Cancer Phenotyping

CMU 10-742: Machine Learning in Healthcare

24 October 2024

Harry Hochheiser
harryh@pitt.edu
Biomedical Informatics
Member, UPMC Cancer Center, Cancer Epidemiology and Prevention Program

Back to the future: IBM Watson

2013

The future of cancer treatment and research: What IBM Watson means for our patients

“The Oncology Expert Adviser can extract patient information from various data sources and synthesize all available medical records into a clear, concise and accurate synopsis. It can analyze clinical information, medical history, as well as leukemia-related information, such as specific genetic and molecular features, and look at all available information in the context of published evidence-based guidelines and available clinical trials.

The OEA also allows us to look at changes in a patient's condition over time, enabling us to learn a wealth of information within seconds and answer complex medical questions with speed, accuracy and confidence. This has an enormous potential to help us make the best cancer treatment decisions for our patients and the best research decisions that can help us make progress in our fight against cancer.”

https://www.mdanderson.org/cancerwise/what-ibm-watson-means-for-our-patients.h00-158834379.html

Watson

2017

MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine

Forbes, Feb 19, 2017

https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/

What is a cancer phenotype?

“The Oncology Expert Adviser can extract patient information from various data sources and synthesize all available medical records into a clear, concise and accurate synopsis. It can analyze clinical information, medical history, as well as leukemia-related information, such as specific genetic and molecular features, and look at all available information in the context of published evidence-based guidelines and available clinical trials.

The OEA also allows us to look at changes in a patient's condition over time, enabling us to learn a wealth of information within seconds and answer complex medical questions with speed, accuracy and confidence. This has an enormous potential to help us make the best cancer treatment decisions for our patients and the best research decisions that can help us make progress in our fight against cancer.”

https://www.mdanderson.org/cancerwise/what-ibm-watson-means-for-our-patients.h00-158834379.html

What is cancer?

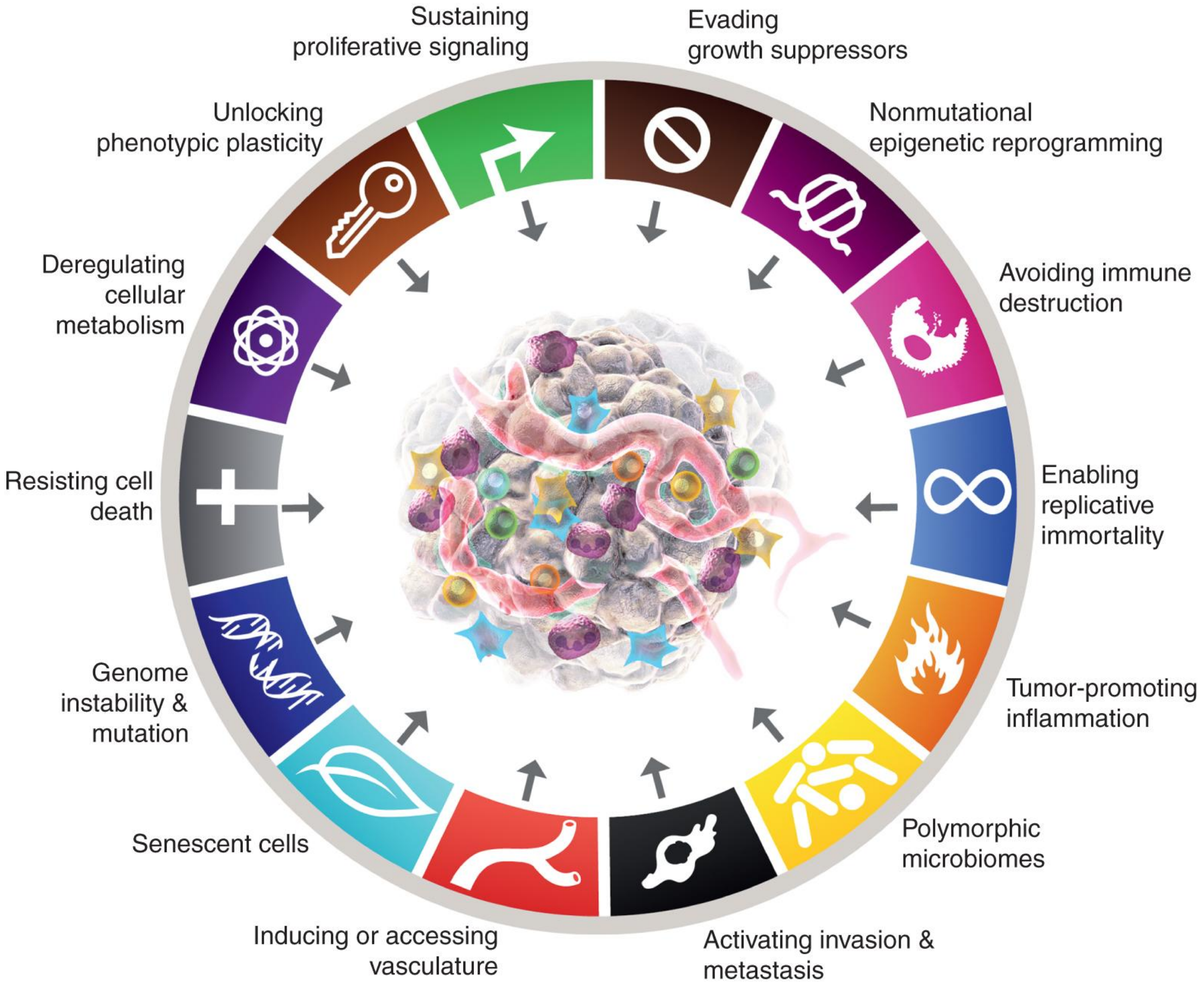
Brown, et al. 2023 “Updating the definition of cancer” *Molecular Cancer Research*
https://doi.org/10.1158/1541-7786.MCR-23-0411

“Cancer is a disease of uncontrolled proliferation by transformed cells subject to evolution by natural selection.”

Disease:	Illness/abnormality that disrupts bodily function
Uncontrolled Cell Proliferation:	Replicative immortality
Transformed Cells:	Genetic/epigenetic mutations causing malignancy
Evolution:	Genetic modification
Natural Selection:	Heritable evolution, competition, struggle.. leading to resistance/adaptations

Hallmarks of Cancer

Hanahan, 2022 *Cancer Discovery* https://doi.org/10.1158/2159-8290.CD-21-1059



Not one disease, but…

SNOMED-CT Terms..

Malignant tumor of breast

- Carcinoma of breast
- Familial cancer of breast
- Hormone receptor positive malignant neoplasm of breast
- Infiltrating lobular carcinoma of breast
- Local recurrence of malignant tumor of breast
- Locally advanced breast cancer
- Malignant lymphoma of breast
- Malignant melanoma of breast
- Malignant neoplasm of axillary tail of breast
- Malignant neoplasm of bone, connective tissue, skin and breast
- Malignant neoplasm of breast in remission
- Malignant neoplasm of breast lower inner quadrant
- Malignant neoplasm of breast lower outer quadrant
- Malignant neoplasm of breast upper inner quadrant
- Malignant neoplasm of breast upper outer quadrant
- Malignant neoplasm of female breast
- Malignant neoplasm of male breast
- Malignant neoplasm of overlapping sites of breast
- Malignant phyllodes tumor of breast
- Metastatic malignant neoplasm to breast
- Paget's disease of nipple
- Primary malignant neoplasm of breast
- Sarcoma of breast

Malignant tumor of lung

- Acinar cell carcinoma of lung
- Acinar cell cystadenocarcinoma of lung
- Alpha heavy chain disease, respiratory form
- Carcinoma of lung
- Carcinosarcoma of lung
- Epithelioid hemangioendothelioma of lung
- Kaposi's sarcoma of lung
- Local recurrence of malignant tumor of lung
- Malignant carcinoid tumor of left lung
- Malignant carcinoid tumor of lung
- Malignant carcinoid tumor of right lung
- Malignant neoplasm of lower lobe of lung
- Malignant neoplasm of middle lobe of right lung
- Malignant neoplasm of upper lobe of lung
- Malignant neoplasm of upper lobe, bronchus or lung
- Malignant tumor of lung parenchyma
- Metastatic malignant neoplasm to lung
- Non-Hodgkin's lymphoma of lung
- Overlapping malignant neoplasm of bronchus and lung
- Pleuropulmonary blastoma
- Primary malignant neoplasm of lung
- Primary pulmonary lymphoma
- Pulmonary blastoma
- Signet ring cell carcinoma of lung

What is a phenotype?

https://www.genome.gov/genetics-glossary/Phenotype

“Phenotype refers to an individual’s observable traits, such as height, eye color and blood type.

A person’s phenotype is determined by both their genomic makeup (genotype) and environmental factors.”

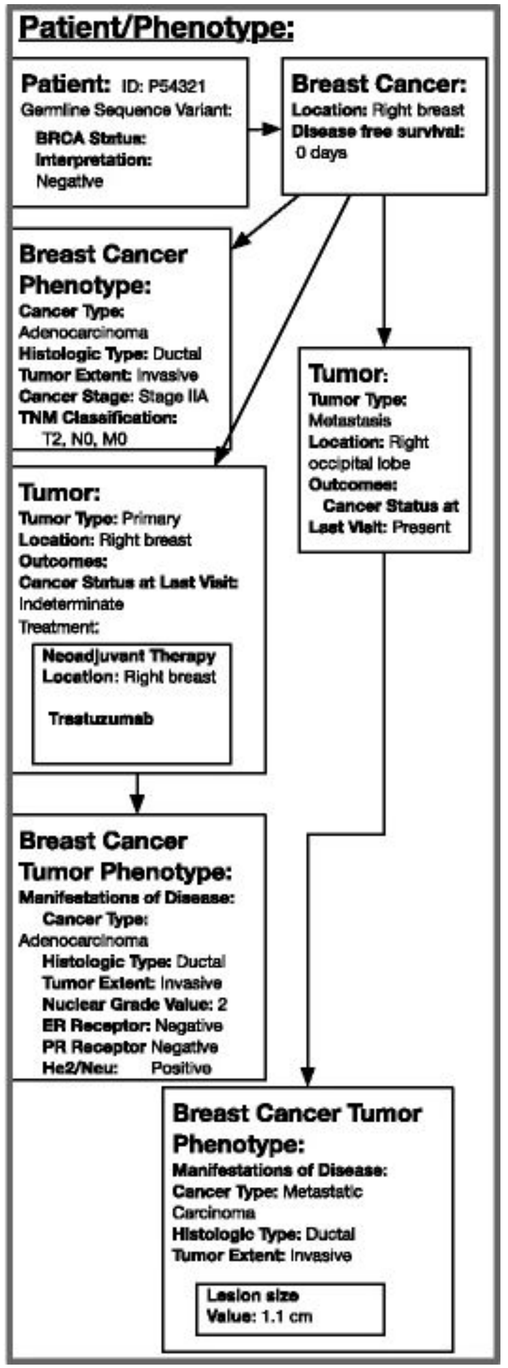
What can be observed, and how? -> data used to assess a phenotype

Cancer phenotypes are complex

Hochheiser, et al. 2016
https://doi.org/10.1186/s12911-016-0358-4c

Other factors:

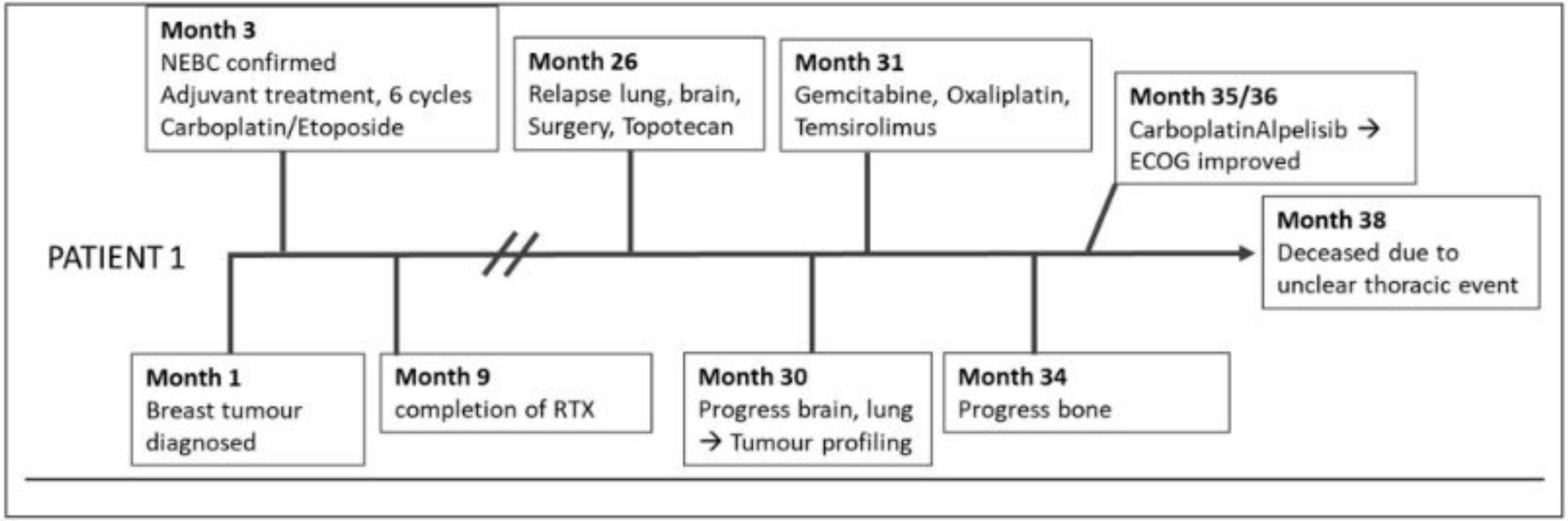
- Comorbidities?
- Response to treatment?



Cancer Phenotypes change over time

Tumors evolve - new mutations

Metastases



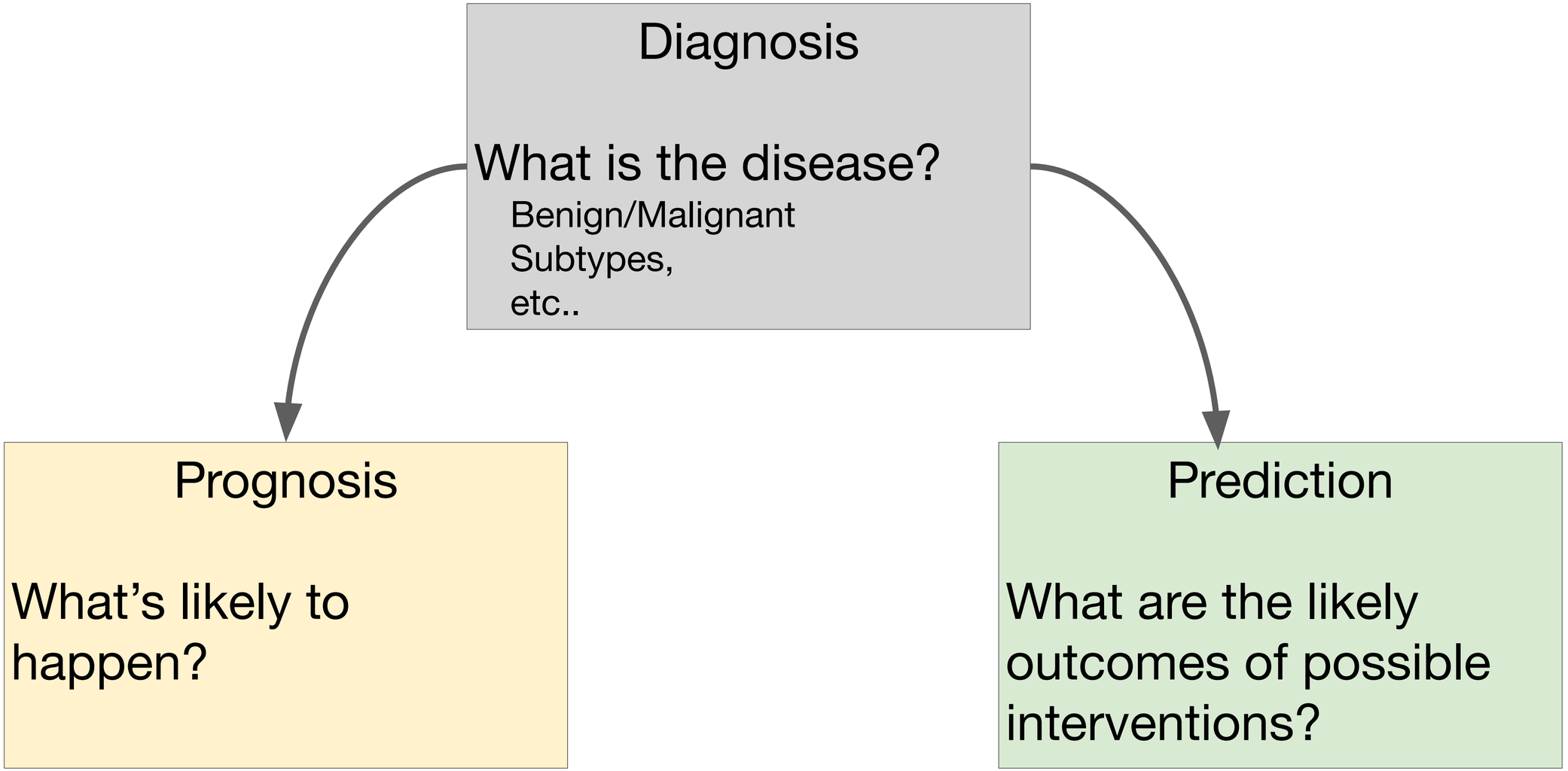
Schaffrin-Nabe, et al. 2022 *Frontiers in Medicine* https://doi.org/10.3389/fmed.2022.841441

Individual phenotypes

People with cancer change

- Reactions to treatment
  - poor tolerance
- Comorbidities
- Preferences

Types of phenotypes



Prognosis vs. Prediction

Echle, et al. 2022 British Journal of Cancer https://doi.org/10.1038/s41416-020-01122-x

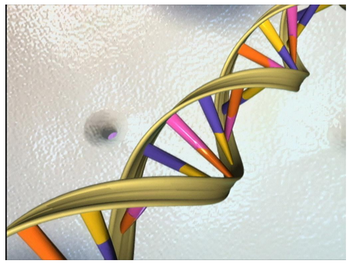
- Microsatellite instability (MSI - characterization of genomic behavior)
-
- Prognostic:
  - Stage II Colorectal Cancer
  - MSI -> better likely outcome
  - lower intensity of adjuvant chemotherapy

- Predictive:
  - Treatment-refractory Stage IV
  - MSI approved biomarker for immune-checkpoint-inhibitor based immunotherapy
  - MSI is a predictive of positive response to immunotherapy

Cancer Phenotypes at multiple scales

Omics

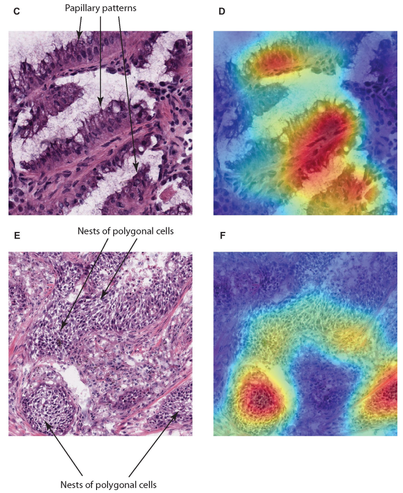
DNA/genome
epigenetics,
transcriptome
pathways
microbiome



www.genome.gov

Tumor/
Cellular

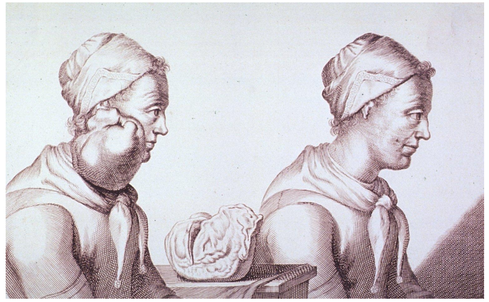
Imaging
Tumor microenvironment



Yu et al. 2020 DOI: 10.1093/jamia/ocz230

Patient/
Clinical

Response to Treatment
Comorbidities
Social Determinants
Needs/Preferences



https://en.wikipedia.org/wiki/History_of_cancer#/media/File:Clara_Jacobi-Tumor.jpg

Basic Research

Clinical

Data used in Cancer Phenotyping

- Genomic/Epigenetic
  - Somatic
  - Tumor
  - Methylation
- Transcriptomic
- Proteomic
- Imaging
  - Radiology
  - Histology
- Clinical
  - Symptoms
  - Procedures
  - Medication
  - Patient reported outcomes
  - Social Determinants of Health
- Exposures

Questions to ask as we discuss specific applications?

