Learning in Limited Data Settings

Advancing Personalized Medicine in Cancer Treatment Planning*

Vaibhav Rajan
Google DeepMind

*work funded at National University of Singapore

One day....

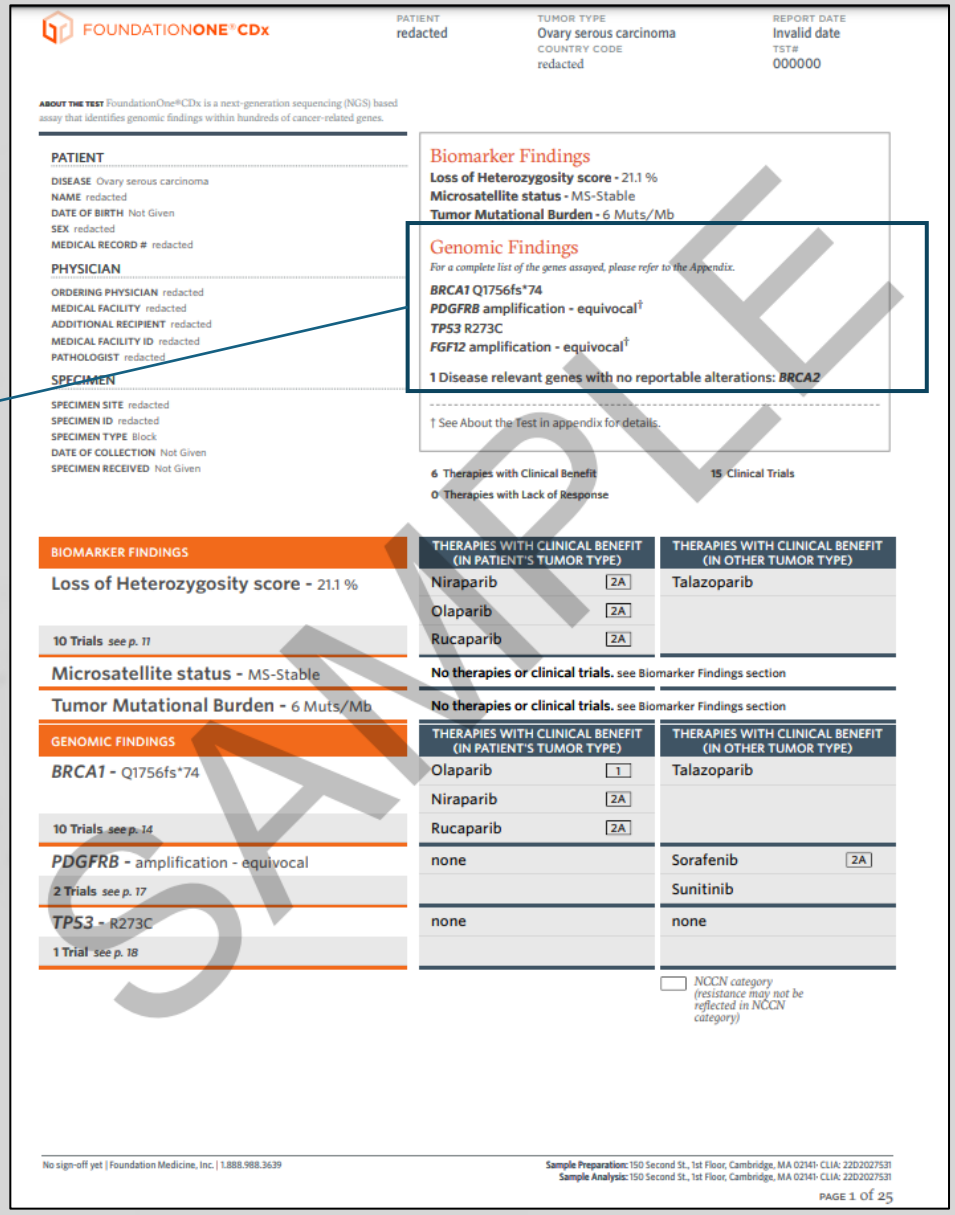
BRCA1 Q1756fs*74

PDGFRB amplification

TP53 R273C

FGF12 amplification

Mutations identified



FOUNDATIONONE®CDx

PATIENT redacted

TUMOR TYPE Ovary serous carcinoma

COUNTRY CODE redacted

REPORT DATE Invalid date

TST# 000000

ABOUT THE TEST FoundationOne®CDx is a next-generation sequencing (NGS) based assay that identifies genomic findings within hundreds of cancer-related genes.

