Pan-Cancer Survival Classification with Clinicopathologic and Targeted Gene Expression Features

Patrick Yu, Department of Computer Science, University of Illinois at Urbana-Champaign

16th Annual AACR Undergraduate Student Caucus and Poster Competition

Background

- Many factors influence a cancer patient’s survival rate and outcome - **what** are they and **how** do they affect the patient’s survival?

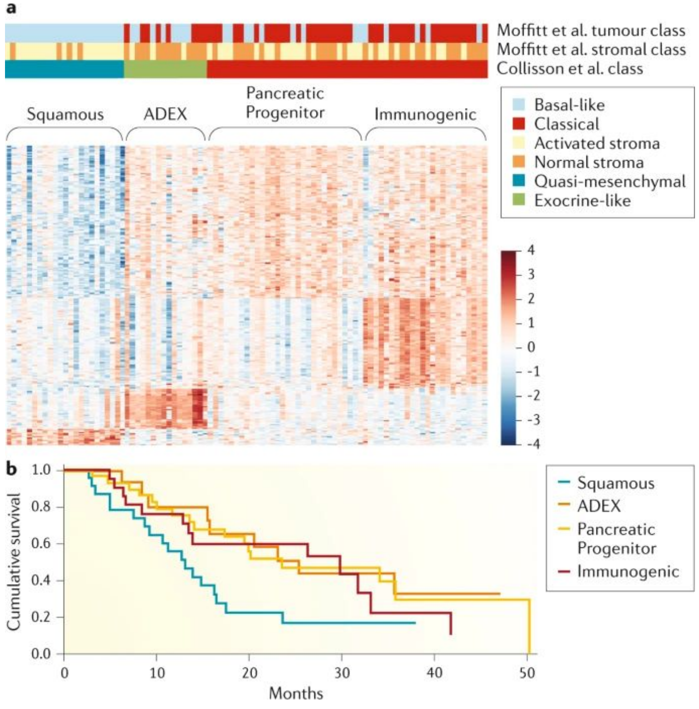
- Types of risk factors to consider:
  - **Clinical/demographic/behavioral** (e.g. age, sex, smoking habits, cancer stage, etc)
  - **Pathological** (e.g. tumor morphology)
  - **Biomolecular** (gene expression, mutations, CNVs, etc)

- The goal is to create a machine learning model that will:
  - 1) Predict Overall Survival (OS) from the considered risk factors, over various cancer types and time points (e.g. at 1 year, 3 year)
  - 2) Identify the most significant risk factors affecting survival outcome

Previous Work

- General *trends* between **molecular** data and survival outcomes have been found, but fail to yield survival outcome prediction at the *individual patient* level
  - Adding on **clinicopathological data** may help to predict patient survival

- Previous studies (3,4) have focused on predicting Overall Survival for individual cancer types
  - Instead, we aim to predict OS at *varying time* points across *various cancer types*



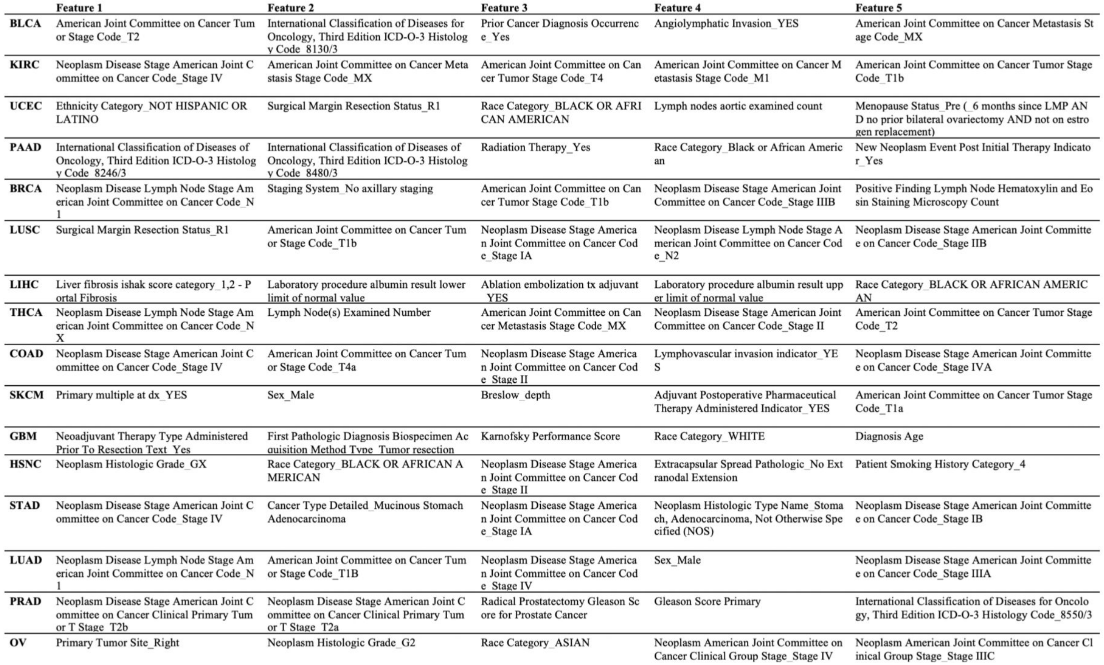
Collisson et al, *Nat Rev Gastro & Hep 2019*