- What data is involved?

- What biases might be present?
  - Who is represented or not?

- What outcomes are we looking for?

‘Omics phenotypes

Changes in …

… gene expression

… protein abundance

… microbiome

… metabolome

… immune behavior

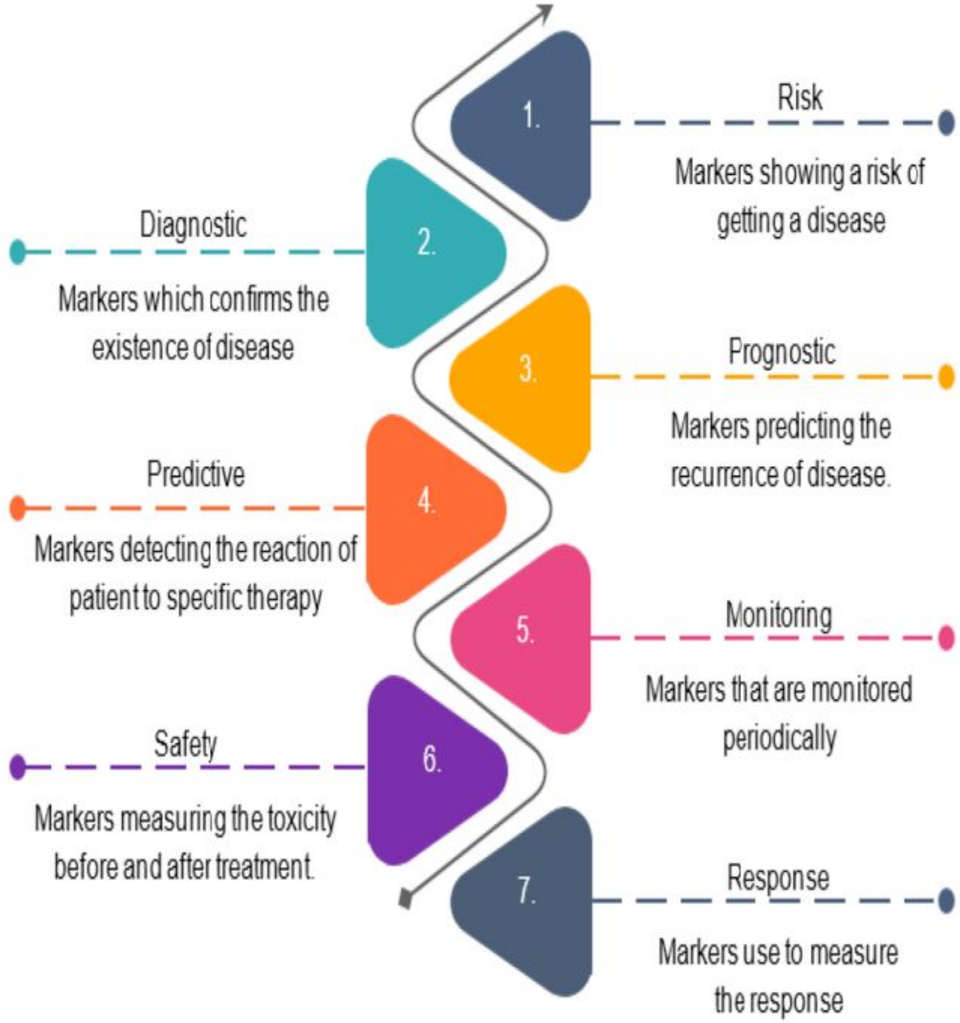
associated with cancer.

Biomarker searches

‘Omics analysis for finding genes/RNA/proteins, etc. informative relative to cancer

Categories are often blurred.

Dhillon, et al. 2022
https://doi.org/10.1007/s11831-022-09821-9



Cancer ‘omics challenges

Swanson, et al. 2023 https://dx.doi.org/10.1016/j.cell.2023.01.035

- Sample Size (maybe < 50 /subtype)

- High-dimensionality
  - regularization becomes important
- Low signal/noise

- Case/control
- n-of-1

Diagnosis

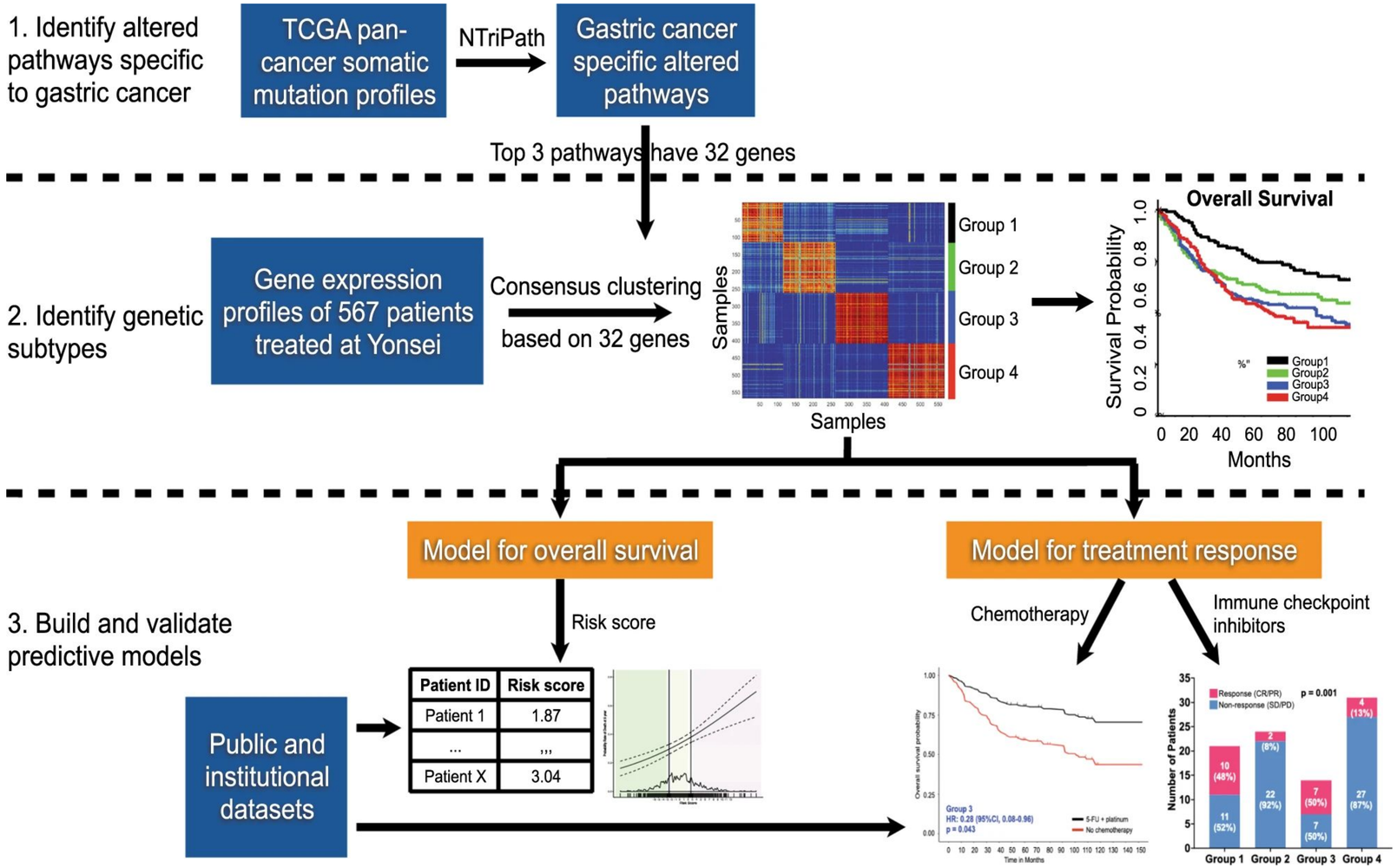
- Challenge - develop finer-grain understanding of patients
  - .. often when we don’t know what the classifications are

- Unsupervised learning
  - sparse data (few samples, many dimensions)

  - All items may be similar distant…

Development and validation of a prognostic and predictive 32-gene signature for gastric cancer

Cheong, et al. 2022 https://doi.org/10.1038/s41467-022-28437-y

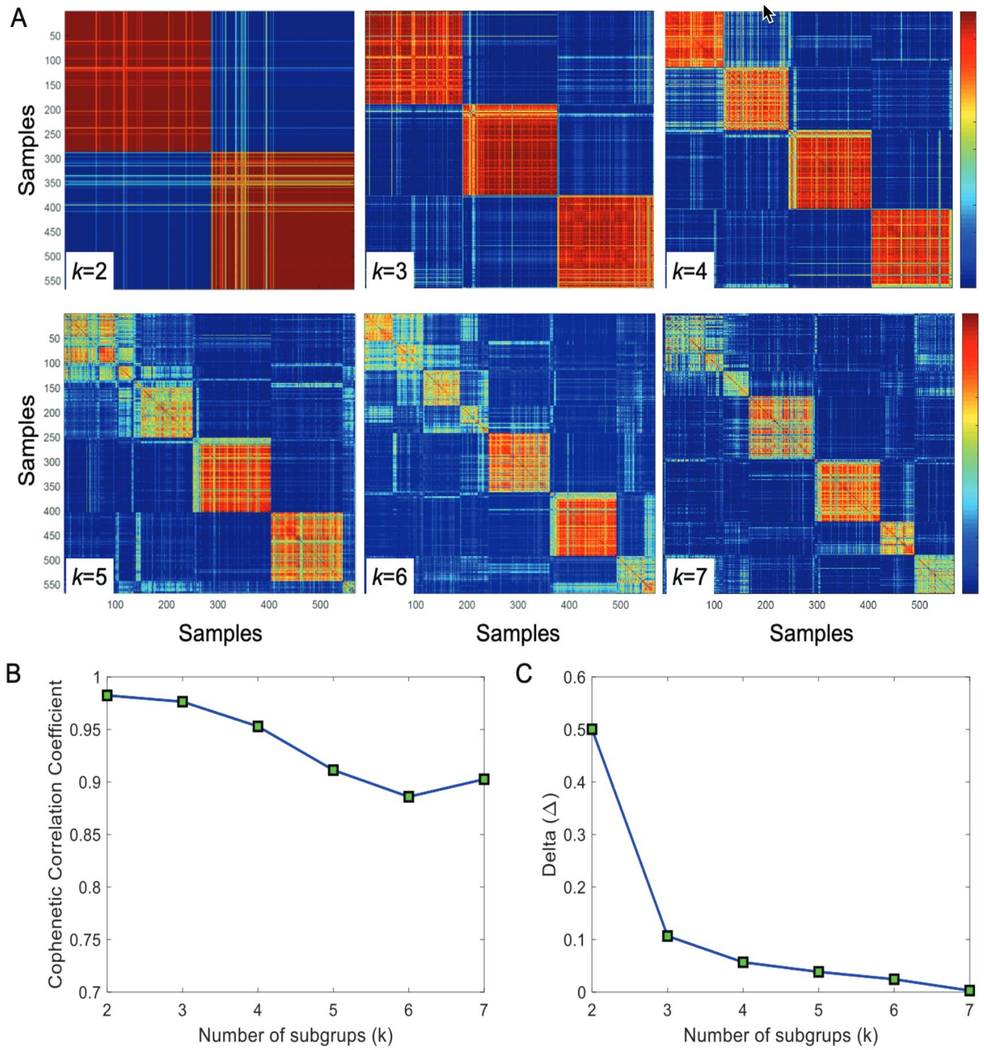


gene expression data
published cohorts
TCGA

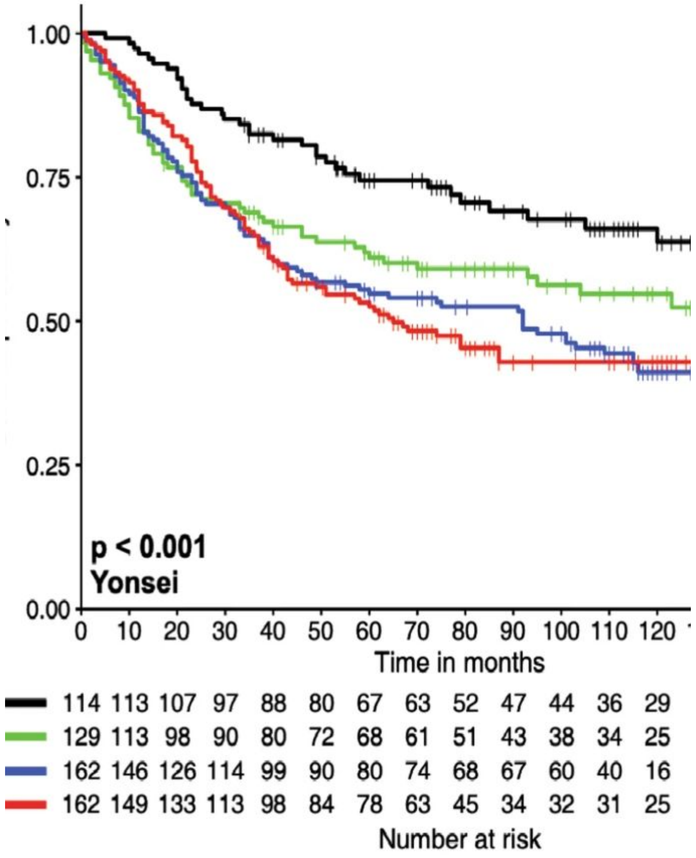
Development and validation of a prognostic and predictive 32-gene signature for gastric cancer

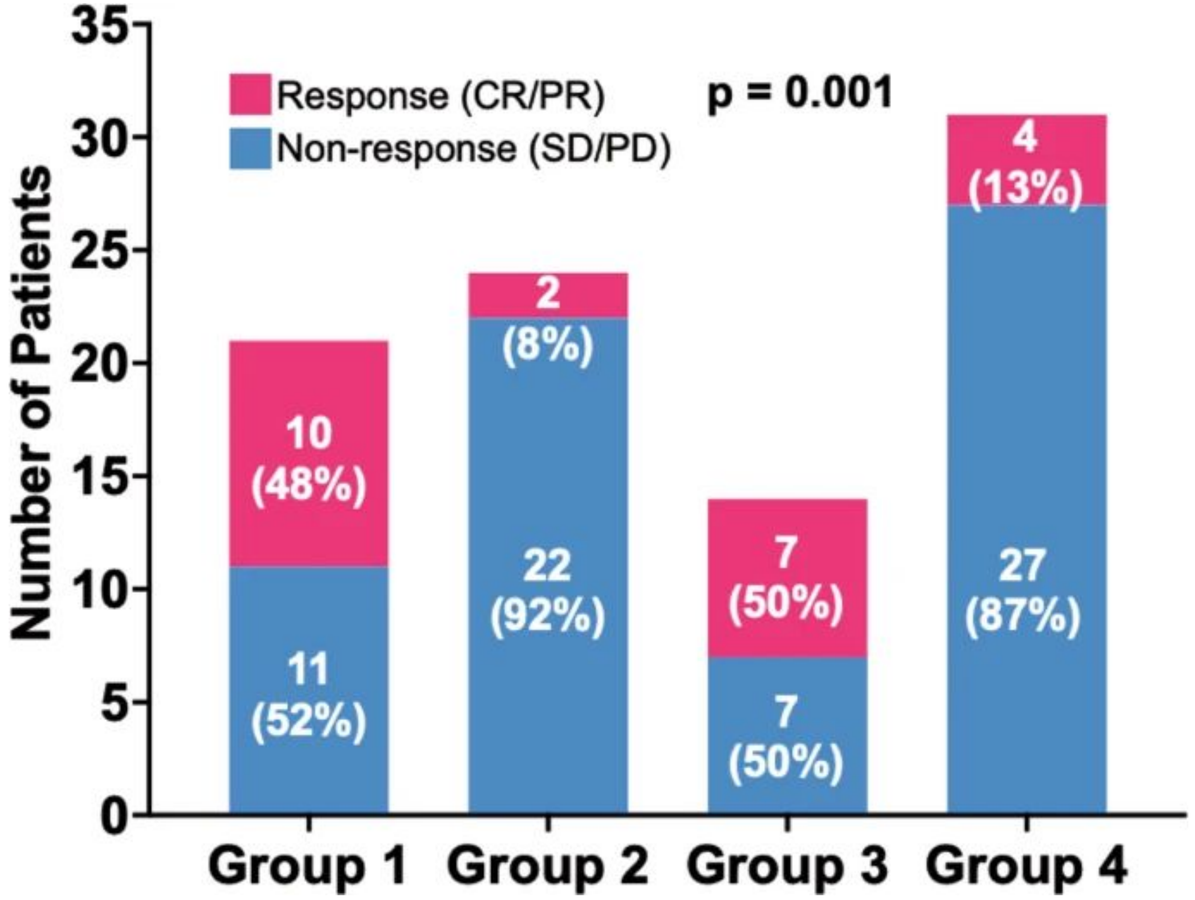
Clustering - non-negative matrix factorization,

validated through a multi-class SVM classifier (AUC 0.98)