PATIENT

DISEASE Ovary serous carcinoma
NAME redacted
DATE OF BIRTH Not Given
SEX redacted
MEDICAL RECORD # redacted

PHYSICIAN

ORDERING PHYSICIAN redacted
MEDICAL FACILITY redacted
ADDITIONAL RECIPIENT redacted
MEDICAL FACILITY ID redacted
PATHOLOGIST redacted

SPECIMEN

SPECIMEN SITE redacted
SPECIMEN ID redacted
SPECIMEN TYPE Block
DATE OF COLLECTION Not Given
SPECIMEN RECEIVED Not Given

Biomarker Findings

Loss of Heterozygosity score - 21.1 %
Microsatellite status - MS-Stable
Tumor Mutational Burden - 6 Muts/Mb

Genomic Findings

For a complete list of the genes assayed, please refer to the Appendix.

BRCA1 Q1756fs*74
PDGFRB amplification - equivocal†
TP53 R273C
FGF12 amplification - equivocal†

1 Disease relevant genes with no reportable alterations: ***BRCA2***

† See About the Test in appendix for details.

6 Therapies with Clinical Benefit
0 Therapies with Lack of Response
15 Clinical Trials

BIOMARKER FINDINGS	THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)	THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)
Loss of Heterozygosity score - 21.1 %	Niraparib 2A Olaparib 2A Rucaparib 2A	Talazoparib
10 Trials see p. 11		
Microsatellite status - MS-Stable	No therapies or clinical trials. see Biomarker Findings section	
Tumor Mutational Burden - 6 Muts/Mb	No therapies or clinical trials. see Biomarker Findings section	
GENOMIC FINDINGS	**THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)**	**THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)**
BRCA1 - Q1756fs*74	Olaparib 1 Niraparib 2A Rucaparib 2A	Talazoparib
10 Trials see p. 14		
PDGFRB - amplification - equivocal	none	Sorafenib 2A Sunitinib
2 Trials see p. 17		
TP53 - R273C	none	none
1 Trial see p. 18		

NCCN category (resistance may not be reflected in NCCN category)

SAMPLE

No sign-off yet | Foundation Medicine, Inc. | 1.888.988.3639

Sample Preparation: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531
Sample Analysis: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531

PAGE 1 of 25

Targeted therapy: “one gene, one target”

Majority of patients remain untreated by targeted therapies

Can AI help us identify the right drug for such cancer patients?