Methods - Overall Survival Prediction

- Our dataset consists of 8,068 patients across 16 cancer types from TCGA
    - Each patient is tied to a set of clinicopathological features

- First, we predicted survival outcomes (at 1-year and 3-year timepoints) with only 15 clinicopathologic features via **Sequential Forward Search (SFS)**

- Then, we sequentially added on expression data from 25 selected genes to optimize model accuracy



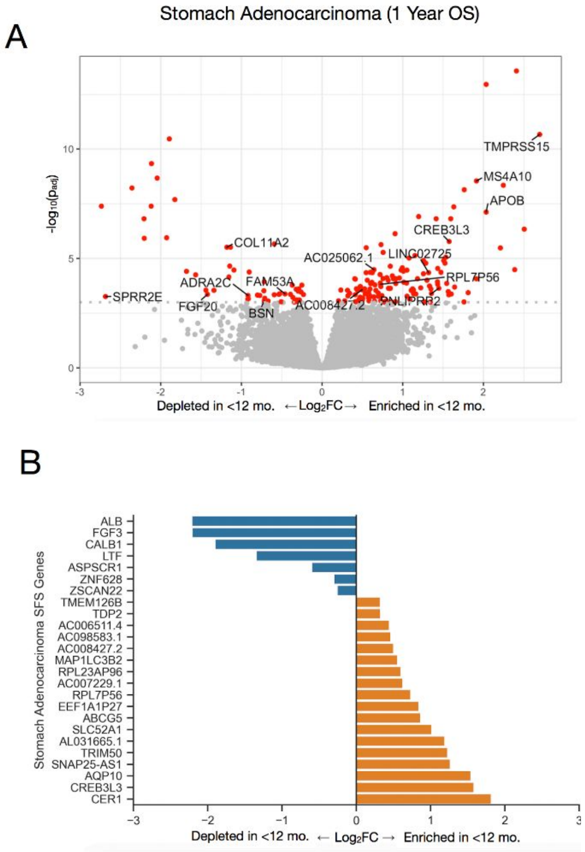
	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
BLCA	American Joint Committee on Cancer Tumor Stage Code_T2	International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code_8130/3	Prior Cancer Diagnosis Occurrence_Yes	Angiolymphatic Invasion_YES	American Joint Committee on Cancer Metastasis Stage Code_MX
KIRC	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IV	American Joint Committee on Cancer Metastasis Stage Code_MX	American Joint Committee on Cancer Tumor Stage Code_T4	American Joint Committee on Cancer Metastasis Stage Code_M1	American Joint Committee on Cancer Tumor Stage Code_T1b
UCEC	Ethnicity Category_NOT HISPANIC OR LATINO	Surgical Margin Resection Status_R1	Race Category_BLACK OR AFRICAN AMERICAN	Lymph nodes aortic examined count	Menopause Status_Pre (_6 months since LMP AND no prior bilateral ovariectomy AND not on estrogen replacement)
PAAD	International Classification of Diseases of Oncology, Third Edition ICD-O-3 Histology Code_8246/3	International Classification of Diseases of Oncology, Third Edition ICD-O-3 Histology Code_8480/3	Radiation Therapy_Yes	Race Category_Black or African American	New Neoplasm Event Post Initial Therapy Indicator_Yes
BRCA	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code_N1	Staging System_No axillary staging	American Joint Committee on Cancer Tumor Stage Code_T1b	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IIIB	Positive Finding Lymph Node Hematoxylin and Eosin Staining Microscopy Count
LUSC	Surgical Margin Resection Status_R1	American Joint Committee on Cancer Tumor Stage Code_T1b	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IA	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code_N2	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IIB
LIHC	Liver fibrosis ishak score category_1,2 - Portal Fibrosis	Laboratory procedure albumin result lower limit of normal value	Ablation embolization tx adjuvant_YES	Laboratory procedure albumin result upper limit of normal value	Race Category_BLACK OR AFRICAN AMERICAN
THCA	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code_NX	Lymph Node(s) Examined Number	American Joint Committee on Cancer Metastasis Stage Code_MX	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage II	American Joint Committee on Cancer Tumor Stage Code_T2
COAD	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IV	American Joint Committee on Cancer Tumor Stage Code_T4a	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage II	Lymphovascular invasion indicator_YES	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IVA
SKCM	Primary multiple at dx_YES	Sex_Male	Breslow_depth	Adjuvant Postoperative Pharmaceutical Therapy Administered Indicator_YES	American Joint Committee on Cancer Tumor Stage Code_T1a
GBM	Neoadjuvant Therapy Type Administered Prior To Resection Text_Yes	First Pathologic Diagnosis Biospecimen Acquisition Method Type_Tumor resection	Karnofsky Performance Score	Race Category_WHITE	Diagnosis Age
HSNC	Neoplasm Histologic Grade_GX	Race Category_BLACK OR AFRICAN AMERICAN	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage II	Extracapsular Spread Pathologic_No Extranodal Extension	Patient Smoking History Category_4
STAD	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IV	Cancer Type Detailed_Mucinous Stomach Adenocarcinoma	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IA	Neoplasm Histologic Type Name_Stomach, Adenocarcinoma, Not Otherwise Specified (NOS)	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IB
LUAD	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code_N1	American Joint Committee on Cancer Tumor Stage Code_T1B	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IV	Sex_Male	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IIIA
PRAD	Neoplasm Disease Stage American Joint Committee on Cancer Clinical Primary Tumor T Stage_T2b	Neoplasm Disease Stage American Joint Committee on Cancer Clinical Primary Tumor T Stage_T2a	Radical Prostatectomy Gleason Score for Prostate Cancer	Gleason Score Primary	International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code_8550/3
OV	Primary Tumor Site_Right	Neoplasm Histologic Grade_G2	Race Category_ASIAN	Neoplasm American Joint Committee on Cancer Clinical Group Stage_Stage IV	Neoplasm American Joint Committee on Cancer Clinical Group Stage_Stage IIIC