Clusters associated with better outcomes and responsiveness to immune checkpoint inhibitors





complete response, partial response, stable disease, progressive disease

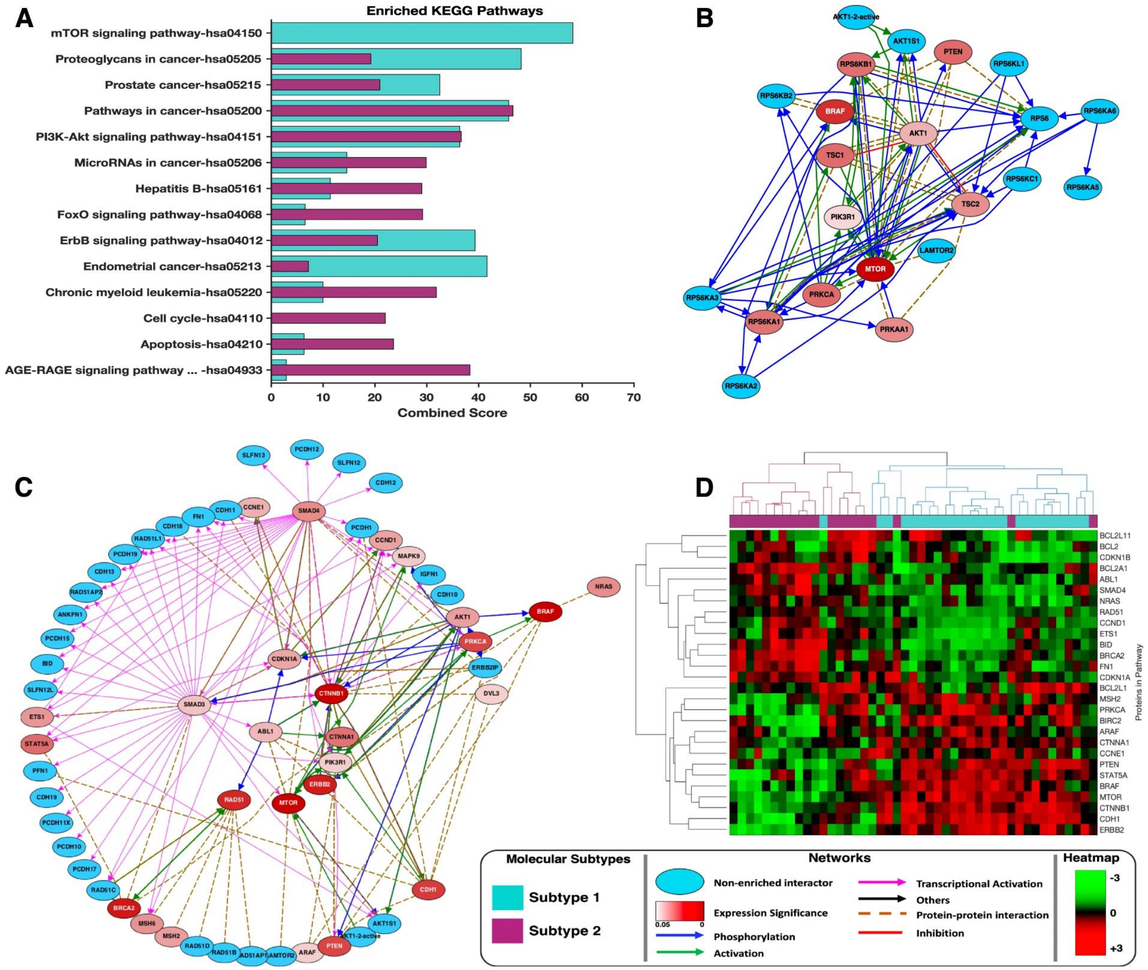
Machine Learning and Network Analysies Reveal Disease Subtypes of Pancreatic Cancer and their Moledcular Characteristics

Sinkala, et al. 2020
https://doi.org/10.1038/s41598-020-58290-2

Data - proteomics

Methods - K-means clustering

Analysis - pathway enrichment, examination of key proteins (mTOR/SMAD4)



A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns

Jiao, et al. 2020 https://doi.org/10.1038/s41467-019-13825-8

Challenge - 3% of patients present with metastasis and no obvious primary

Multi-class deep learning classifier trained on 2606 tumors, 24 cancer types

Features

- 150 mutational features for single nucleotide variation (type, flanking, etc.)
- Structural variations - copy number, indel,
- Counts of events/genomome, etc.
- total 2897 SNV +indel, 2926 CNV, 2929 SV features

Fully-connected Neural Net - softmax probability for mutli-class prediction

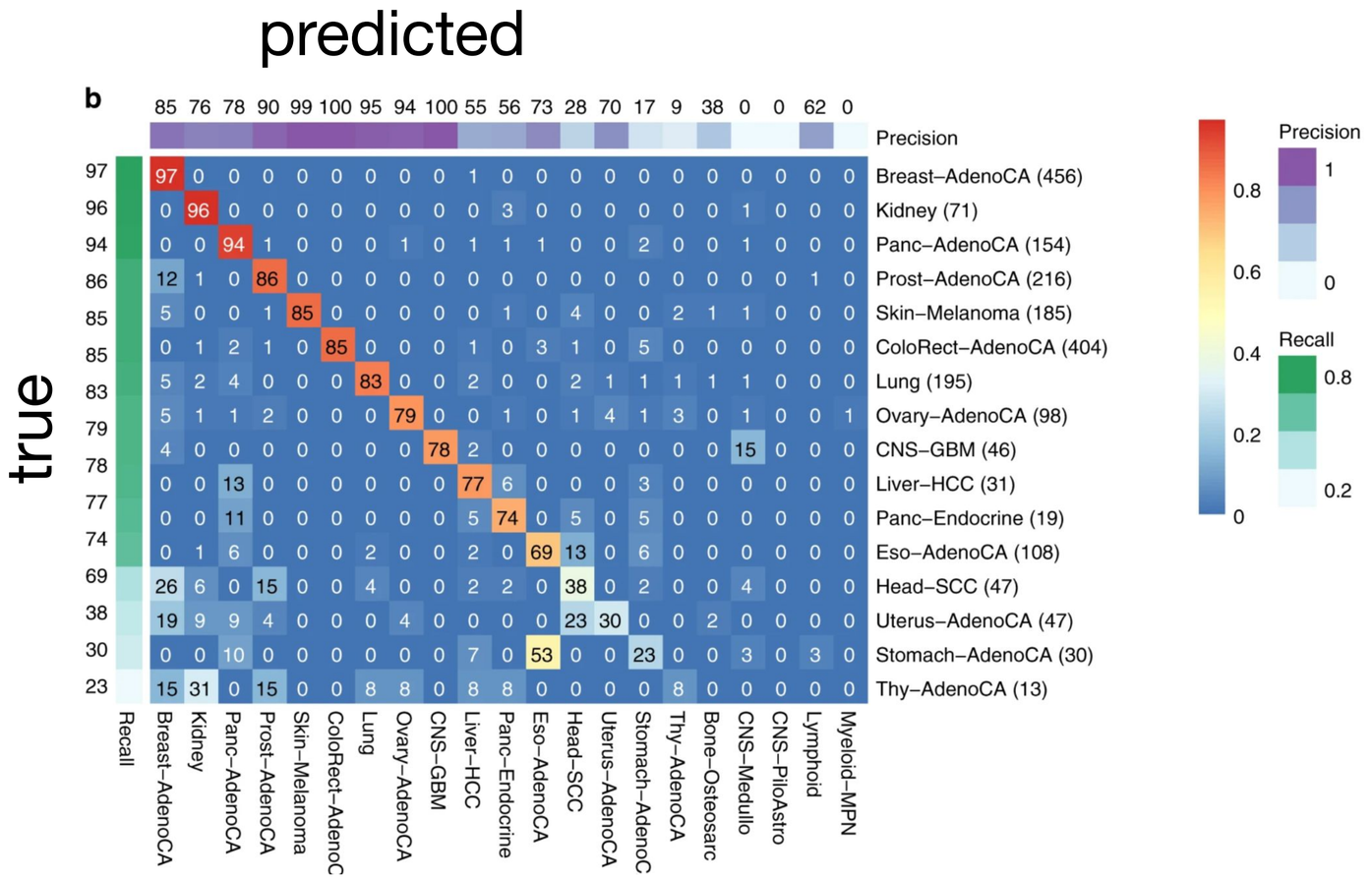
- 88%/83% accuracy on independent primary/metastatic samples - 2x trained pathologists

Addition of drivers did not improve performance

A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns

Jiao, et al. 2020 https://doi.org/10.1038/s41467-019-13825-8

Validation against independent data sets - primary and metastatic



Tumor-Specific Causal Inference

Cai, et al. 2019 https://doi.org/10.1371%2Fjournal.pcbi.1007088

Searching for genetic mutations that cause cancer - *cancer drivers.*

Tumor-specific causal association between

Somatic Genome Alterations (SGAs) and

Differentially Expressed Genes (DEGs)

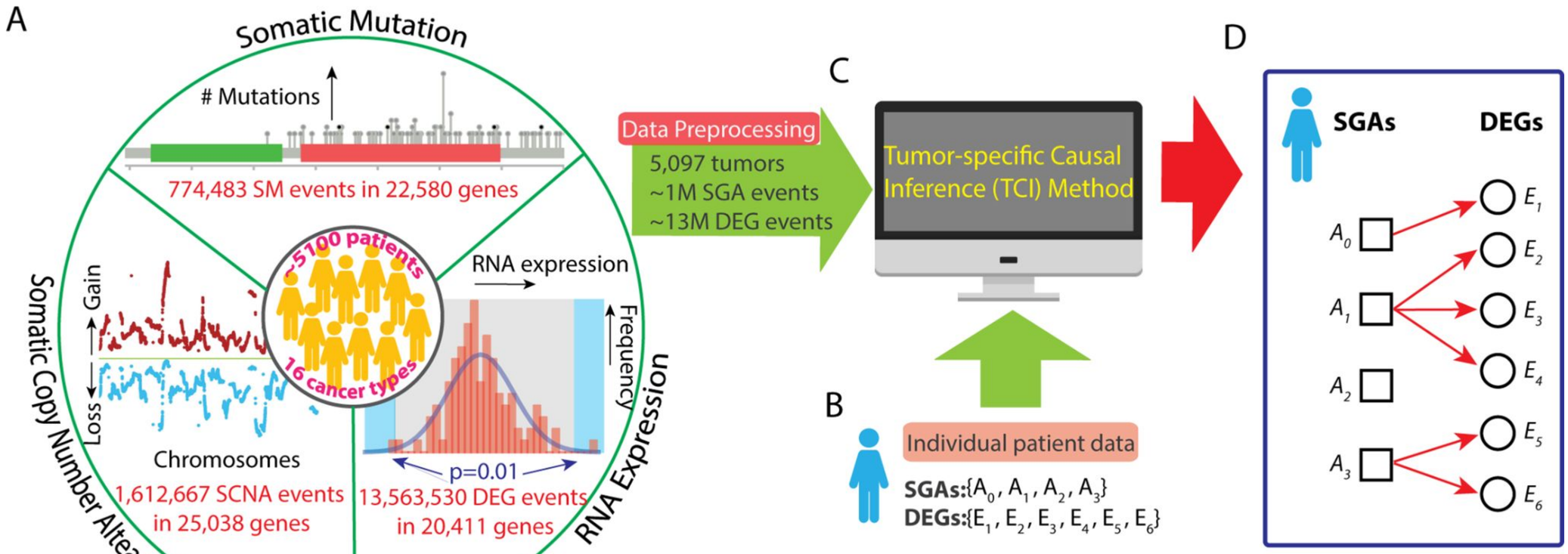
Calculate most-probable SGA for each DEG

Data sources: The Cancer Genome Atlas (TCGA),

Library of Integrated Network-Based Cellular Signatures (LINCS)

Tumor-Specific Causal Inference

Cai, et al. 2019 https://doi.org/10.1371%2Fjournal.pcbi.1007088



Tumor-Specific Causal Inference

Cai, et al. 2019 https://doi.org/10.1371%2Fjournal.pcbi.1007088

$$P\left(A_h \rightarrow E_i \middle| D\right) = \frac{1}{Z} P\left(A_h \rightarrow E_i\right) P\left(D \middle| A_h \rightarrow E_i\right)$$

A_h - potentially causal SGA
E_i - differentially expressed gene
D - genomic dataset

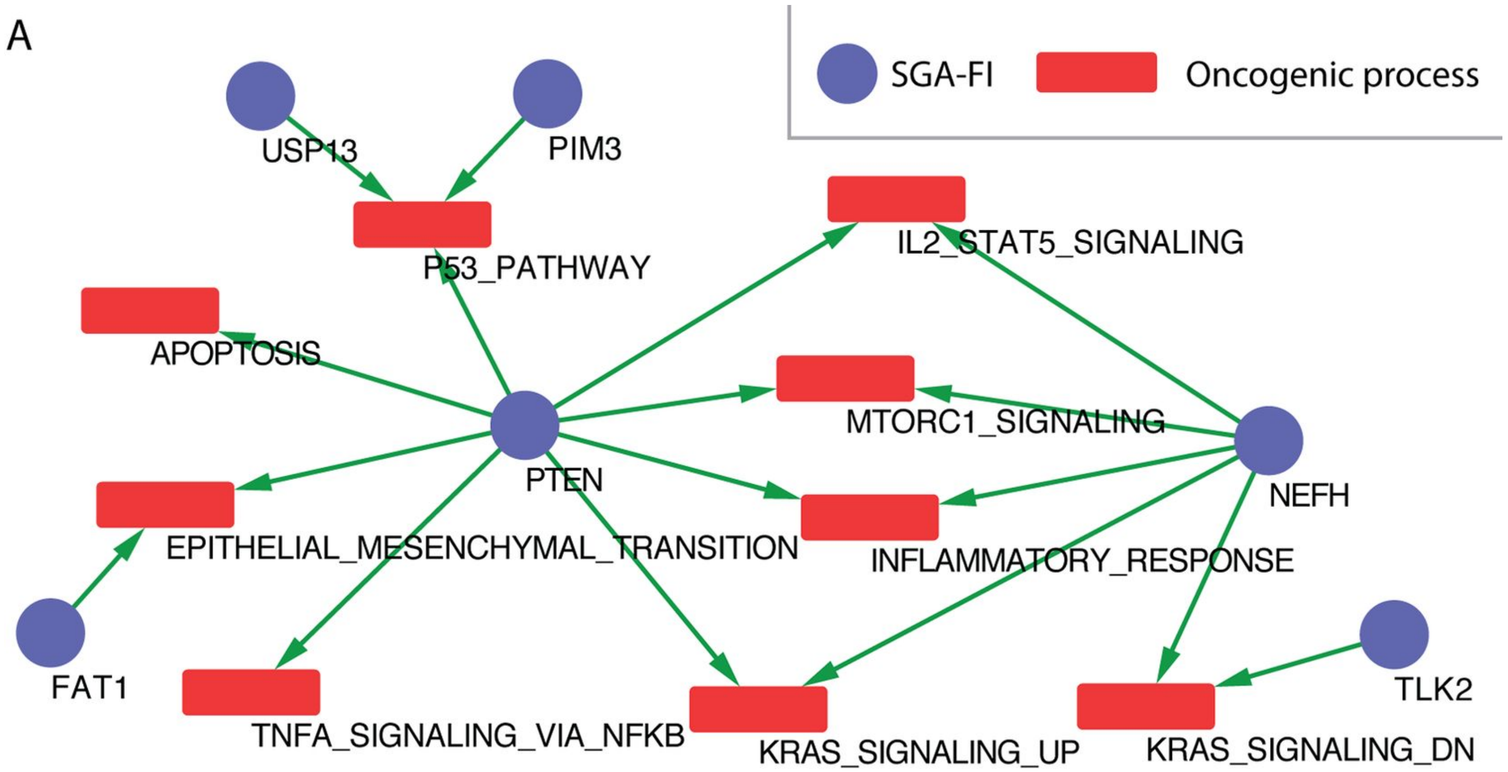
population prior

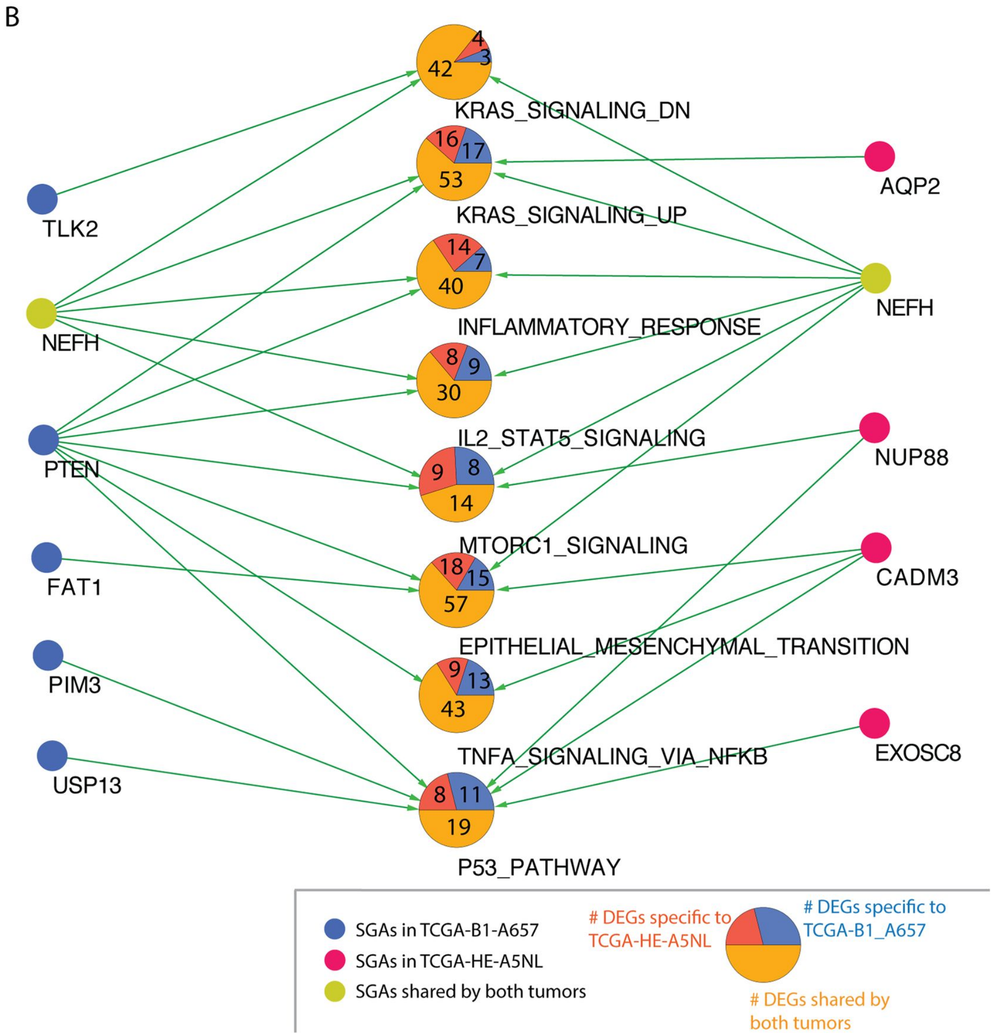
marginal likelihood of data - tumor specific

extension by Xue et al - 2019 clusters DEGs to infer disease mechanisms https://doi.org/10.1038%2Fs41598-019-48318-7