NORMAL CELL AND CANCER CELL DEVELOPMENT

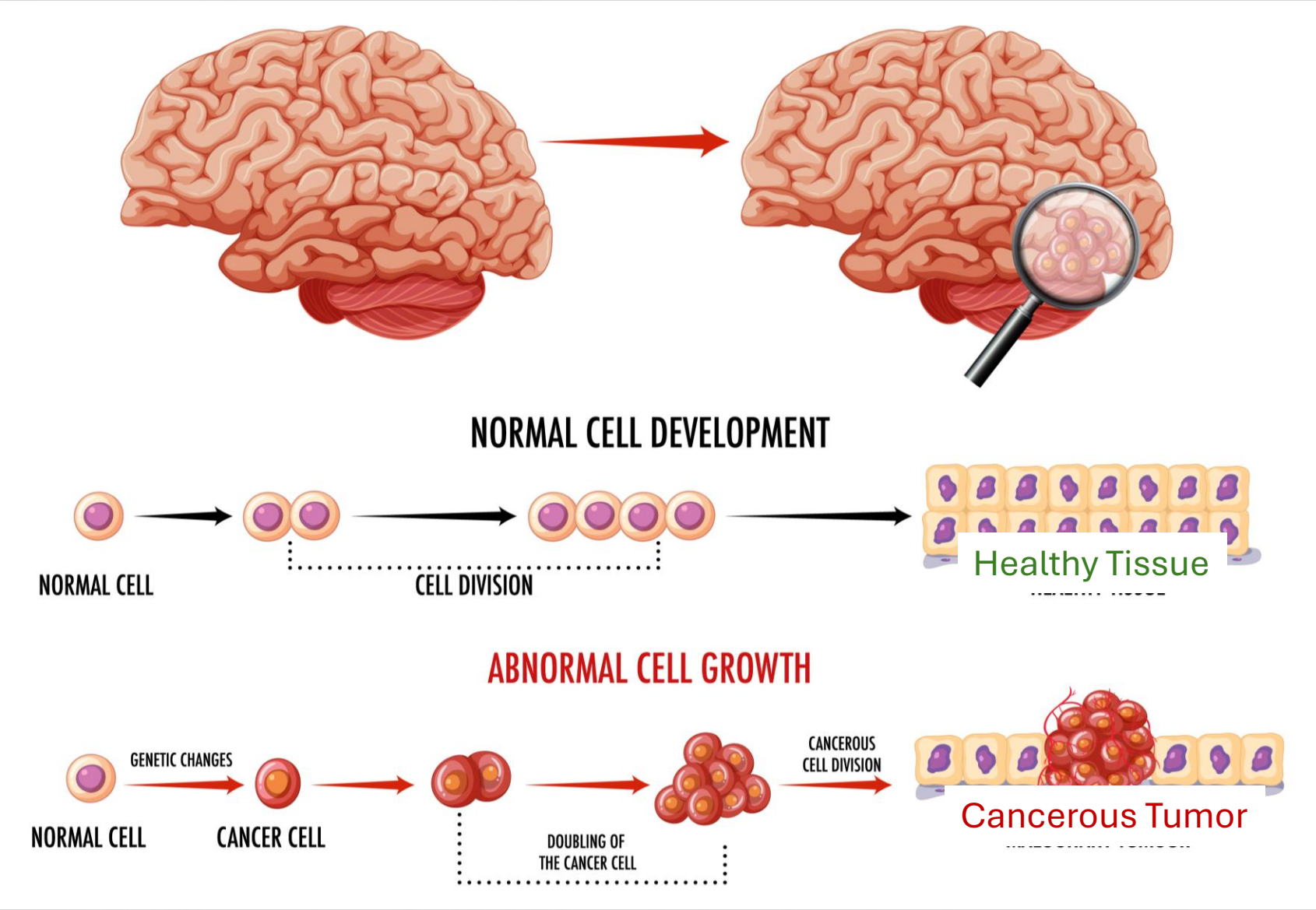
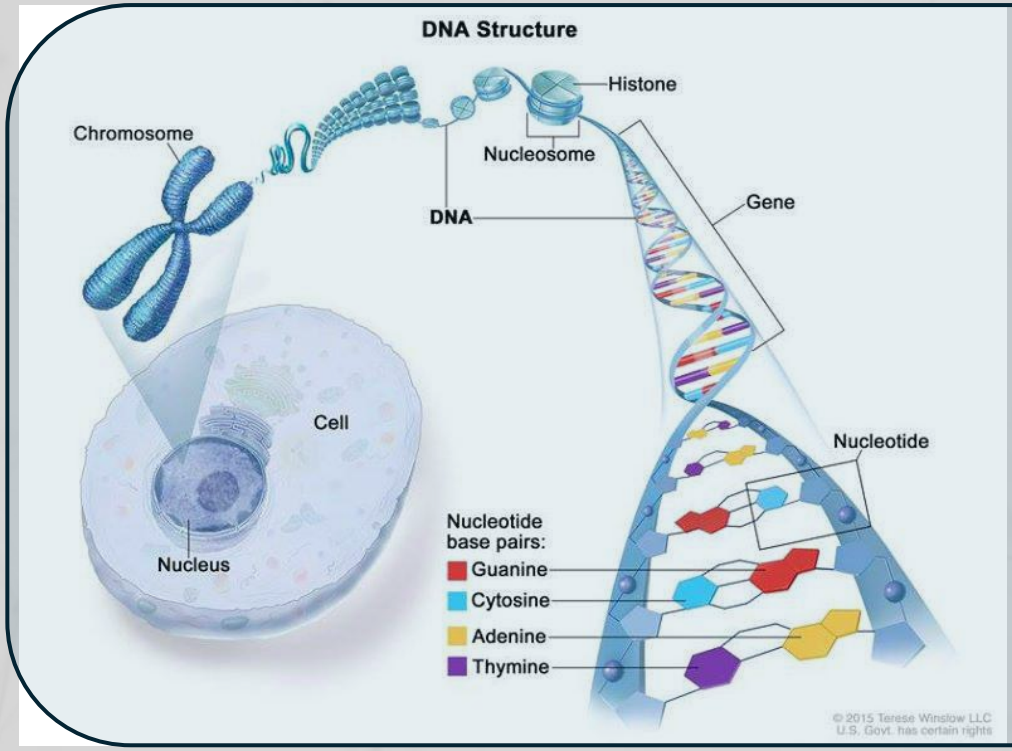