Top ranked features for 1 year model. Abbreviations: BLCA = bladder urothelial carcinoma, KIRC = kidney clear cell carcinoma, UCEC = uterine corpus endometrial carcinoma, PAAD = pancreatic adenocarcinoma, BRCA = breast invasive carcinoma, LUSC = lung squamous cell carcinoma, LIHC = liver hepatocellular carcinoma, THCA = thyroid carcinoma, COAD = colon adenocarcinoma, SKCM = skin cutaneous melanoma, GBM = glioblastoma multiforme, HSNC = head and neck squamous cell carcinoma, STAD = stomach adenocarcinoma, LUAD = lung adenocarcinoma, PRAD = prostate adenocarcinoma, OV = ovarian serous cystadenocarcinoma.

Methods - Gene Selection

- A differential analysis can reveal the most *significant* genes with the largest expression differences between surviving and deceased cohorts

- To determine which genes to select, we performed a Differential Expression analysis (DESeq2) comparing between patients who survive <1 year vs. >1 year after diagnosis, as well as <3 year vs. >3 year post-diagnosis



Implementation

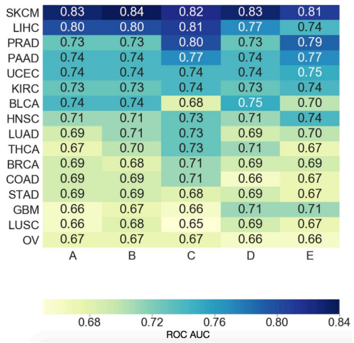
- **Preprocessing:**
  - Imputed missing data from patient with XGBoost's imputation, median, and K-Nearest Neighbors
  - Omitted data missing from >40% of patients & removed features hinting at survival outcomes (e.g. "Disease Free Status", "Overall Survival Status")