Tumor-specific causal inference -validation

Cai, et al. 2019 https://doi.org/10.1371%2Fjournal.pcbi.1007088

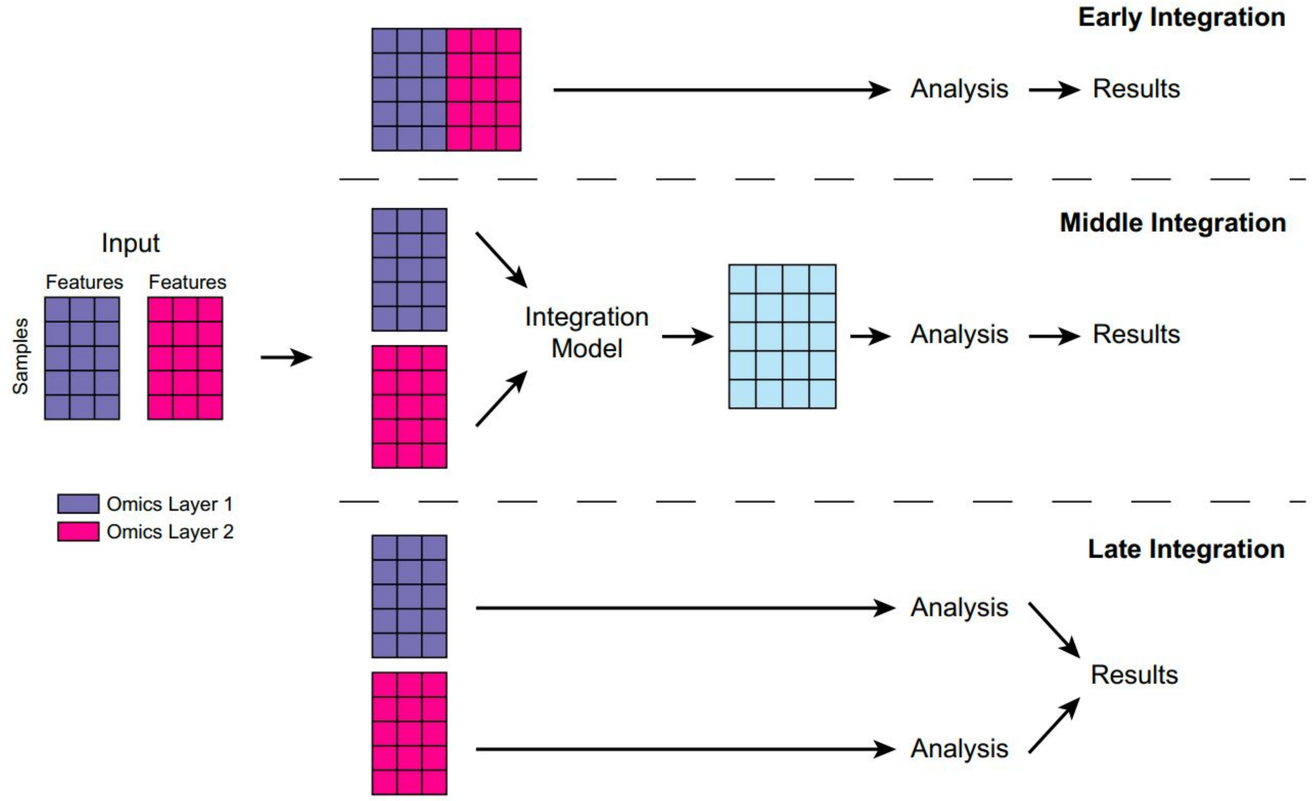




SGA-FI: somatic gene alterations with functional impact

Other ‘Omics targets

Multiomics - Cai 2022 https://doi.org/10.1016/j.isci.2022.103798



Metastasis

Recurrence

Liquid Biopsy (diagnosis from cell-free DNA)

Single-cell/cell-type specific

Understanding the tumor microenvironment

Epigenetics

An interpretable deep learning framework for genome-informed precision oncology

Ren, et al. 2024 https://doi.org/10.1038/s42256-024-00866-y

Use somatic alterations to predict drug response

Transformers for reduced-dimensionality representation

Cell lines

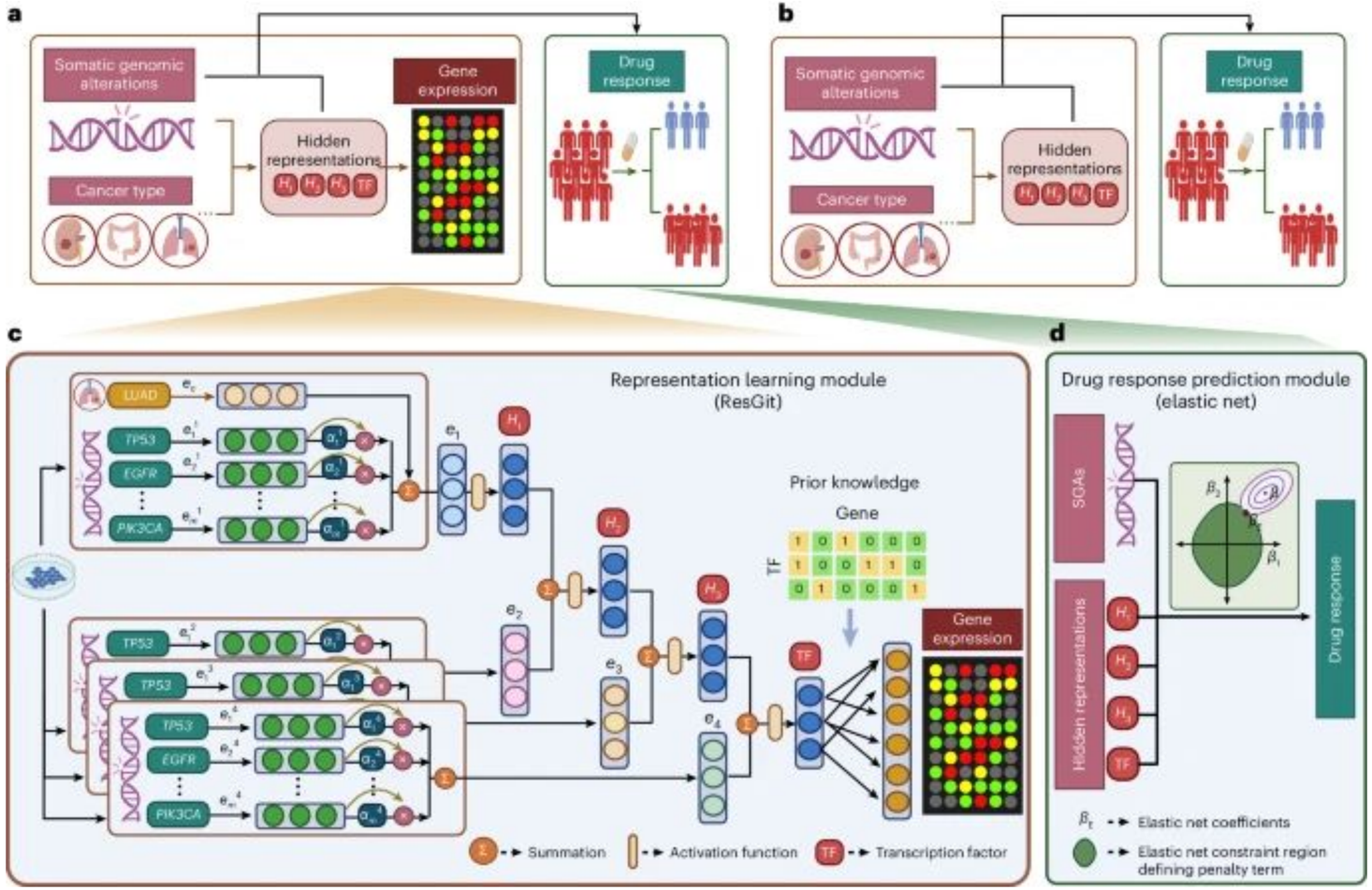
TCGA data

transcription-factor target gene matrix.

Gene2Vec: word2vec embeddings of SGAs

An interpretable deep learning framework for genome-informed precision oncology

Ren, et al. 2024 https://doi.org/10.1038/s42256-024-00866-y



Imaging https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis

Histology - stained cellular images from tissue

CT

MRI

Nuclear Scan

Bone Scan

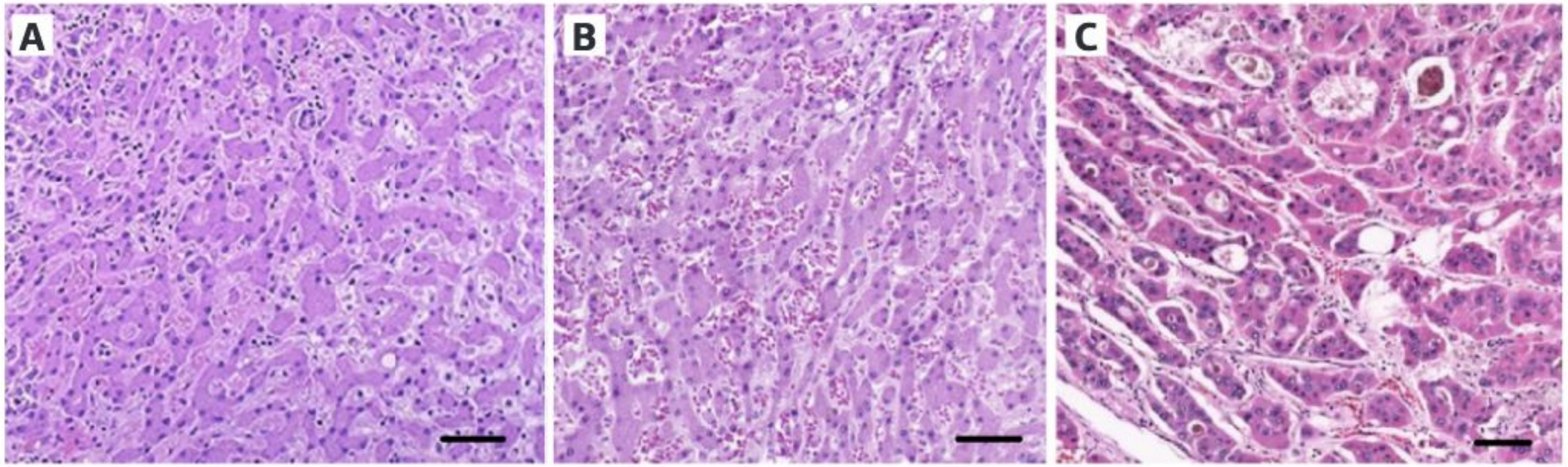
Ultrasound

X-Rays

Histology example

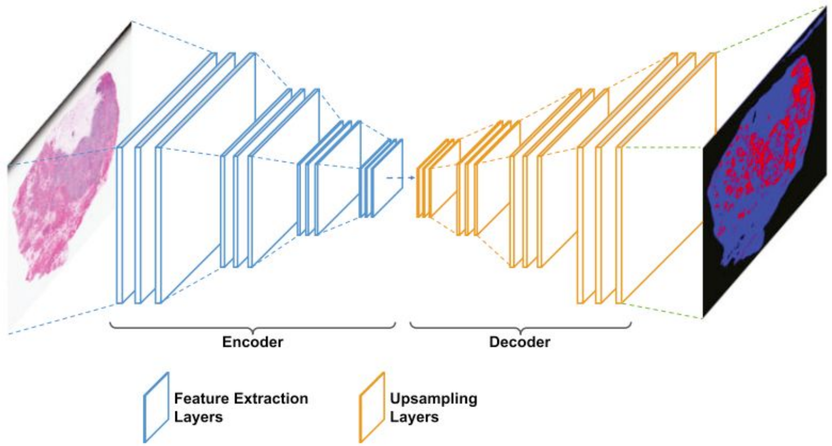
Wang, et al. 2019 https://www.jbuon.com/archive/24-4-1408.pdf

Mouse liver



Normal **Primary cancer** **Metastatic**

Common approaches



Convolutional Neural Net ->
colored map of major classes

figure: Tran, et al. 2021
https://doi.org/10.1186/s13073-021-00968-x

Feature type	Feature family
Shape	Morphology
First-order	Local Intensity
	Intensity-based Statistics
	Intensity Histogram
Second-order	Grey Level Co-occurrence Matrix (GLCM)
	Grey Level Run Length Matrix (GLRLM)
	Grey Level Size Zone Matrix (GLSZM)
	Grey Level Distance Zone Matrix (GLDZM)
	Neighborhood Grey Tone Difference Matrix (NGTDM)
	Neighboring Grey Level Dependence Matrix (NGLDM)
	Laws
	Gabor
Higher-order	Wavelets
	Laplacian of Gaussian (LoG)

Radiomic features
Ge & Zhang 2022 https://doi.org/10.1002/acm2.13869

Challenge: aligning multiple devices & institutions

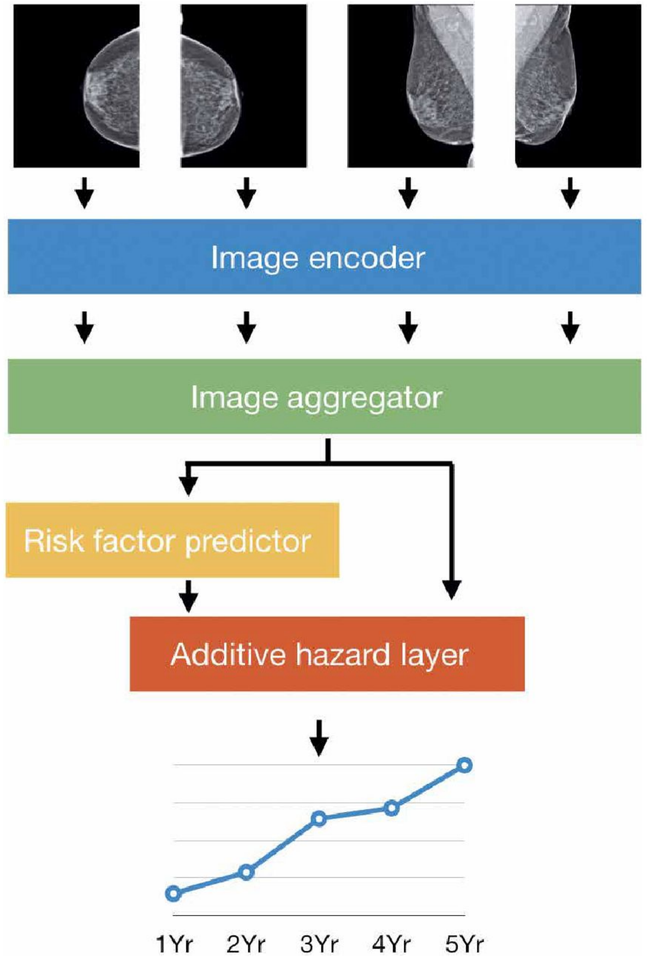
Mammography breast cancer risk prediction

Yala, et al. Toward robust mammography-based models for breast cancer risk. 2021 DOI: 10.1126/scitranslmed.aba4373

Challenges:

- consistently predict risk at multiple time points
- incorporate clinical factors (age and hormonal status)
- demonstrate generalizable performance