Image by brgfx on Freepik

CANCER

Cancer is a ***genetic disease***, i.e., it is caused by changes to genes (mutations)

Cancer is a leading cause of death worldwide (one-in-six deaths, 2020)



Each cell in our body contains 23 pairs of chromosomes

Each chromosome is a sequence of “base pairs” , bases are A, C, G, T

Gene: **subsequence of the chromosome** which has functional importance

~20,000 genes have been identified

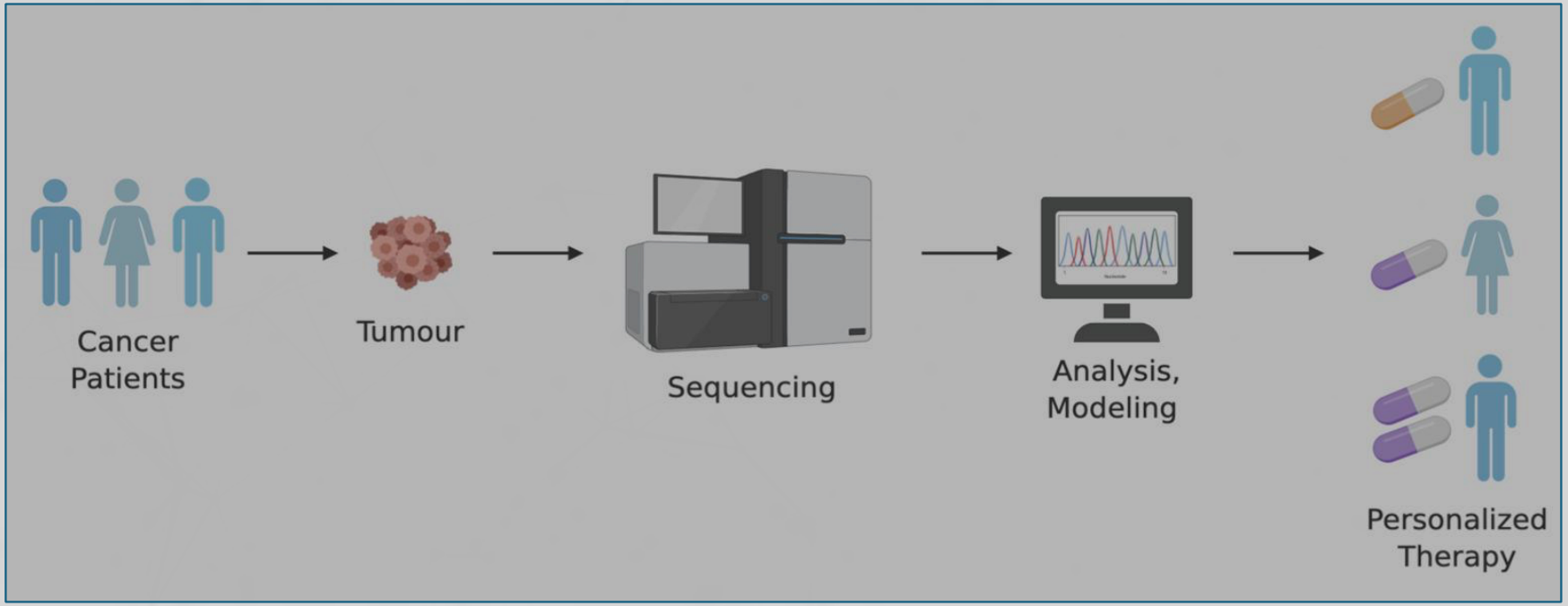
https://www.cancer.gov/about-cancer/understanding/what-is-cancer

https://www.who.int/news-room/fact-sheets/detail/cancer

CANCER TREATMENT

- Treatment remains challenging
  - Complex disease: ***Every cancer has an individual set of mutations***
  - A drug that works for one cancer patient, might have absolutely no effect on another

- Treatment must be tailored to each patient: **personalized therapy**



https://www.worldwidecancerresearch.org/news-opinion/2021/march/why-havent-we-cured-cancer-yet/
https://en.wikipedia.org/wiki/Personalized_medicine

Cancer patient

Genomic sequencing

Mutation 1
Mutation 2
...
Mutation N

Challenge in precision oncology

Clinical trials

Mutation 1

Mutation 2

Mutation 1
Mutation 2
...
Mutation N

Clinical trials intractable!