- **Model Training:**
  - Tested 40 chosen features (15 clinicopathological vs. 15 clinicopathological + 25 genes) on either XGBoost or Random Forest in predicting Overall Survival with a 80%/20% test-train cross-validation split
  - Implemented a grid search on 2 models (XGBoost, Random Forest) and 3 imputation techniques (XGBoost-imputation, median, and KNN) to find the optimal model

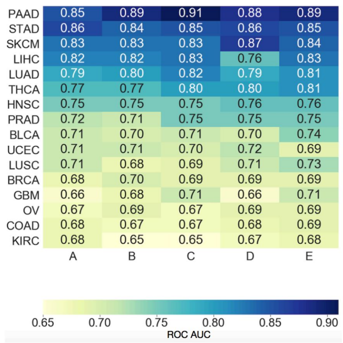
Results - Overall Survival Prediction

- While including the 15 clinical features alone yielded a relatively low AUC measure (~ 0.6-0.7 range) for lower-performing cancers (GBM, OV, etc), some cancers (e.g. PAAD) performed well even *without* the 25 genes

- AUC increased noticeably after including the 25 genes (up to the 0.75-0.78 range)
    - AUC’s for Glioblastoma (GBM), Stomach Adenocarcinoma (STAD), Ovarian Carcinoma (OV) increased from **0.71, 0.62, 0.66** to **0.76, 0.77, and 0.77,** respectively
    - These equate to a ~7 to 23% increase in AUC across the 3 lowest-performing cancers

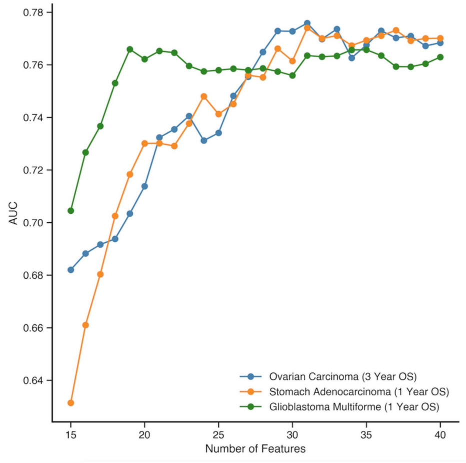


	A	B	C	D	E
SKCM	0.83	0.84	0.82	0.83	0.81
LIHC	0.80	0.80	0.81	0.77	0.74
PRAD	0.73	0.73	0.80	0.73	0.79
PAAD	0.74	0.74	0.77	0.74	0.77
UCEC	0.74	0.74	0.74	0.74	0.75
KIRC	0.73	0.73	0.74	0.73	0.74
BLCA	0.74	0.74	0.68	0.75	0.70
HNSC	0.71	0.71	0.73	0.71	0.74
LUAD	0.69	0.71	0.73	0.69	0.70
THCA	0.67	0.70	0.73	0.71	0.67
BRCA	0.69	0.68	0.71	0.69	0.69
COAD	0.69	0.69	0.71	0.66	0.67
STAD	0.69	0.69	0.68	0.69	0.67
GBM	0.66	0.67	0.66	0.71	0.71
LUSC	0.66	0.68	0.65	0.69	0.67
OV	0.67	0.67	0.67	0.66	0.66