  Different devices



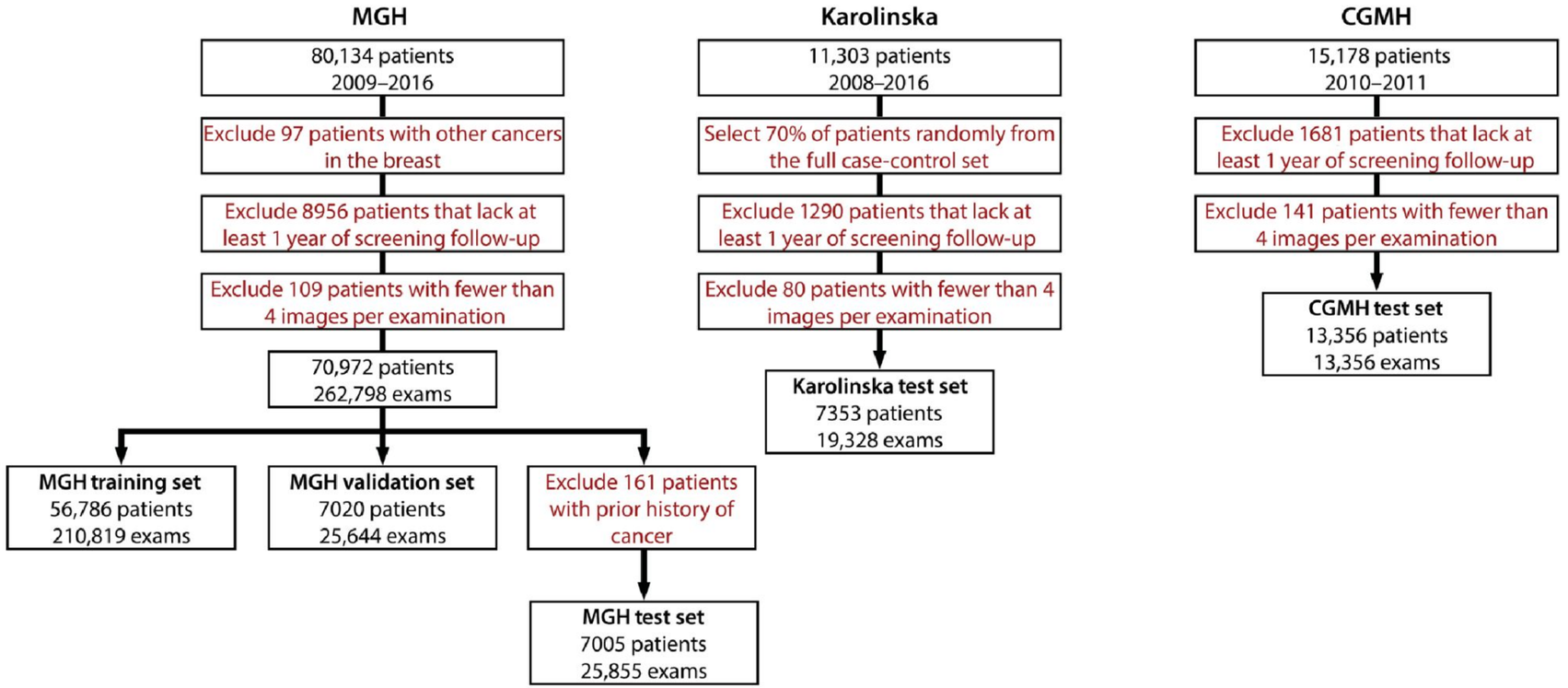
ResNet

Transformer w/attention-pooling

H_i layer predicts risk at year *i* based on other factors and previous year's risks

Generalization

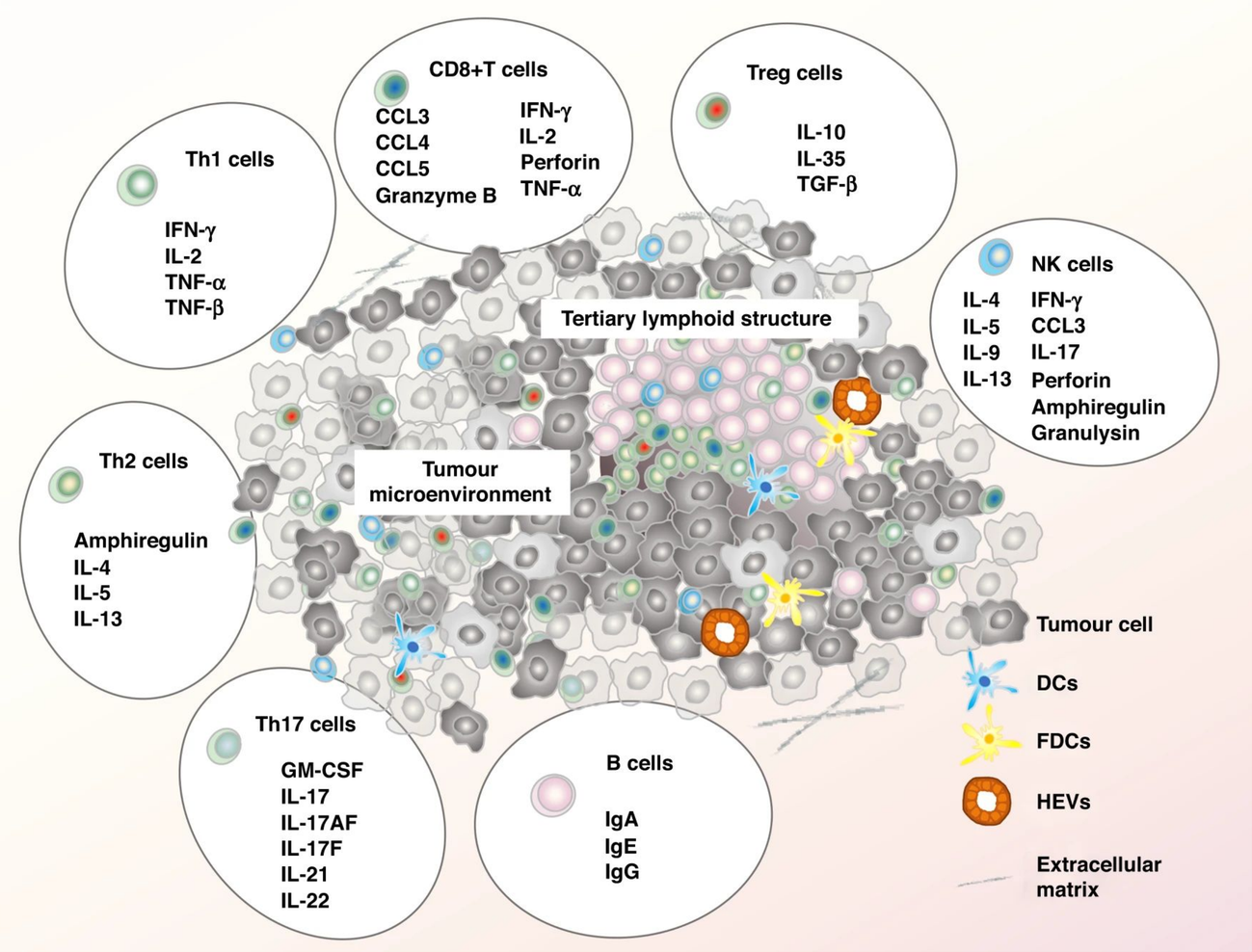
Yala, et al. Toward robust mammography-based models for breast cancer risk. 2021 **DOI: 10.1126/scitranslmed.aba4373**



Device bias: trained device classifier + conditional adversarial training

Immunology - Tumor Infiltrating Lymphocytes and the Tumor Microenvironment

Liu, et al. 2023 https://doi.org/10.1038/s41416-023-02321-y



TILS - immune cells
blood stream -> tumor

Purpose: tumor elimination

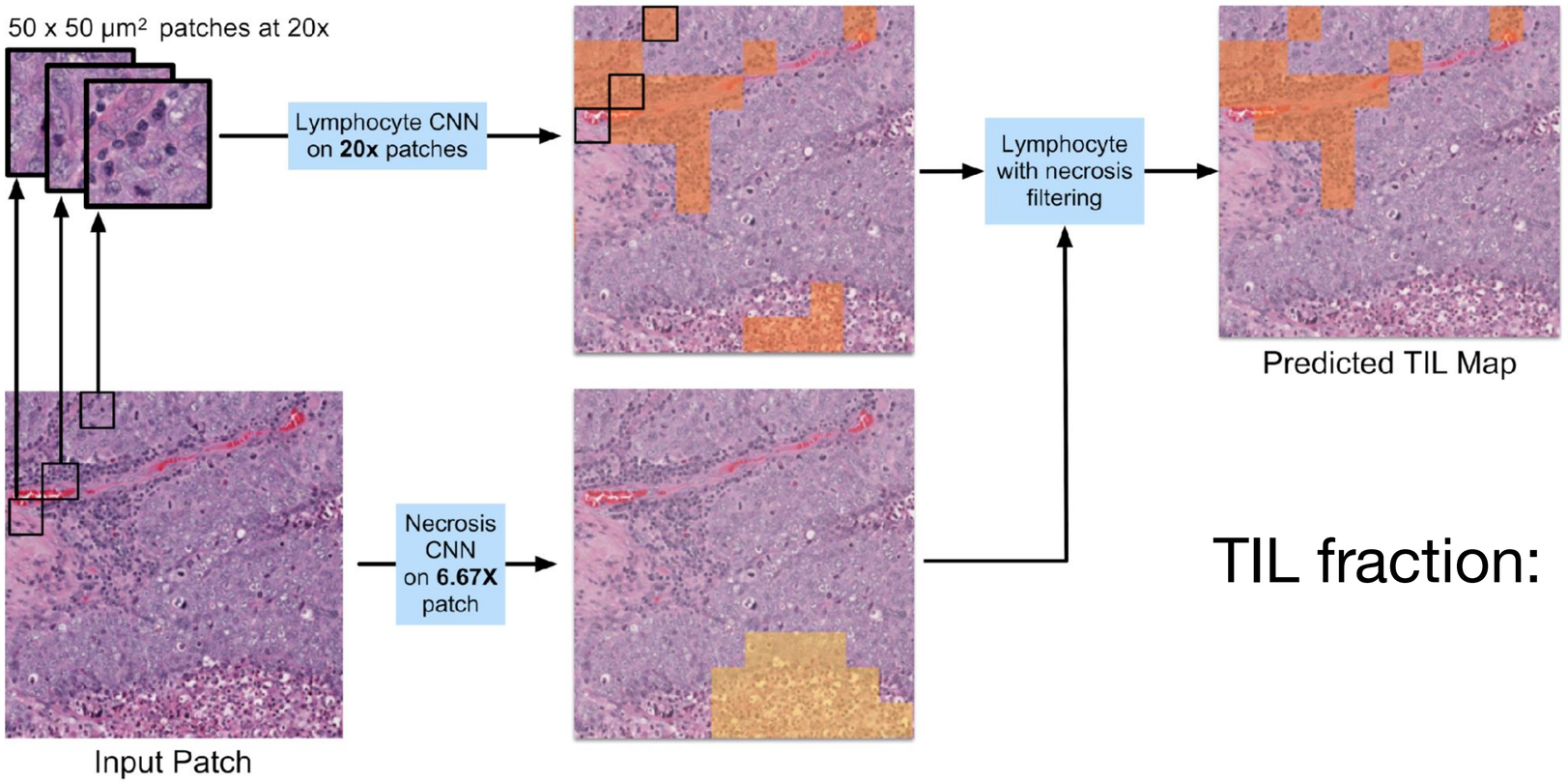
Spatial context is important: tumor micro-environment may suppress TIL activity

Existence may be rough indicator of survival and impact of immune checkpoint blockade therapy

Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images

Saltz, et al. 2019 https://doi.org/10.1016/j.celrep.2018.03.086

"Computational staining"



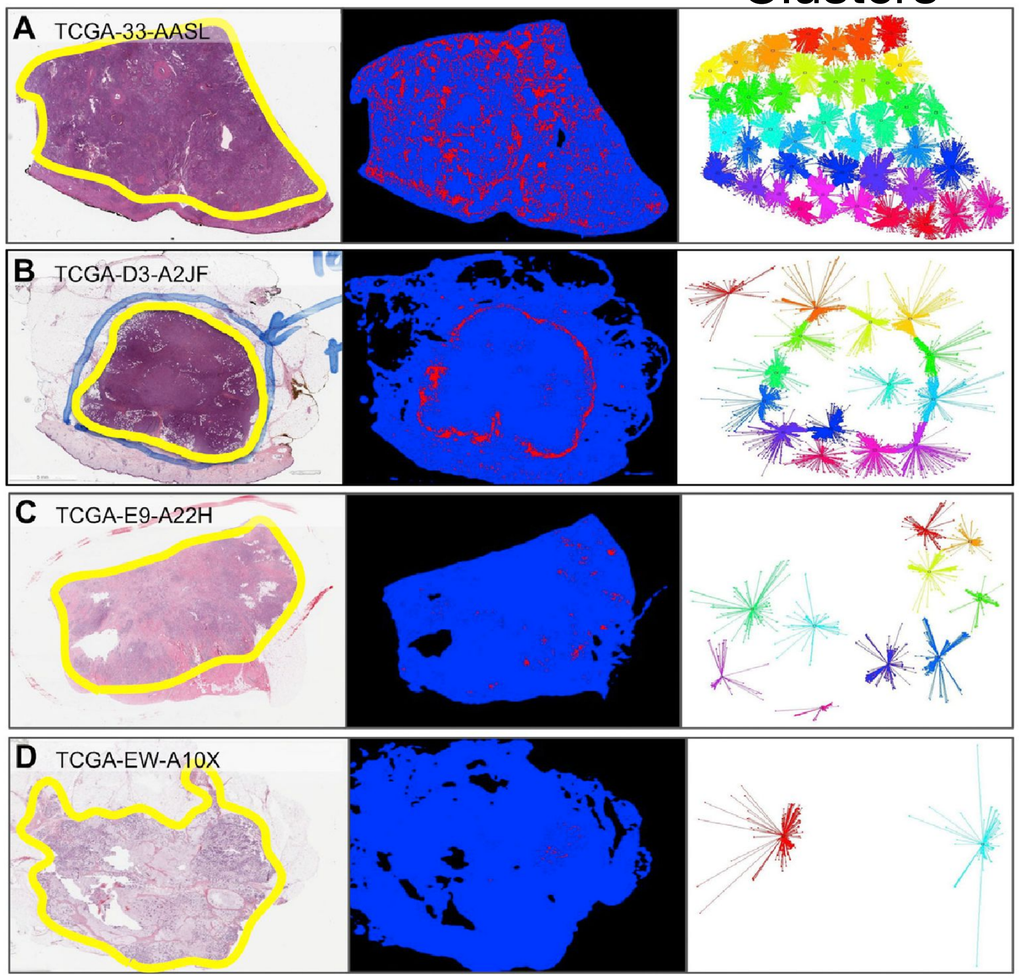
VGG-based CNNs

Images from TCGA

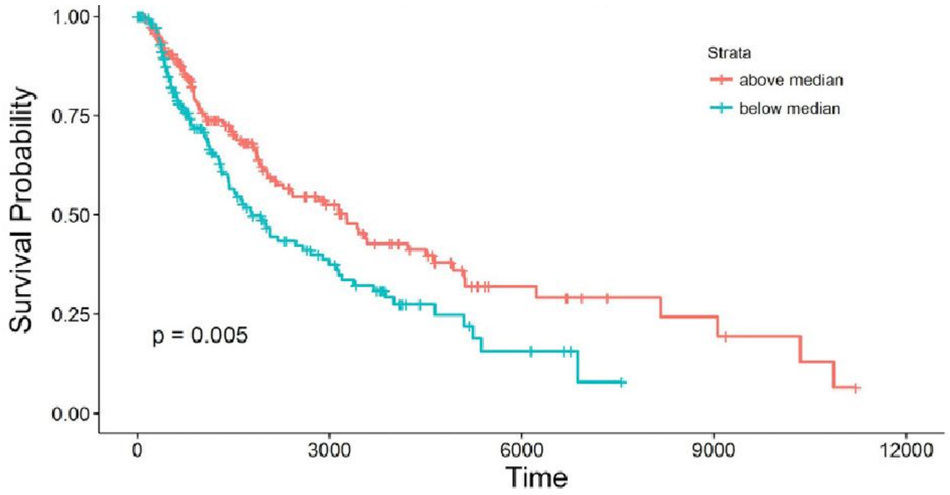
TIL fraction: = # positive patches/ total # of patches