Icons from Flaticon

CANCER GENOMICS DATA

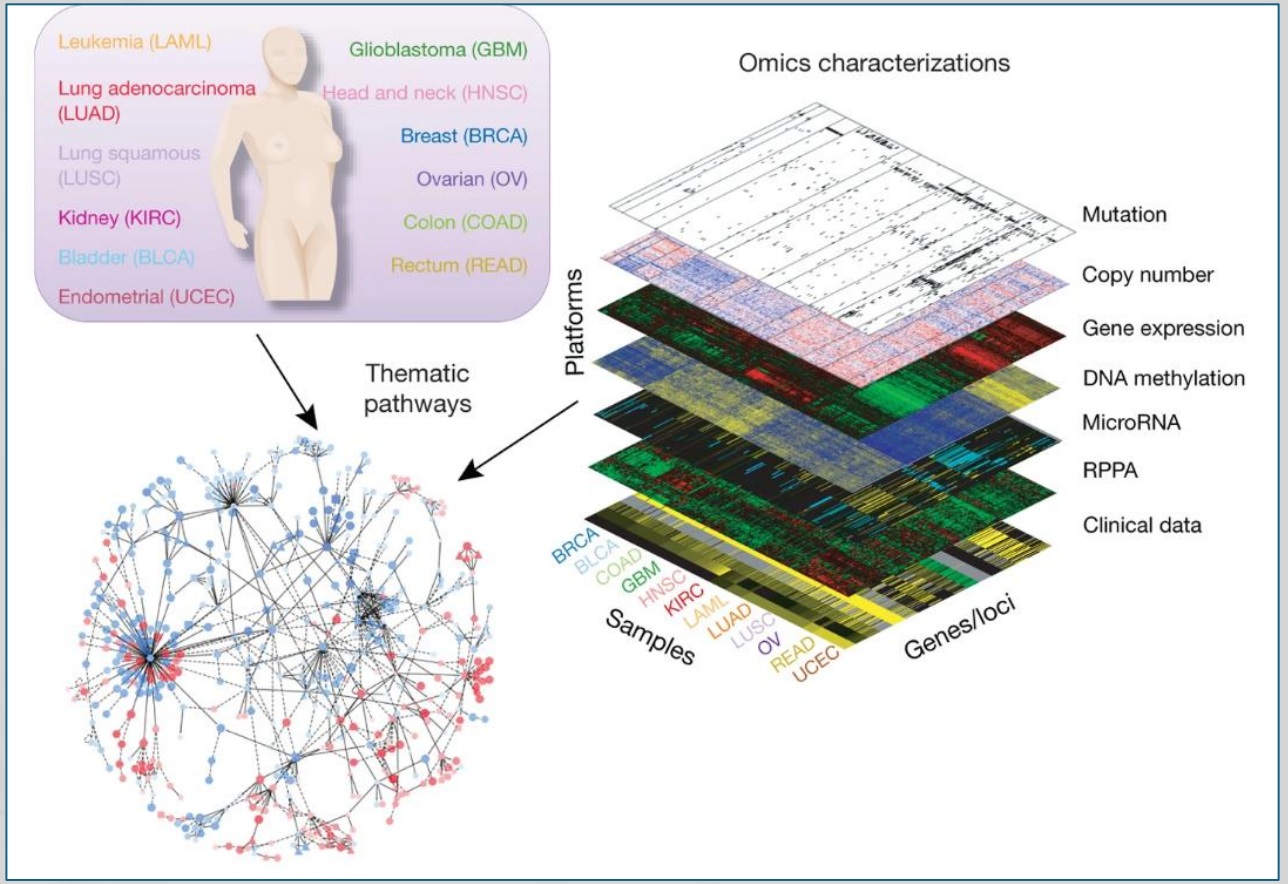
The Cancer Genome Atlas (TCGA)

Since 2006

> 11,000 patients

2.5 PetaBytes of Data

33 cancer types



- Many similar data collection efforts to understand cancer

Weinstein, J. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 2013

REPRESENTING GENOMIC DATA

Raw sequence (rarely used)

...ACCTTTCGGCCGGACCCCC...

Mutation Vector

Genes of interest:

G1	G2	G3	G4	G5	G6	G7	G8	G9	...
1	0	1	0	0	0	0	1	0	...

Binary indicator: 1 → mutation in gene, 0 → no mutation

Gene Expression Vector

Genes of interest:

G1	G2	G3	G4	G5	G6	G7	G8	G9	...
6	2.1	3	0	0	0	0	1	0	...

Count or real value: indicates activity level of gene

Sequence of Mutations

G1: R273C, G1: S1372L, G2: L145V ...