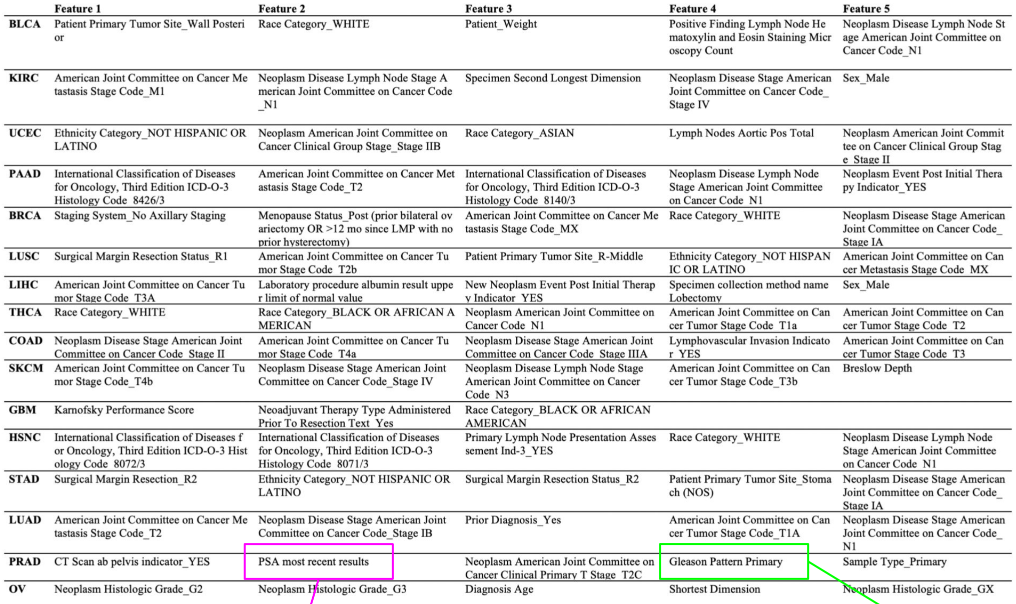
	A	B	C	D	E
PAAD	0.85	0.89	0.91	0.88	0.89
STAD	0.86	0.84	0.85	0.86	0.85
SKCM	0.83	0.83	0.83	0.87	0.84
LIHC	0.82	0.82	0.83	0.76	0.83
LUAD	0.79	0.80	0.82	0.79	0.81
THCA	0.77	0.77	0.80	0.80	0.81
HNSC	0.75	0.75	0.75	0.76	0.76
PRAD	0.72	0.71	0.75	0.75	0.75
BLCA	0.71	0.70	0.71	0.70	0.74
UCEC	0.71	0.71	0.70	0.72	0.69
LUSC	0.71	0.68	0.69	0.71	0.73
BRCA	0.68	0.70	0.69	0.69	0.69
GBM	0.66	0.68	0.71	0.66	0.71
OV	0.67	0.69	0.67	0.69	0.69
COAD	0.68	0.67	0.67	0.68	0.69
KIRC	0.68	0.65	0.65	0.67	0.68

A = XGBoost + XGBoost imputation; **B** = XGBoost + median imputation; **C** = Random Forest + median imputation; **D** = XGBoost + K-nearest neighbors; **E** = Random Forest + K-nearest neighbors



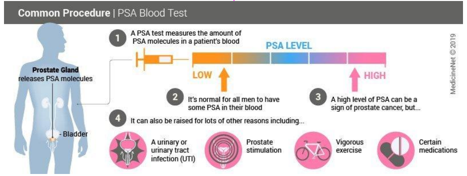
Results - Top Clinical Factors Influencing Survival

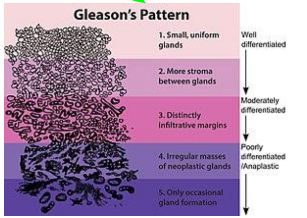
- Our analyses also showed the **top 5 features** that were strongly related to a lower survival in the 1 and 3 year timeframes
    - Many were **disease-specific** features - e.g. for PRAD (Prostate Cancer) our model utilized the **Gleason** prostate biopsy score and PSA (Prostate-Specific Antigen) to predict survival outcomes