Spatial Organization and Molecular Correlation…

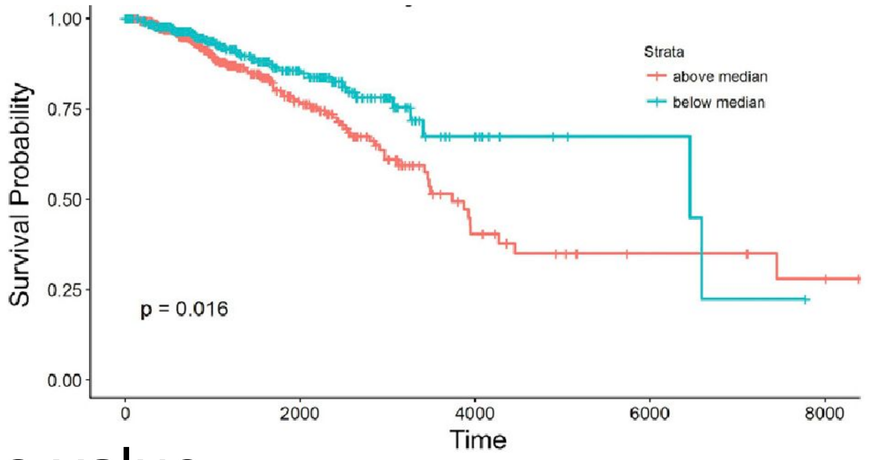
Tumor TIL map TIL Clusters



Melanoma



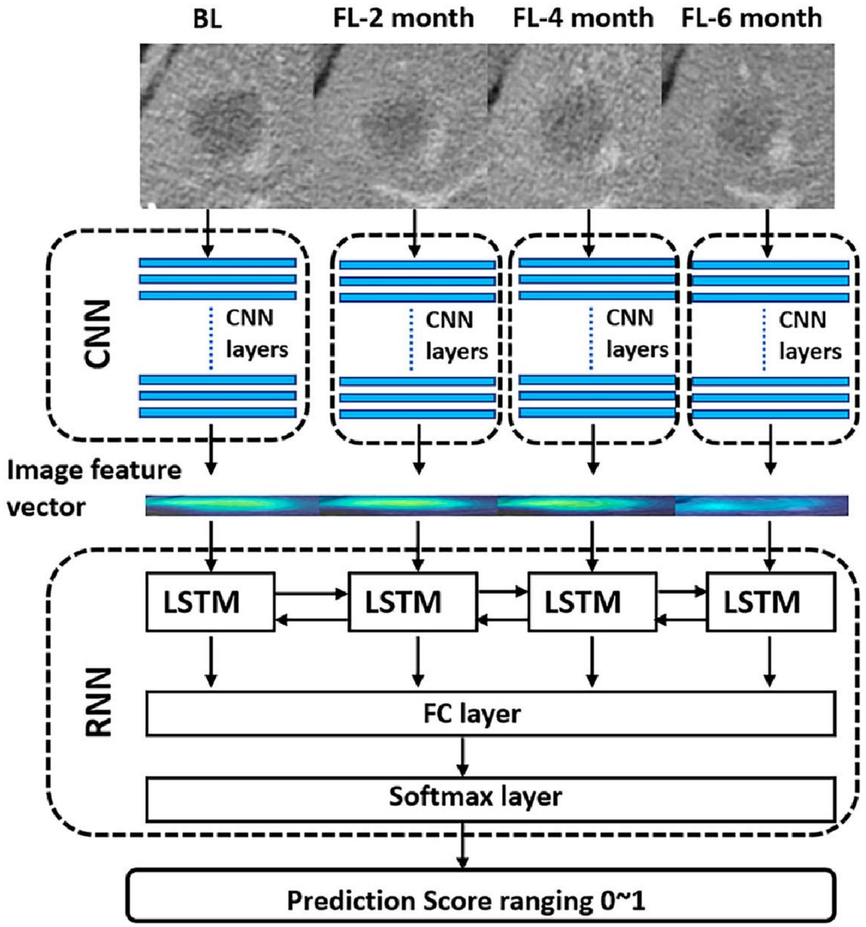
BRCA

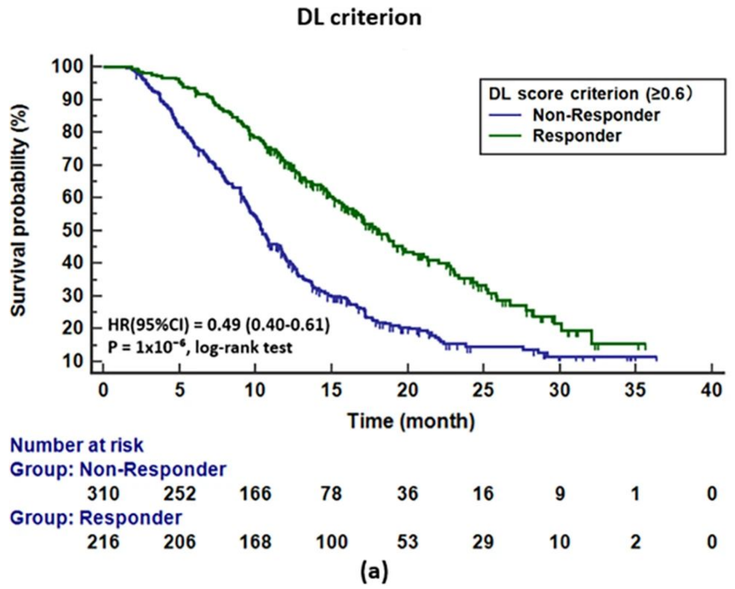


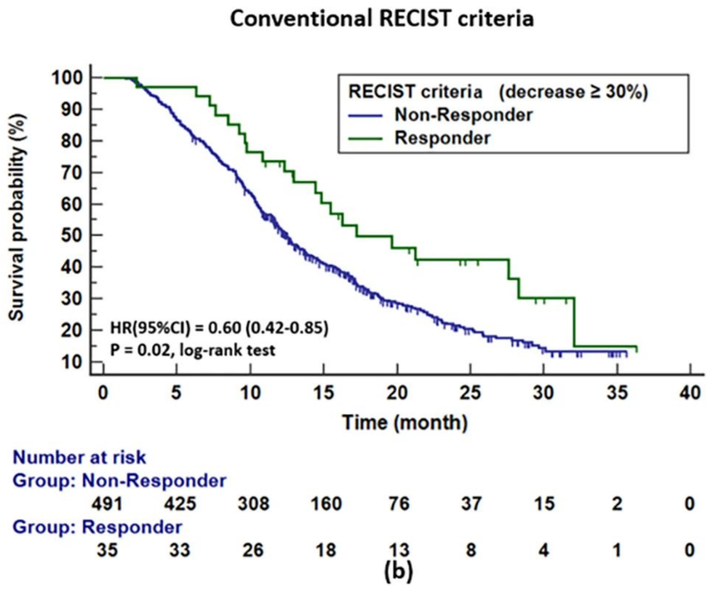
Moore, et al. 2021 - demonstrates prognostic value

Deep learning for the prediction of early on- treatment response in metastatic colorectal cancer from serial medical imaging

Lu, et al. 2021 https://doi.org/10.1038/s41467-021-26990-6







Resolving challenges in deep learning-based analyses of histopathological images using explanation methods

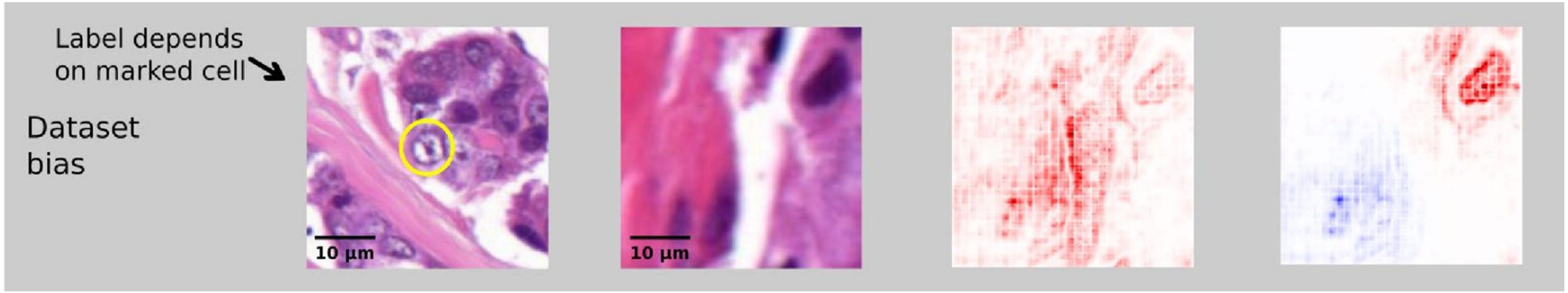
Hägele. et al. 2020 https://doi.org/10.1038/s41598-020-62724-2

- Can visual explanation methods help with systemic biases?
- melanoma datasets from TCGA
- manual annotation
- GoogLeNet CNN
- Layer-wise Relevance Propagation - distribute output backward
- Examine bias

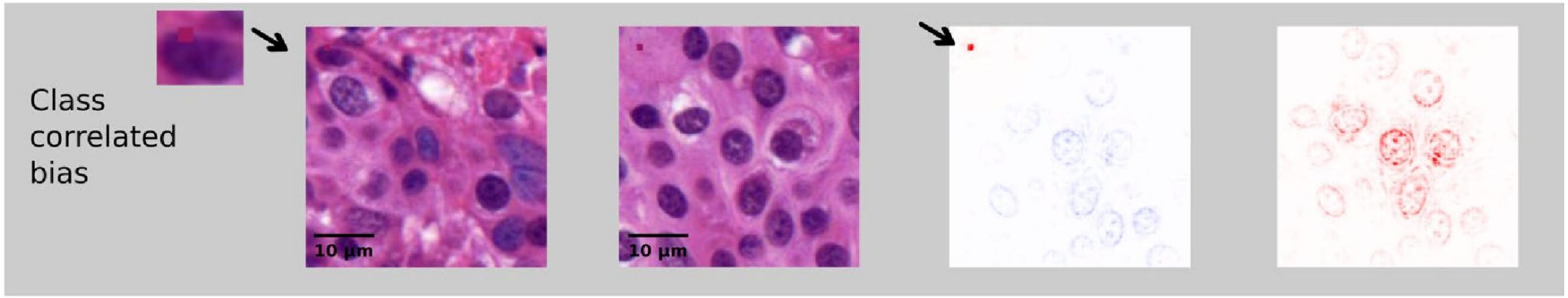
Resolving challenges in deep learning-based analyses of histopathological images using explanation methods

Hägele. et al. 2020 https://doi.org/10.1038/s41598-020-62724-2

Dataset bias - over focus on center of image. Fixed by spatial translation augmentation



Class-correlated bias - unintentionally correlated image features

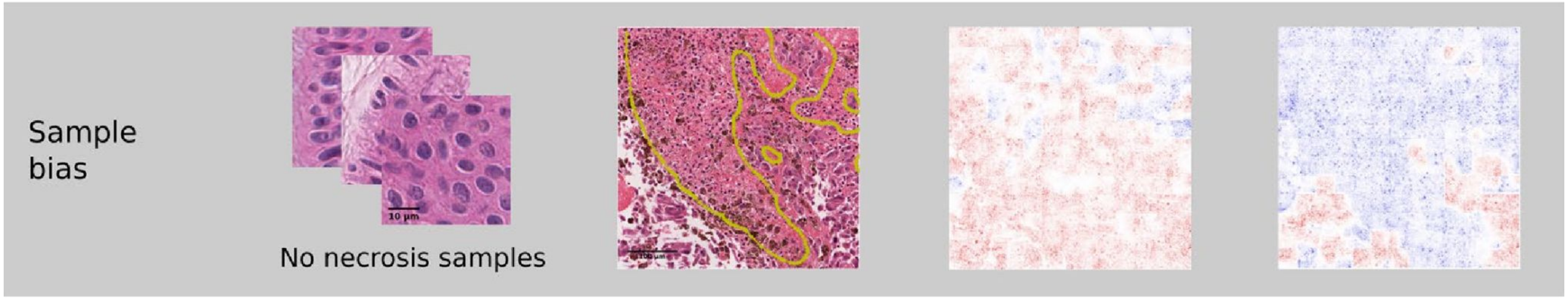


Resolving challenges in deep learning-based analyses of histopathological images using explanation methods

Hägele. et al. 2020 https://doi.org/10.1038/s41598-020-62724-2

Sample bias

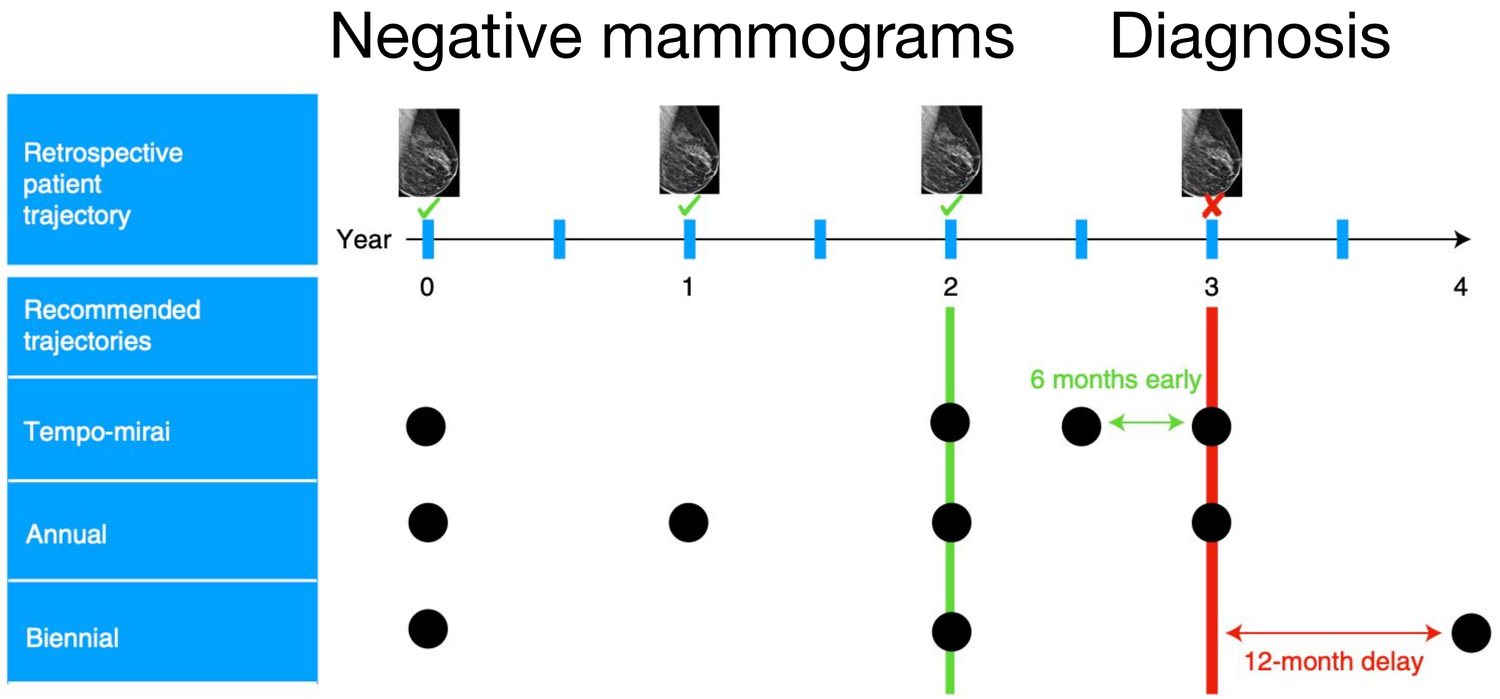
- omission of necrosis in training data leads to false positives



Optimizing risk-based breast cancer screening policies with reinforcement learning

Yala, et al. 2022 https://doi.org/10.1038/s41591-021-01599-w

Manage follow-up time for breast cancer

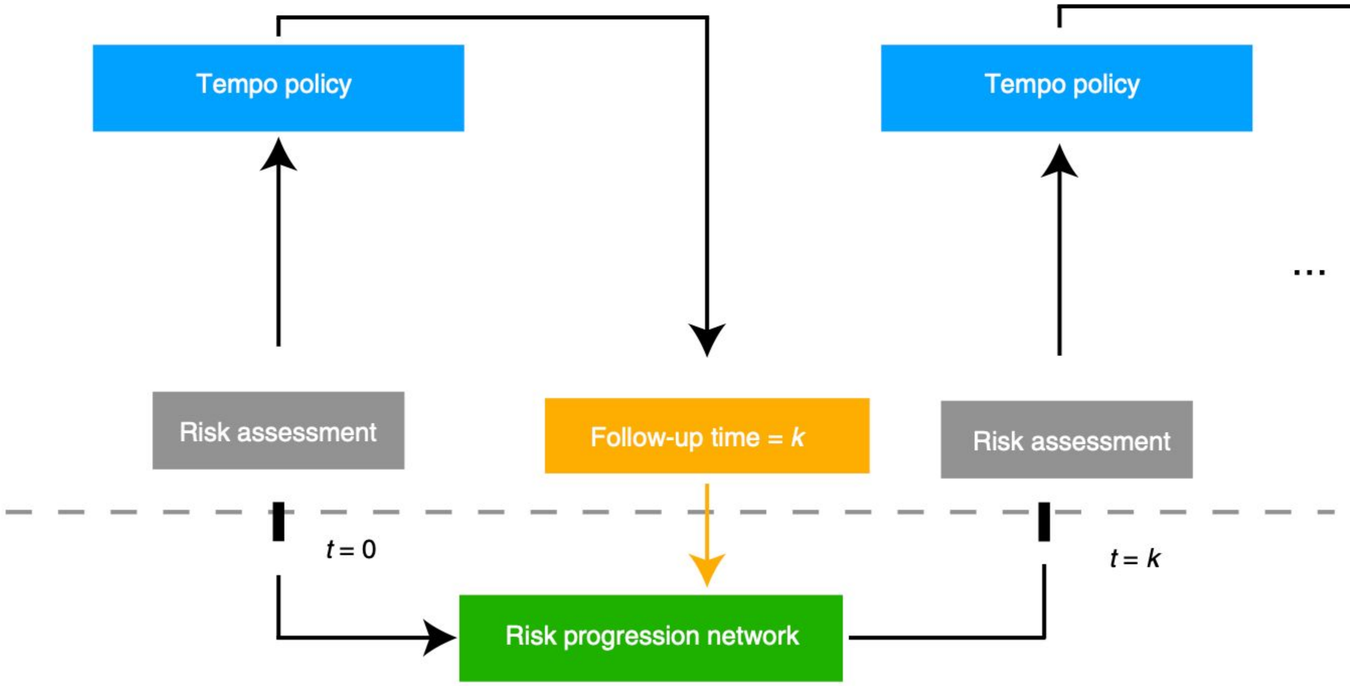


Can we increase early detection benefit?