In gene G1, at location 273 a mutation changed R to C in the protein

DRUG RESPONSE MEASUREMENTS

1. Response Evaluation Criteria In Solid Tumors (RECIST)

- Standard way to measure how well a cancer patient responds to treatment.

Good response (label +1)

Bad response (label -1)

RECIST	
CR	Complete Response
PR	Partial Response
PD	Progressive Disease
SD	Stable Disease

2. Progression-free Survival (PFS)

- The length of time during and after the treatment (days/months/years), that a patient lives without the cancer getting worse.

DRUG RESPONSE PREDICTION (DRP)

One day…

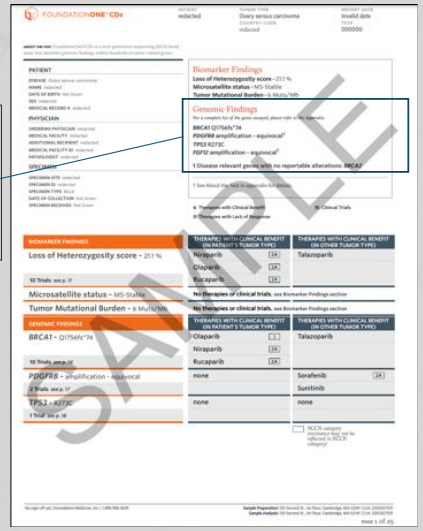
BRCA1 Q1756fs*74

PDGFRB amplification

TP53 R273C

FGF12 amplification

Mutations identified



Targeted therapy: “one gene, one target”

Majority of patients remain untreated by targeted therapies

Can AI help us identify the right drug for such cancer patients?

➢ **Given:**

- a patient’s genomic profile and
- a drug

➢ **Will the response of the patient to the drug be good?**

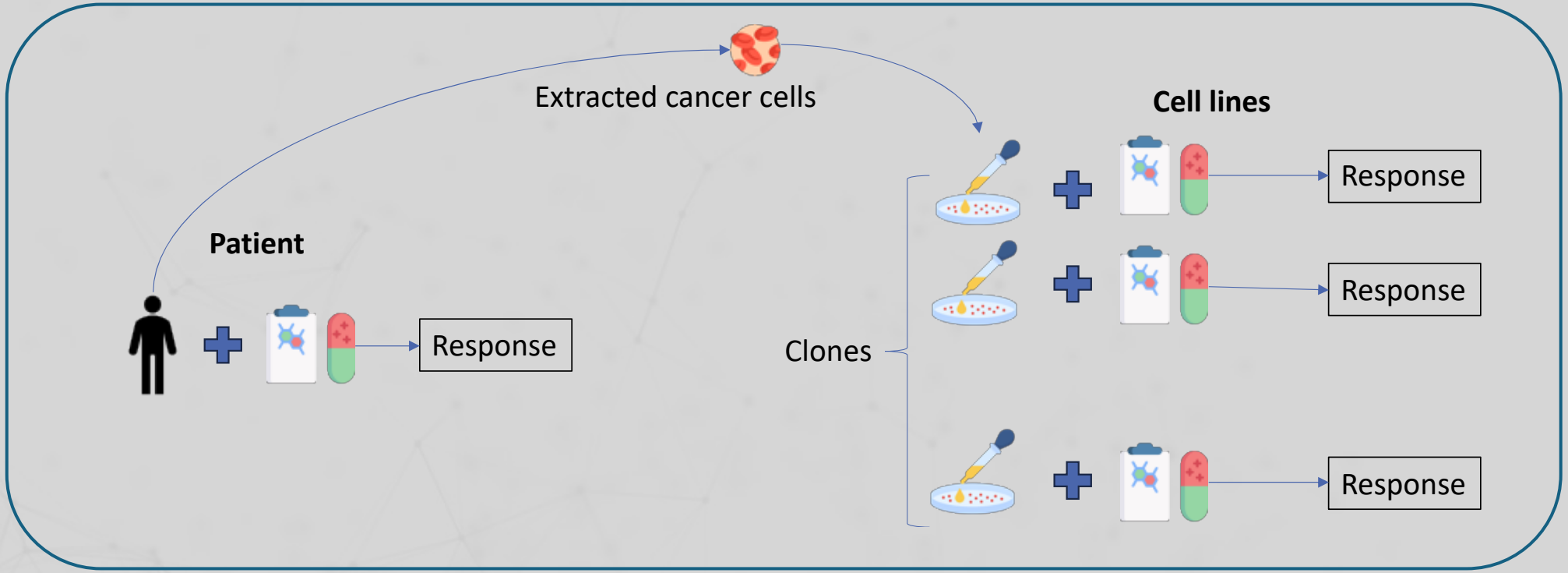
DRUG RESPONSE PREDICTION (DRP)