	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
BLCA	Patient Primary Tumor Site_Wall Posterior	Race Category_WHITE	Patient_Weight	Positive Finding Lymph Node Hematoxylin and Eosin Staining Microscopy Count	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code_N1
KIRC	American Joint Committee on Cancer Metastasis Stage Code_M1	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code_N1	Specimen Second Longest Dimension	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IV	Sex_Male
UCEC	Ethnicity Category_NOT HISPANIC OR LATINO	Neoplasm American Joint Committee on Cancer Clinical Group Stage_Stage IIB	Race Category_ASIAN	Lymph Nodes Aortic Pos Total	Neoplasm American Joint Committee on Cancer Clinical Group Stage Stage II
PAAD	International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code 8426/3	American Joint Committee on Cancer Metastasis Stage Code_T2	International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code 8140/3	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code N1	Neoplasm Event Post Initial Therapy Indicator_YES
BRCA	Staging System_No Axillary Staging	Menopause Status_Post (prior bilateral ovariectomy OR >12 mo since LMP with no prior hysterectomy)	American Joint Committee on Cancer Metastasis Stage Code_MX	Race Category_WHITE	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IA
LUSC	Surgical Margin Resection Status_R1	American Joint Committee on Cancer Tumor Stage Code T2b	Patient Primary Tumor Site_R-Middle	Ethnicity Category_NOT HISPANIC OR LATINO	American Joint Committee on Cancer Metastasis Stage Code MX
LIHC	American Joint Committee on Cancer Tumor Stage Code T3A	Laboratory procedure albumin result upper limit of normal value	New Neoplasm Event Post Initial Therapy Indicator YES	Specimen collection method name Lobectomy	Sex_Male
THCA	Race Category_WHITE	Race Category_BLACK OR AFRICAN AMERICAN	Neoplasm American Joint Committee on Cancer Code N1	American Joint Committee on Cancer Tumor Stage Code T1a	American Joint Committee on Cancer Tumor Stage Code T2
COAD	Neoplasm Disease Stage American Joint Committee on Cancer Code Stage II	American Joint Committee on Cancer Tumor Stage Code T4a	Neoplasm Disease Stage American Joint Committee on Cancer Code Stage IIIA	Lymphovascular Invasion Indicator YES	American Joint Committee on Cancer Tumor Stage Code T3
SKCM	American Joint Committee on Cancer Tumor Stage Code_T4b	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IV	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code N3	American Joint Committee on Cancer Tumor Stage Code_T3b	Breslow Depth
GBM	Karnofsky Performance Score	Neoadjuvant Therapy Type Administered Prior To Resection Text Yes	Race Category_BLACK OR AFRICAN AMERICAN		
HSNC	International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code 8072/3	International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code 8071/3	Primary Lymph Node Presentation Assessment Ind-3_YES	Race Category_WHITE	Neoplasm Disease Lymph Node Stage American Joint Committee on Cancer Code N1
STAD	Surgical Margin Resection_R2	Ethnicity Category_NOT HISPANIC OR LATINO	Surgical Margin Resection Status_R2	Patient Primary Tumor Site_Stomach (NOS)	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IA
LUAD	American Joint Committee on Cancer Metastasis Stage Code_T2	Neoplasm Disease Stage American Joint Committee on Cancer Code_Stage IB	Prior Diagnosis_Yes	American Joint Committee on Cancer Tumor Stage Code_T1A	Neoplasm Disease Stage American Joint Committee on Cancer Code_N1
PRAD	CT Scan ab pelvis indicator_YES	PSA most recent results	Neoplasm American Joint Committee on Cancer Clinical Primary T Stage T2C	Gleason Pattern Primary	Sample Type_Primary
OV	Neoplasm Histologic Grade_G2	Neoplasm Histologic Grade_G3	Diagnosis Age	Shortest Dimension	Neoplasm Histologic Grade_GX