Optimizing risk-based breast cancer screening policies with reinforcement learning

Yala, et al. 2022 https://doi.org/10.1038/s41591-021-01599-w

Manage follow-up time for breast cancer



deep learning risk assessment - mammograms

Rewards:

1. imaging cost - decrease in recommended mammograms/year
2. early detection benefit difference between recommended screening and diagnosis

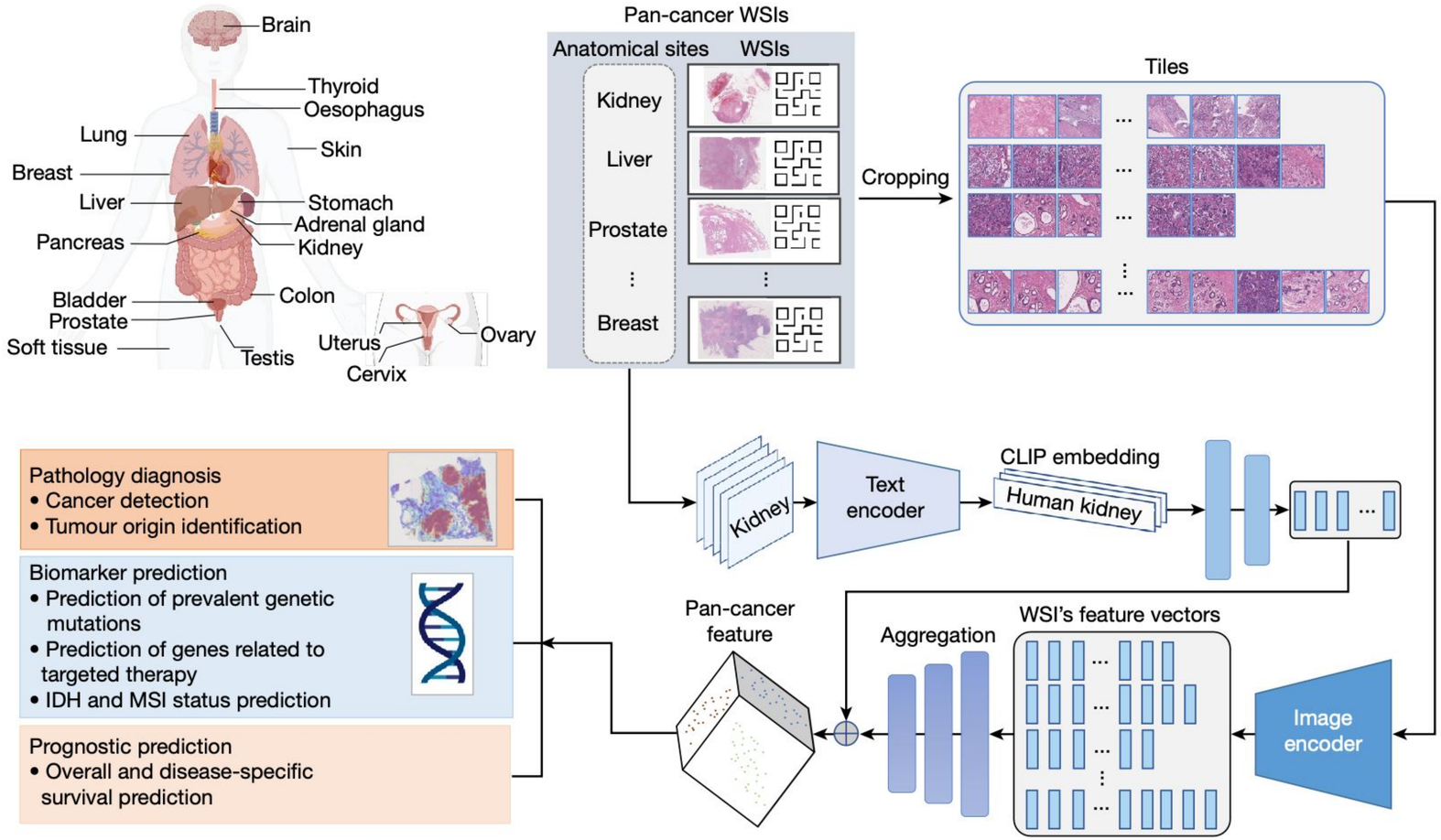
Deep RL

MGH + 2 validation sets

A pathology foundation model for cancer diagnosis and prognosis prediction

Wang, et al. 2024

https://doi.org/10.1038/s41586-024-07894-z



- Contrastive language-image pretraining - features for textual descriptions (images and captions)
- Unsupervised pretraining on 15M unlabelled tiles
- Heterogenous public databases
- Weakly supervised pretraining on 60K whole slide images
- 44 TB training data
- Validation - 19K whole slide images from 32 independent sets from 24 cohorts internationally.

Clinical data

Structured data

Tabular data from electronic medical record

- Demographics
- Diagnoses
- Procedures
- Medications (orders, administrations)
- Labs
- Vitals

Structured data presents phenotyping challenges

- Incorrect

- Incomplete: particularly for complex patients who receive care in multiple locations/systems

- Inconclusive: Does a C.50 ICD10 code (Malignant neoplasm of breast) mean that the patient has BrCa?

hence, NLP - clinical notes…

Many targets of Cancer NLP

Tumor/Cancer characteristics

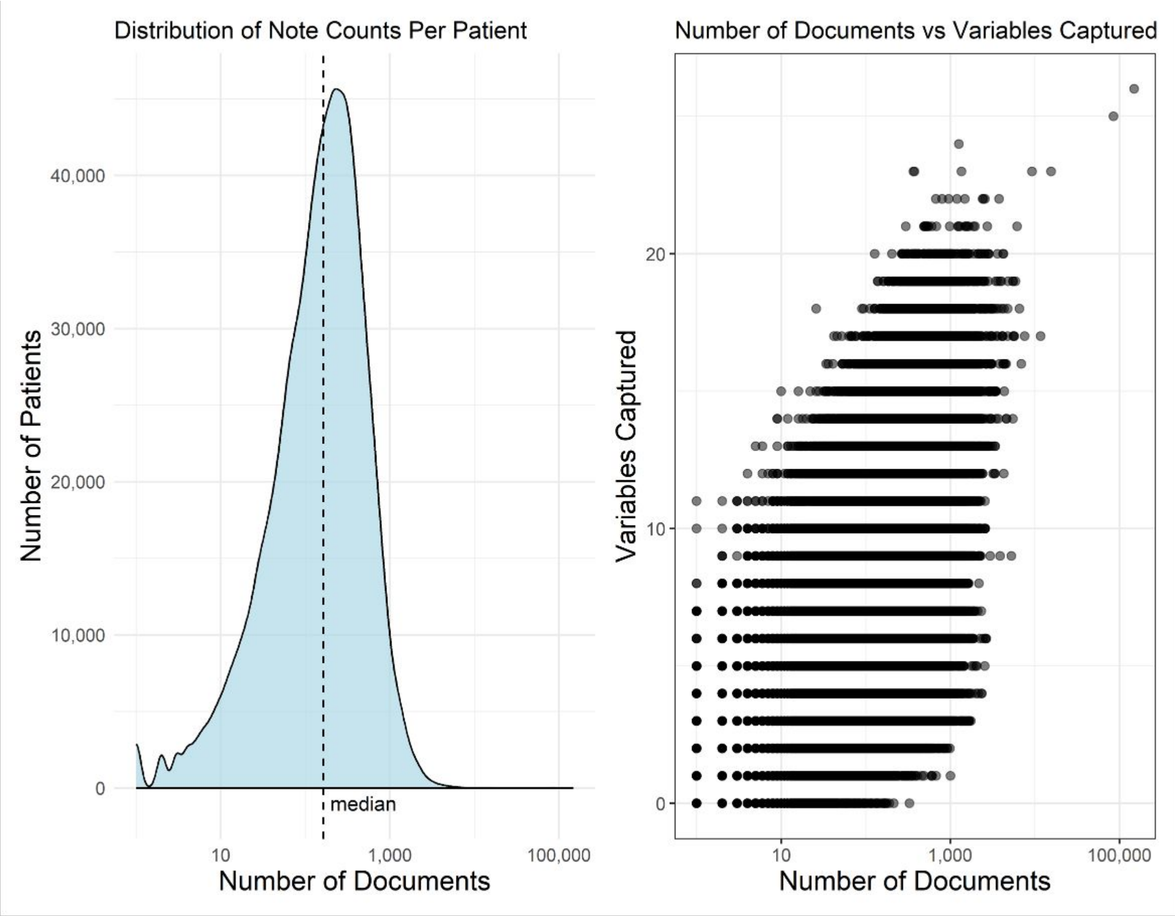
- Diagnosis
- Site
- Grade
- Laterality
- Stage

Patient Characteristics

- History
- Family History
- Treatments
- Procedures
- Comorbidities
- Reactions
- SDoH

All vary over time!

Volume of data makes cancer NLP challenging



Potentially 1000s of documents over years

“pt has history of BrCa”

meaning evolves over time

Too much detail to look at individual notes

need summarization

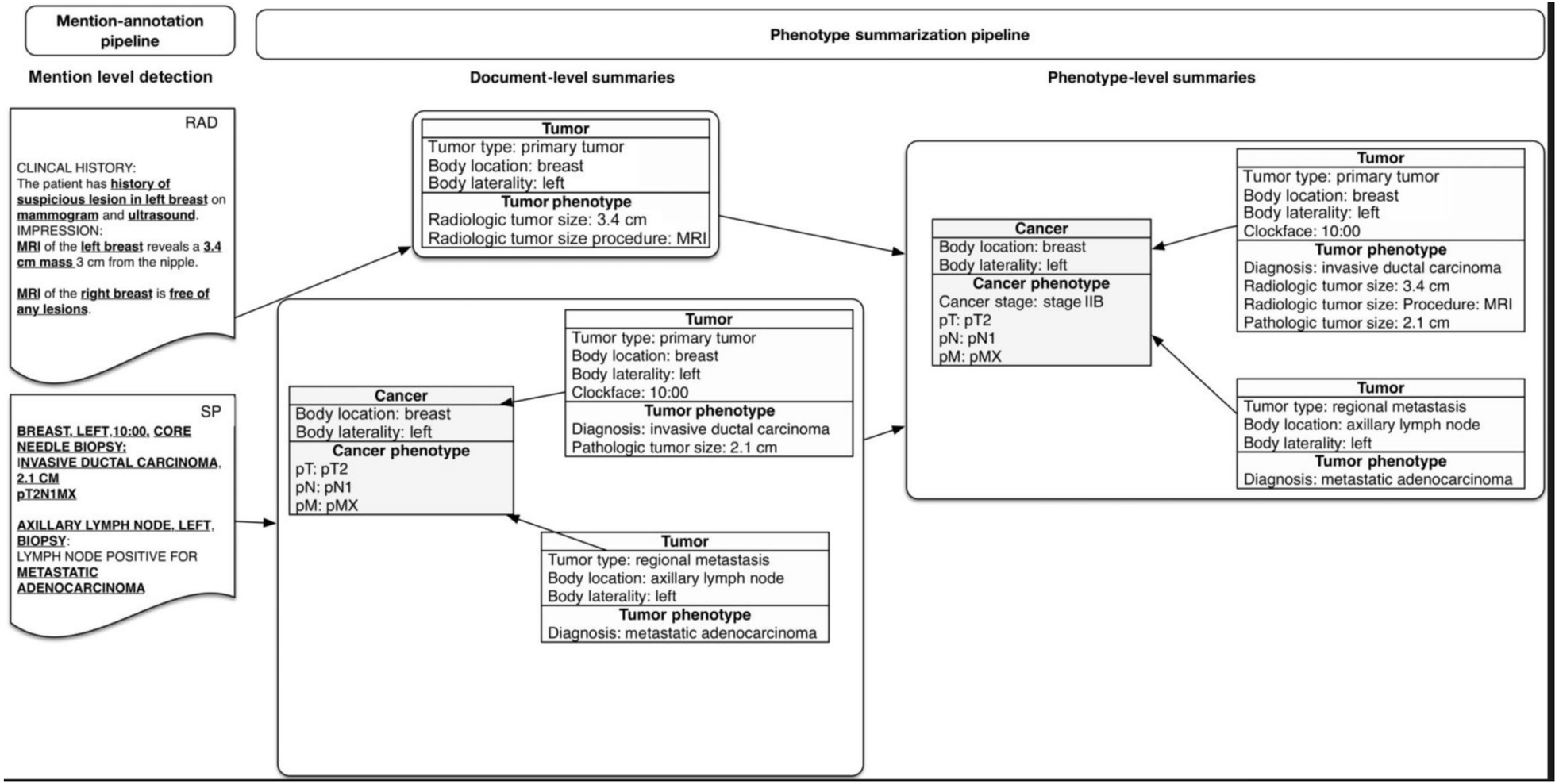
coreference resolution:

when does “the tumor” in multiple notes refer to the same thing?

Figure from A. VanHelene

DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records

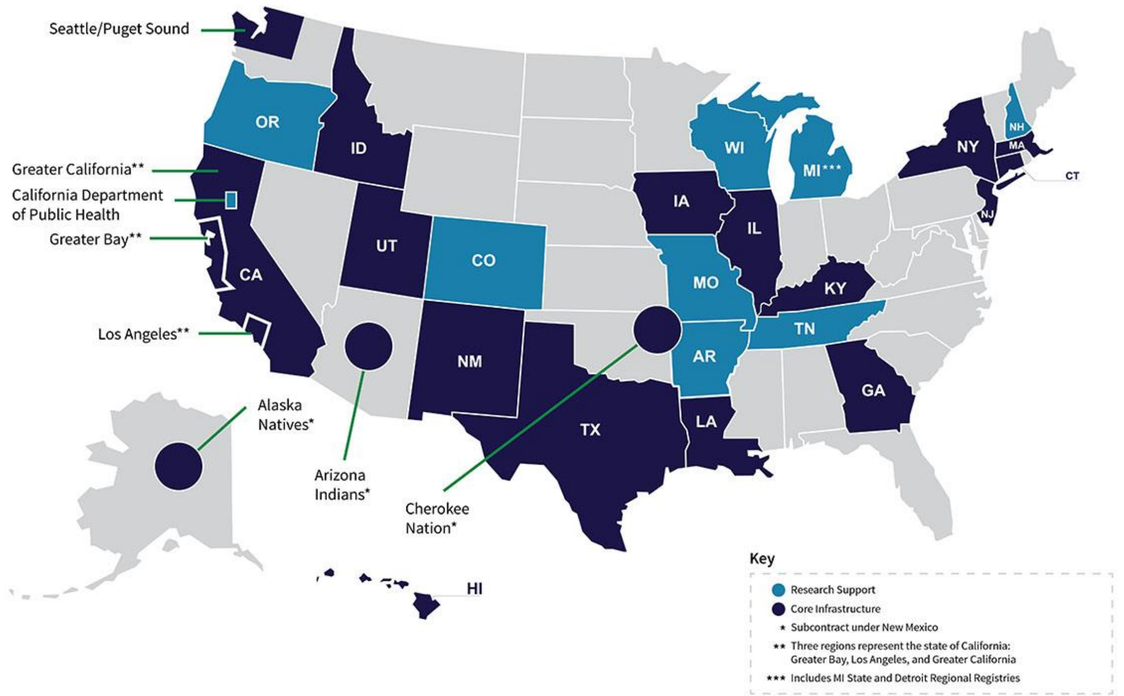
Savova, et al. 2017 https://doi.org/10.1158/0008-5472.CAN-17-0615



Cancer registries and data abstraction

Epidemiological tracking of cancer

- NCI Surveillance, Epidemiology, and End Results Program (SEER) - https://seer.cancer.gov

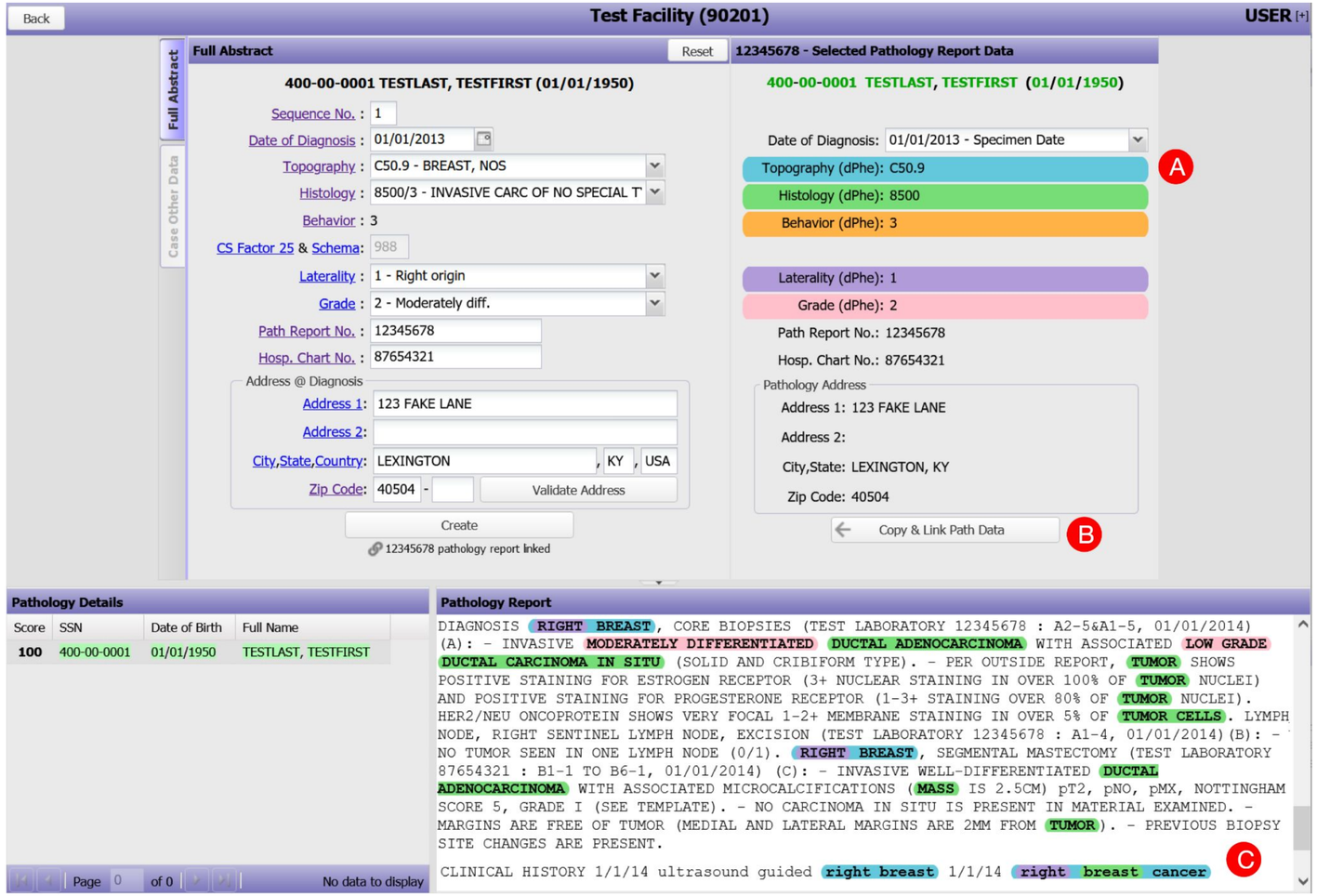


- State and local registries
- Challenge - abstracting data
  - SEER Coding manual - 279 pages
    https://seer.cancer.gov/manuals/2024/SPCSM_2024_MainDoc.pdf

DeepPhe-CR: Natural Language Processing Software for Cancer Registrar Case Abstraction Hochheiser, et al. 2023 https://doi.org/10.1200/CCI.23.00156