- X: Patient's genomic data (e.g., mutation vector or gene expression)
- Y: RECIST value after administering drug d

- $Y \sim f_d(X)$ → binary classification
- Challenge
  - X: abundant, but…
  - Y: extremely limited for any drug d

- Why?
  - Each patient is given one/few drugs, counterfactual unknown

CELL LINES: A RELATED "DOMAIN"



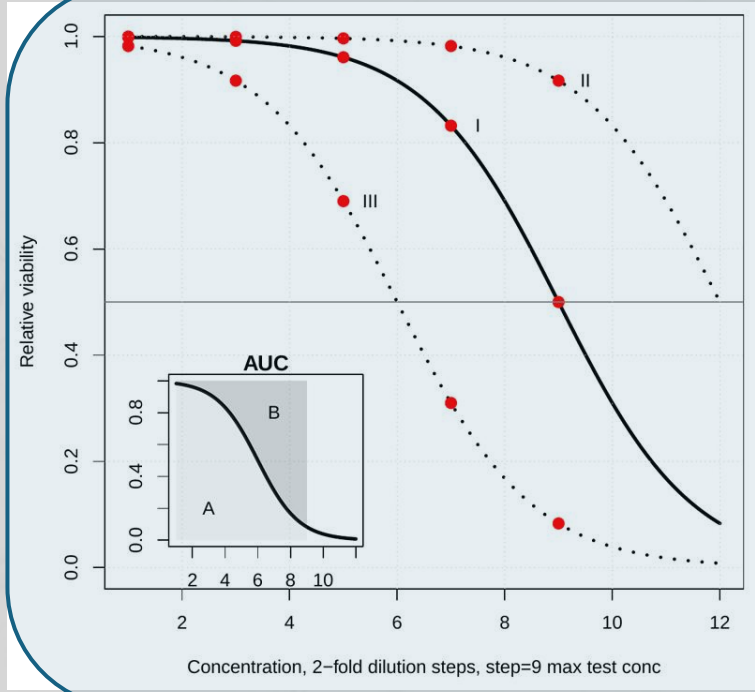
- Extract cancer cells and clone them in lab (living cells, continue growing)
  - Ensures each cell has same genomic data (X)

- Administer multiple drugs on cell lines, measure response Y

DRUG RESPONSE MEASUREMENT IN CELL LINES

Area under the Dose Response Curve (AUDRC)

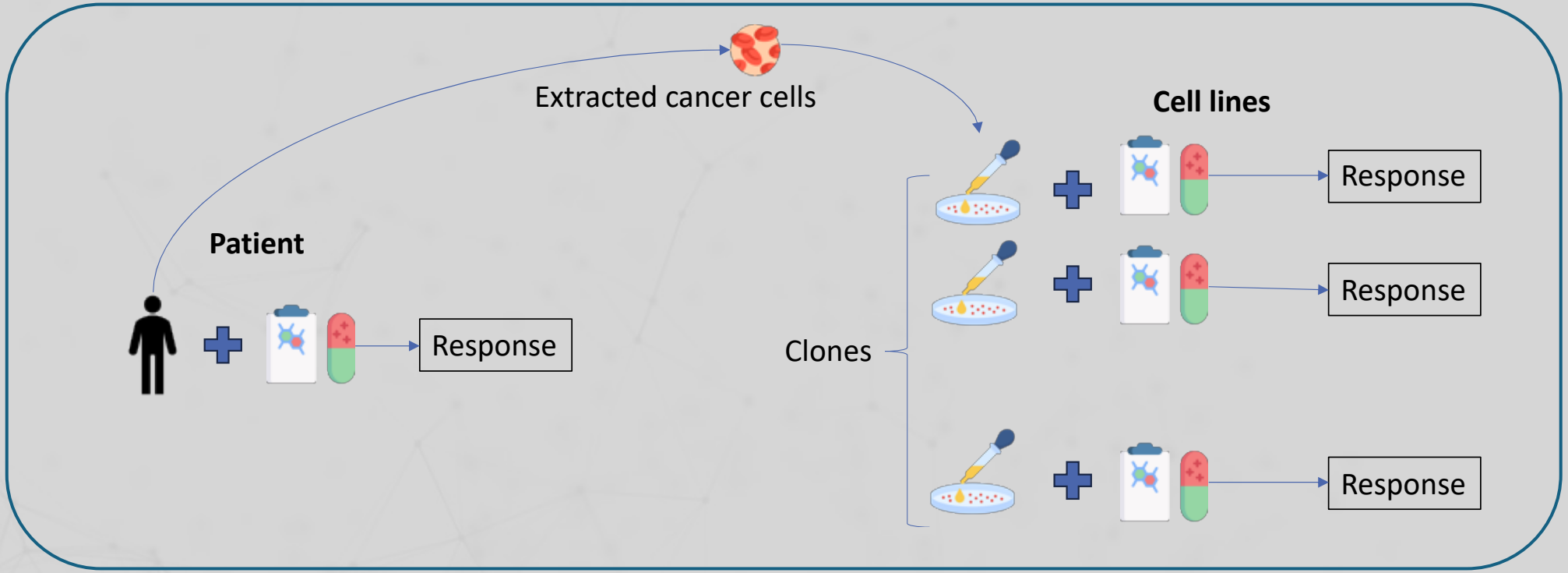
Real-valued [0,1]