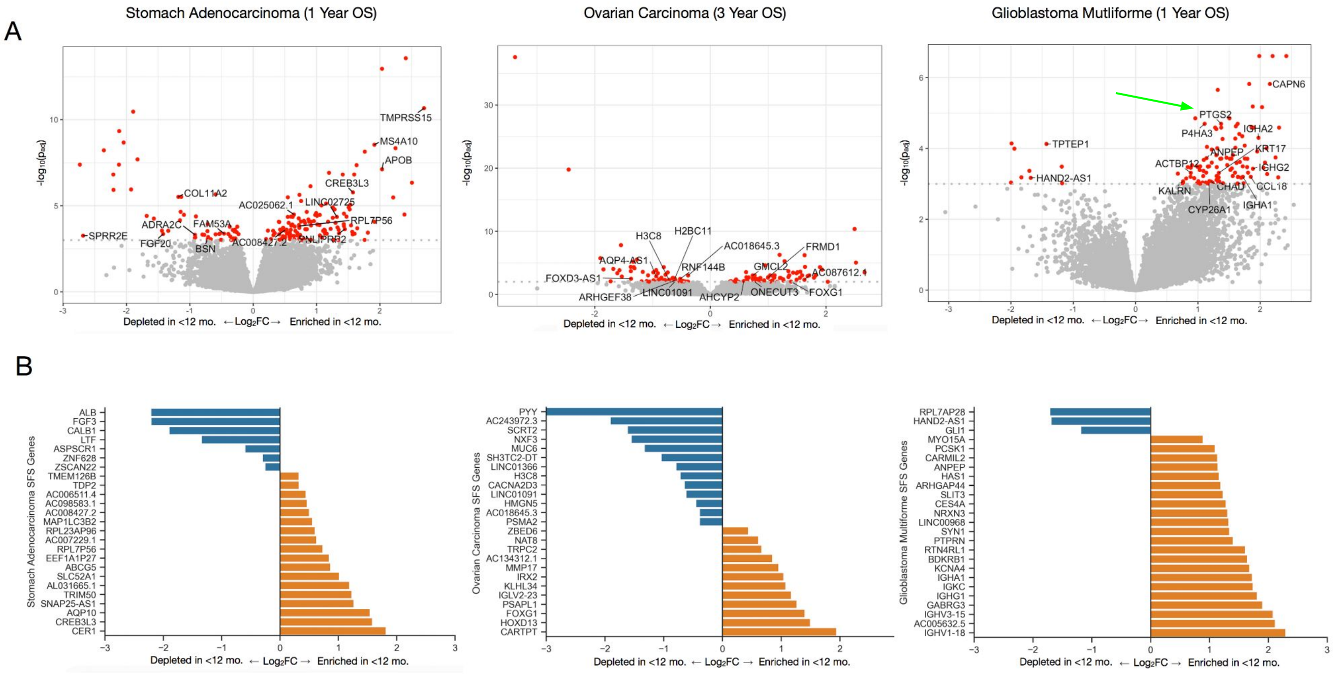
Top ranked features for 3 year model. Abbreviations: BLCA = bladder urothelial carcinoma, KIRC = kidney clear cell carcinoma, UCEC = uterine corpus endometrial carcinoma, PAAD = pancreatic adenocarcinoma, BRCA = breast invasive carcinoma, LUSC = lung squamous cell carcinoma, LIHC = liver hepatocellular carcinoma, THCA = thyroid carcinoma, COAD = colon adenocarcinoma, SKCM = skin cutaneous melanoma, GBM = glioblastoma multiforme, HSNC = head and neck squamous cell carcinoma, STAD = stomach adenocarcinoma, LUAD = lung adenocarcinoma, PRAD = prostate adenocarcinoma, OV = ovarian serous cystadenocarcinoma.


Common Procedure | PSA Blood Test
Prostate Gland
releases PSA molecules
Bladder
1 A PSA test measures the amount of PSA molecules in a patient's blood
PSA LEVEL
LOW
HIGH
2 It's normal for all men to have some PSA in their blood
3 A high level of PSA can be a sign of prostate cancer, but...
4 It can also be raised for lots of other reasons including...
A urinary or urinary tract infection (UTI)
Prostate stimulation
Vigorous exercise
Certain medications
MedicineNet © 2019

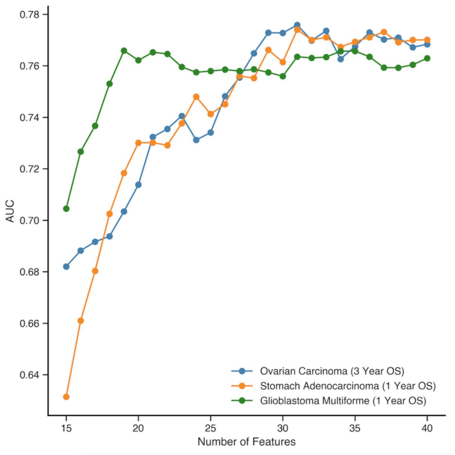

Gleason's Pattern
1. Small, uniform glands
2. More stroma between glands
3. Distinctly infiltrative margins
4. Irregular masses of neoplastic glands
5. Only occasional gland formation
Well differentiated
Moderately differentiated
Poorly differentiated /Anaplastic

Results - Differential Expression Gene Analyses

- Certain genes in our DE analysis represented **known markers** to promote or impede cancer
    - PTSG2 (prostaglandin-endoperoxide synthase 2) is significantly enriched in the <12 month survival cohort in Glioblastoma Multiforme
    - PTSG2 is *also* reported to aggressively facilitate resistance of glioblastoma to chemotherapy treatment methods