Using NLP to reduce costs of abstraction

- Adapt DeepPhe: Containerized API implementation
- Goal > 95% accuracy

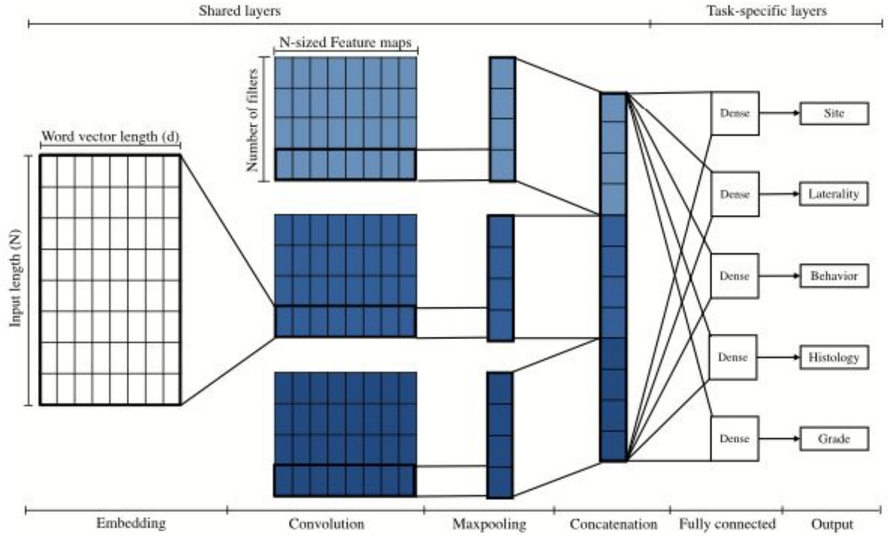


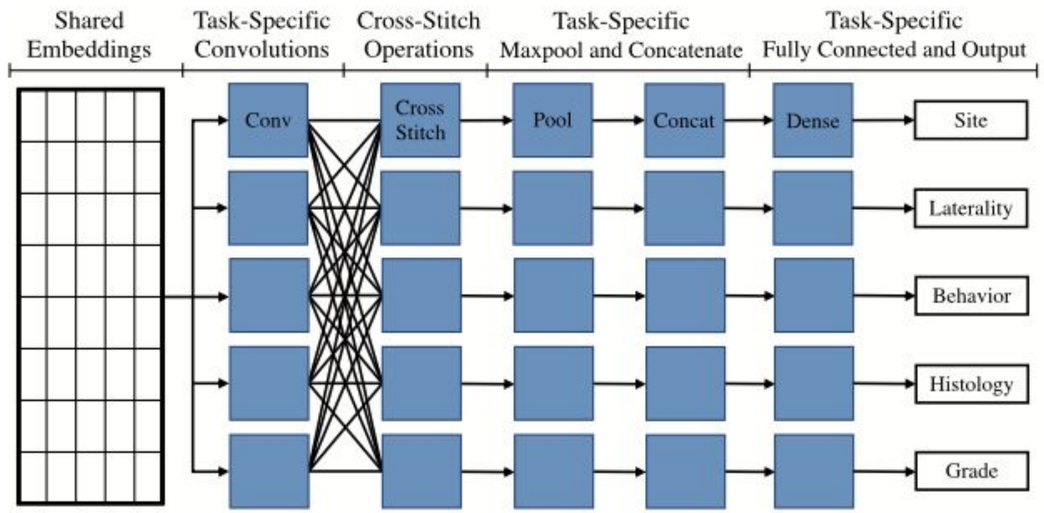
Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks

Alawad, et al. 2019 DOI: 10.1093/jamia/ocz153

Multi-Task Learning

Convolutional nets operate on embeddings of context of /-word vectors





Multi-class CNNs outperform single-class and other models

Transfer learning: fine-tuning and joint training improve performance over baselin

Alawad 2019 (DOI: 10.1109/BHI.2019.8834586)

Large Language models

Article | Open access | Published: 12 July 2023

Large language models encode clinical knowledge

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, ... Vivek Natarajan + Show authors

Nature **620**, 172–180 (2023) | Cite this article

https://doi.org/10.1038/s41586-023-06291-2

Incorporating Clinical Guidelines Through Adapting Multi-modal Large Language Model for Prostate Cancer PI-RADS Scoring

Conference paper | First Online: 04 October 2024

https://doi.org/10.1007/978-3-031-72086-4_34

Large language model (ChatGPT) as a support tool for breast tumor board

Vera Sorin, Eyal Klang, Miri Sklair-Levy, Israel Cohen, Douglas B. Zippel, Nora Balint Lahat, Eli Konen & Yiftach Barash

https://doi.org/10.1038/s41523-023-00557-8

Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer

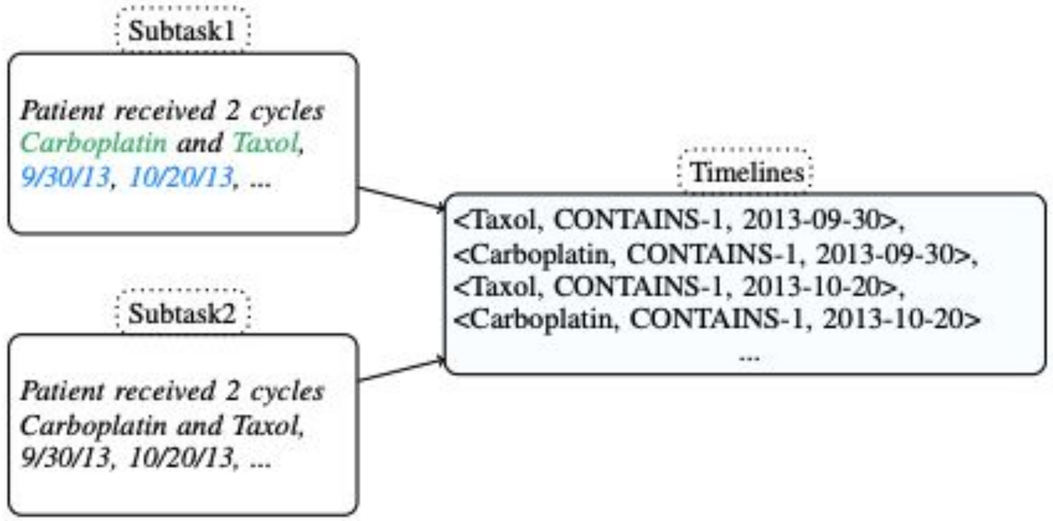
Matthias A. Fink, Arved Bischoff, Christoph A. Fink, Martin Moll, Jonas Kroschke, Luca Dulz, Claus Peter Heußel, Hans-Ulrich Kauczor, Tim F. Weber

Author Affiliations

Published Online: Sep 19 2023 | https://doi.org/10.1148/radiol.231362

Overview of the 2024 Shared Task on Chemotherapy Treatment Timeline Extraction Yao, et al. https://aclanthology.org/2024.clinicalnlp-1.53/

Extracting temporally ordered treatment information



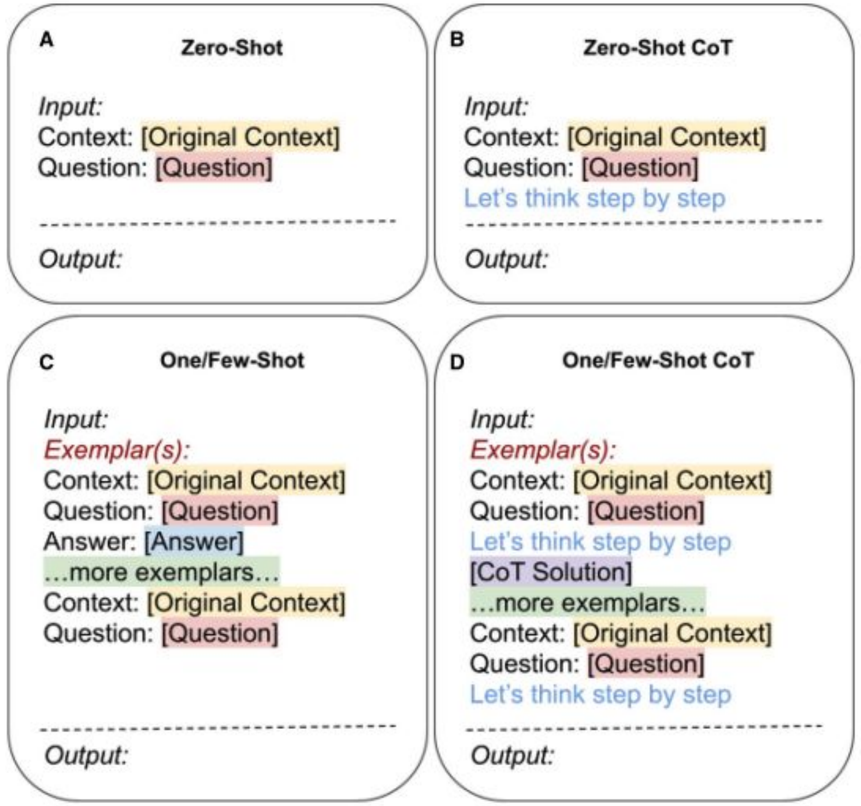
9 teams, 19 submissions

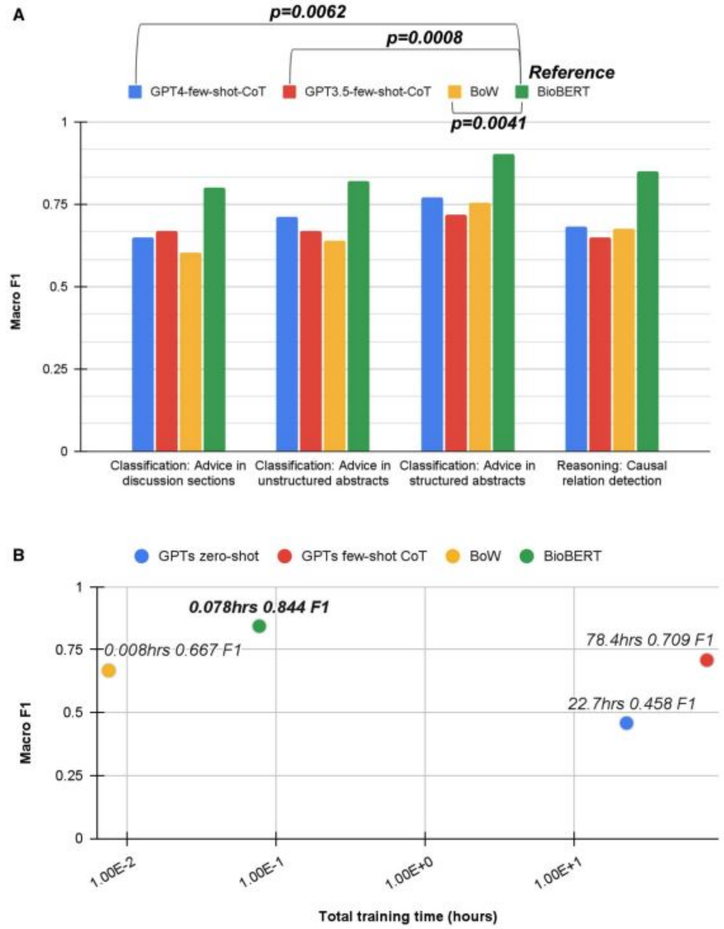
“Perhaps surprising in our current era of very large LMs, fine-tuned smaller LMs achieved superior performance. This discrepancy between prompting LLMs and fine tuning smaller-sized LMs suggests that more sophisticated LLMs or prompting techniques are necessary in order to achieve optimal results for challenging tasks such as patient-level chemotherapy timeline extraction.”

Evaluating the ChatGPT family of models for biomedical reasoning and classification

Chen, et al. 2024 DOI: 10.1093/jamia/ocad256

- Evaluate LLMs on two tasks
  - classification
  - reasoning





Large language models to identify social determinants of health in electronic medical records

Guevera,et al. 2024 https://doi.org/10.1038/s41746-023-00970-0

- “SDoH are estimated to account for 80–90% of modifiable factors impacting health outcomes”

- Employment
- Housing
- Transportation
- Parental status
- Social support

“Our fine-tuned models outperformed ChatGPT-family models with zero- and few-shot learning for most SDoH classes and were less sensitive to the injection of demographic descriptors”

The Cancer Genome Atlas

https://www.cancer.gov/ccg/research/genome-sequencing/tcga

> 20,000 tumor and normal samples

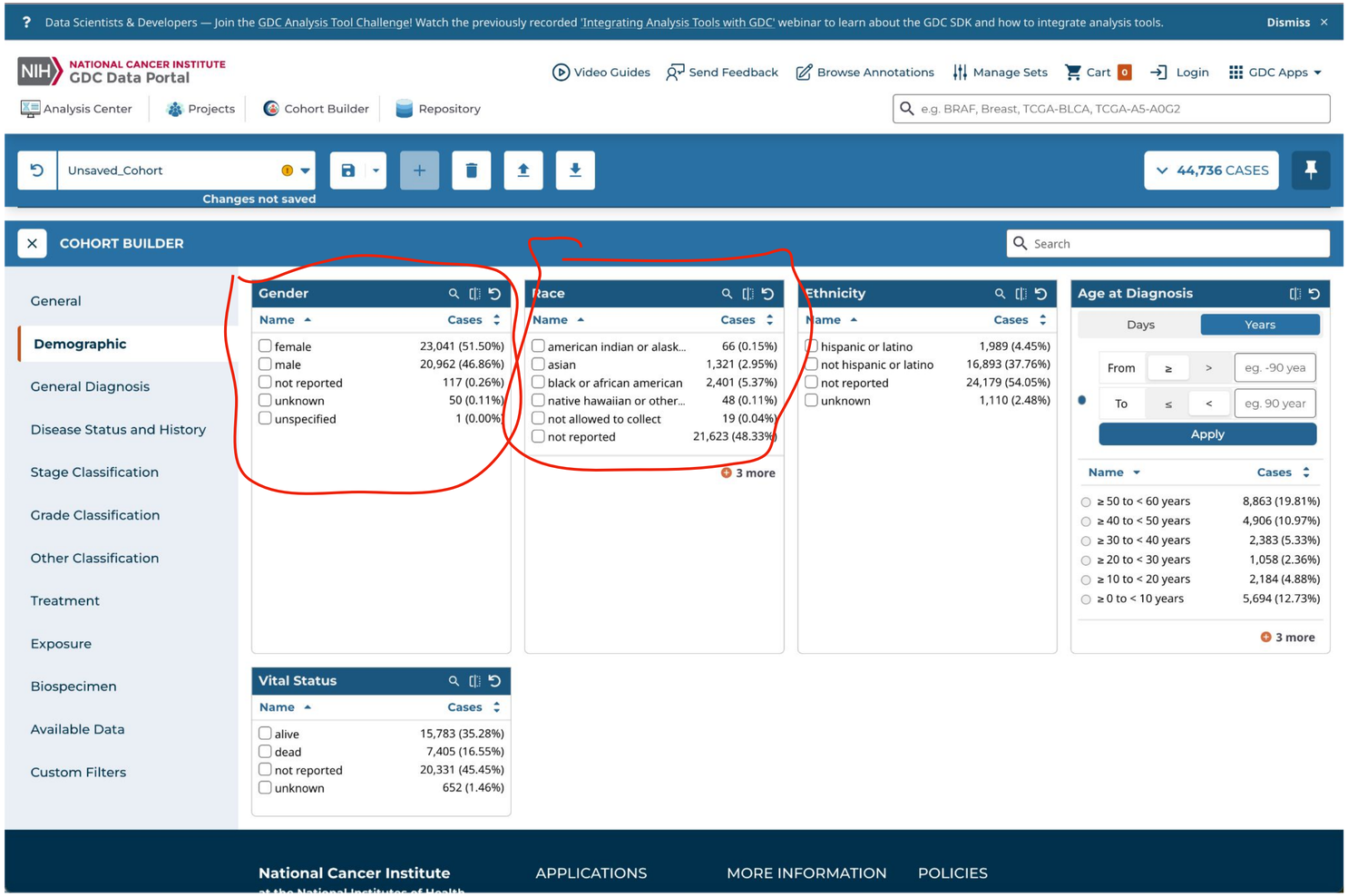
Data types

. clinical information (e.g., smoking status)
. molecular analyte metadata (e.g., sample portion weight)
. molecular characterization data (e.g., gene expression values)

details: https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data/types

TCGA: Potential Biases?

Genomic Data Commons - harmonized cancer datasets: https://portal.gdc.cancer.gov/



TCGA: Additional Biases?

Most (all?) data is from the US

Is this a problem?

Article | Open access | Published: 03 September 2018

Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls

Hamid Behravan ✉, Jaana M. Hartikainen, Maria Tengström, Katri Pylkäs, Robert Winqvist, Veli–Matti Kosma & Arto Mannermaa

Behraven, et al. 2018 https://doi.org/10.1038/s41598-018-31573-5

See also https://www.icgc-argo.org/ -100K patients

More on Bias

Kaushal, et al. 2022 doi:10.1001/jama.2020.12067

74 deep learning studies 2015-2019

Which geographic locations were over-represented?

under-represented?

Table. US Patient Cohorts Used for Training Clinical Machine Learning Algorithms, by State[a]

States	No. of studies
California	22
Massachusetts	15
New York	14
Pennsylvania	5
Maryland	4
Colorado	2
Connecticut	2
New Hampshire	2
North Carolina	2
Indiana	1
Michigan	1
Minnesota	1
Ohio	1
Texas	1
Vermont	1
Wisconsin	1

[a] Fifty-six studies used 1 or more geographically identifiable US patient cohorts in the training of their clinical machine learning algorithm. Thirty-four states were not represented in geographically identifiable cohorts: Alabama, Alaska, Arizona, Arkansas, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Iowa, Kansas, Kentucky, Louisiana, Maine, Mississippi, Missouri, Montana, Nebraska, Nevada, New Jersey, New Mexico, North Dakota, Oklahoma, Oregon, Rhode Island, South Carolina, South Dakota, Tennessee, Utah, Virginia, Washington, West Virginia, and Wyoming.

Additional potential concerns

Which patients?

Where are people enrolled in trials?

Retrospective vs. Prospective?

What would be needed to translate these models into care?

Acknowledgments

NIH Grants: UH3CA243120, U24CA248010

Boston Children's Hospital: Guergana Savova, Sean Finan, Jiarui Yao, Dennis Johns, WonJin Yoon, Eli Goldner

Brown University: Jeremy Warner

Kentucky Cancer Registry: Eric Durbin, Isaac Hands, David Rust, Ram Kavuluru

University of Pittsburgh: John Levander, Zhou Yuan