Multi-cancer Brain Metastasis Risk Score Development and Validation using 220,246 Whole Transcriptomes and Machine Learning

Background
Brain metastases occur in multiple cancer types with higher prevalence in lung, breast, melanoma, and GI cancers. The prognosis of patients who develop brain metastases is very poor and identification of brain metastasis risk could be useful for prognostication, monitoring, and therapy selection.

Methods
Data from the whole transcriptome of 220,246 tumor profiles were analyzed and multiple machine learning models were trained on various molecular subtypes. The dataset was split 50% for training and the other 50% for testing. UMAP was employed for dimensionality reduction and the patterns learned across the entirety of the training dataset irrespective of brain metastasis were leveraged on the testing data set. Patients with brain metastases were identified using the expression of ICD-10 code C79.31 (Secondary malignant neoplasm of the brain). As the absence of C79.31 could be due to the event not happening set, patients without brain metastases were stratified into groups based on 3, 4 or greater than 5 years without a C79.31 ICD-10 code. The brain metastasis risk score was defined by empirical evaluation of the positive predictive value in 7 groups of risk probabilities. The validation set contained 1,017 patients with brain metastases and 4,631 without an observed brain metastasis within 3 years.

Results
In the validation set, the prevalence of brain metastases within the risk scores across all cancer types ranged from 4% with the lowest risk score to 94% in the highest with 71% of cases receiving the lowest 2 risk scores, 15% the 2 intermediate risk scores, and 14% the 3 highest risk scores. For breast, lung and colon cancers, the prevalence of brain metastasis ranged from 4% with the lowest risk score to 94% in the highest with 71% of cases receiving the lowest 2 risk scores, 15% the 2 intermediate risk scores, and 14% the 3 highest risk scores. In the validation set, the prevalence of brain metastases across all risk scores ranged from 4% with the lowest risk score to 96% in the highest with 71% of cases receiving the lowest 2 risk scores, 15% the 2 intermediate risk scores, and 14% the 3 highest risk scores.

Conclusions
Whole transcriptome data can be leveraged by a machine learning platform that employs dimensionality reduction techniques alongside transfer learning to predict the risk of brain metastasis. This tool can be used to augment the clinical picture of cancer patients and unmet clinical opportunity to inform prognosis, monitoring, and therapeutic selection.

References
1. McInnes, L, Healy, J, ArXiv 2018.1802.03426, 2018
3. Lipton Z, Caruana R, Lee D, NIPS 2013
4. Lipton Z, Caruana R, Lee D, NIPS 2013
5. Lipton Z, Caruana R, Lee D, NIPS 2013

Introduction
Application of AI to large molecular data sets covering transcribed genes is in its infancy. To date, there have not been large enough data sets to take advantage of the power of AI technology to predict disease progression for patients with cancer. Here we show that application of MSK methodologies to a large collection of molecular and clinical data, generates insight into disease progression. While meaningful predictions can be made on cohorts of 100,000 patients, it is clear that more data will enable even more accurate predictions. Generation of WTS data on larger patient cohorts is essential to maximize precision medicine and advance the science and medicine of cancer care. These results can have a direct impact on current patients as well as provide insight into future drug development efforts.