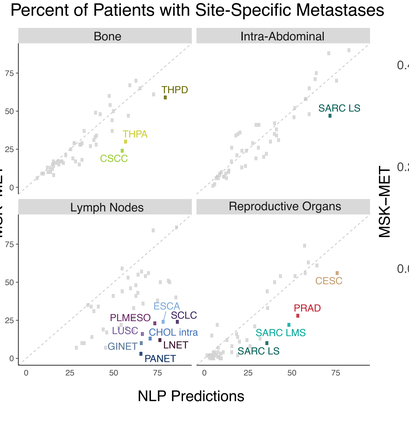


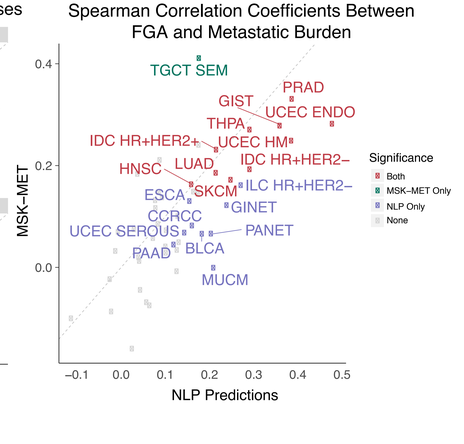
Results











Conclusions

• We confirmed that chromosomal instability, as inferred by the fraction of genome altered (FGA), is correlated with metastatic burden (defined as the number of distinct organs affected by metastases) in several tumor types

• Our models, applied at scale, offer a unique resource for the investigation of the biological basis for metastatic spread.

• Our models can supplement billing codes and sequenced metastases.

• Our automated clinical data extractions can enable further large-scale studies of associations between genomic biomarkers and metastatic behavior.

1 Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.
2 Nguyen, B., Fong, C., Luthra, A., … Sanchez-Vega F., Schultz, N. (2022). Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell, 185(3), 563–575.e11. https://doi.org/10.1016/j.cell.2022.01.003
3 Do R. K. G., Lupton K., Causa Andrieu P. I., Luthra A., Taya M., Batch K., et al. . (2021). Patterns of metastatic disease in patients with cancer derived from natural language processing of structured CT Radiology reports over a 10-year period. Radiology 301, 115–122. 10.1148/radiol.2021210043

Abstract #1158