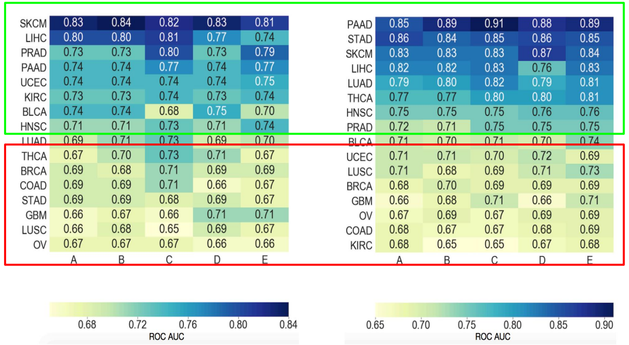


Discussion & Future Work



- Clinical and pathological data alone can accurately predict 1 and 3 year overall survival in many cancers, but the addition of gene expression features significantly improves survival prediction performance in weaker cancers
    - For example, STAD, GBM and OV saw up to **+0.15 increase in AUC**

- Poorly performing cancers (e.g. OV) often suffered from a lack of **disease-specific features/markers** that better-performing cancers had (e.g. Liver fibrosis for LIHC/Liver Hepatocellular Carcinoma), and benefited greatly from additional pathological or gene expression data
    - The AUC for OV increased by 23% increase after adding the 25 additional genes on top of the initial 15 clinicopathological features



	A	B	C	D	E
SKCM	0.83	0.84	0.82	0.83	0.81
LIHC	0.80	0.80	0.81	0.77	0.74
PRAD	0.73	0.73	0.80	0.73	0.79
PAAD	0.74	0.74	0.77	0.74	0.77
UCEC	0.74	0.74	0.74	0.74	0.75
KIRC	0.73	0.73	0.74	0.73	0.74
BLCA	0.74	0.74	0.68	0.75	0.70
HNSC	0.71	0.71	0.73	0.71	0.74
LUAD	0.69	0.71	0.73	0.69	0.70
THCA	0.67	0.70	0.73	0.71	0.67
BRCA	0.69	0.68	0.71	0.69	0.69
COAD	0.69	0.69	0.71	0.66	0.67
STAD	0.69	0.69	0.68	0.69	0.67
GBM	0.66	0.67	0.66	0.71	0.71
LUSC	0.66	0.68	0.65	0.69	0.67
OV	0.67	0.67	0.67	0.66	0.66

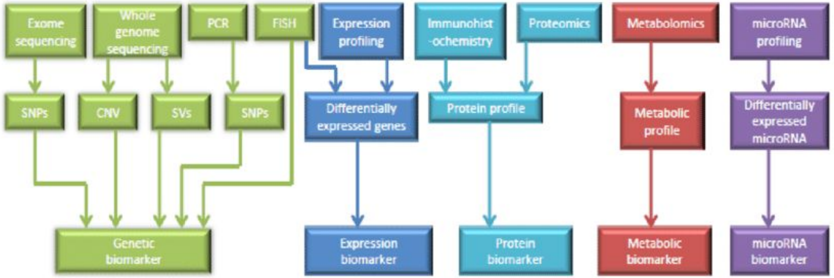
	A	B	C	D	E
PAAD	0.85	0.89	0.91	0.88	0.89
STAD	0.86	0.84	0.85	0.86	0.85
SKCM	0.83	0.83	0.83	0.87	0.84
LIHC	0.82	0.82	0.83	0.76	0.83
LUAD	0.79	0.80	0.82	0.79	0.81
THCA	0.77	0.77	0.80	0.80	0.81
HNSC	0.75	0.75	0.75	0.76	0.76
PRAD	0.72	0.71	0.75	0.75	0.75
BLCA	0.71	0.70	0.71	0.70	0.74
UCEC	0.71	0.71	0.70	0.72	0.69
LUSC	0.71	0.68	0.69	0.71	0.73
BRCA	0.68	0.70	0.69	0.69	0.69
GBM	0.66	0.68	0.71	0.66	0.71
OV	0.67	0.69	0.67	0.69	0.69
COAD	0.68	0.67	0.67	0.68	0.69
KIRC	0.68	0.65	0.65	0.67	0.68

Discussion & Future Work

- Many factors other than clinical/gene expression data (DNA methylation, copy number, spatial information of biomarkers) can influence survival outcomes

- Develop *specialized* models for each cancer subtype that adaptively select features relevant to each specific cancer type

- Extract features directly from TCGA tumor imagery



Thank You!

Questions?

My e-mail: **pzy2@illinois.edu**

Special thanks to my mentor, **Jimmy A. Guo** from the Broad Institute, for guiding and supporting me throughout the project.