- Administer progressively increasing concentration (X-axis) of drug and measure the amount of cancer cells (Y-axis) killed: Dose Response Curve (DRC)
- Lesser concentration kills more cells → more effective drug →Lower AUDRC
- E.g. efficacy of III > I > II

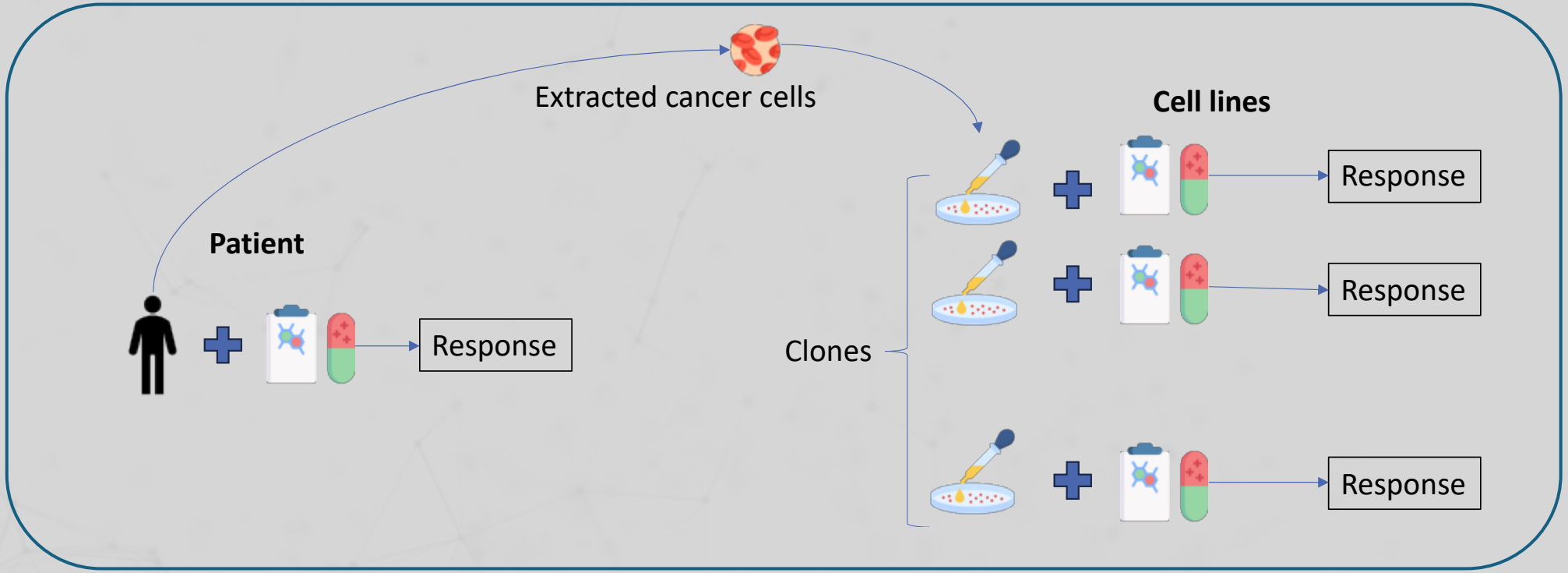
Vis, D. J. et al. Multilevel models improve precision and speed of IC50 estimates. *Pharmacogenomics 2016.*

CELL LINES: A RELATED "DOMAIN"



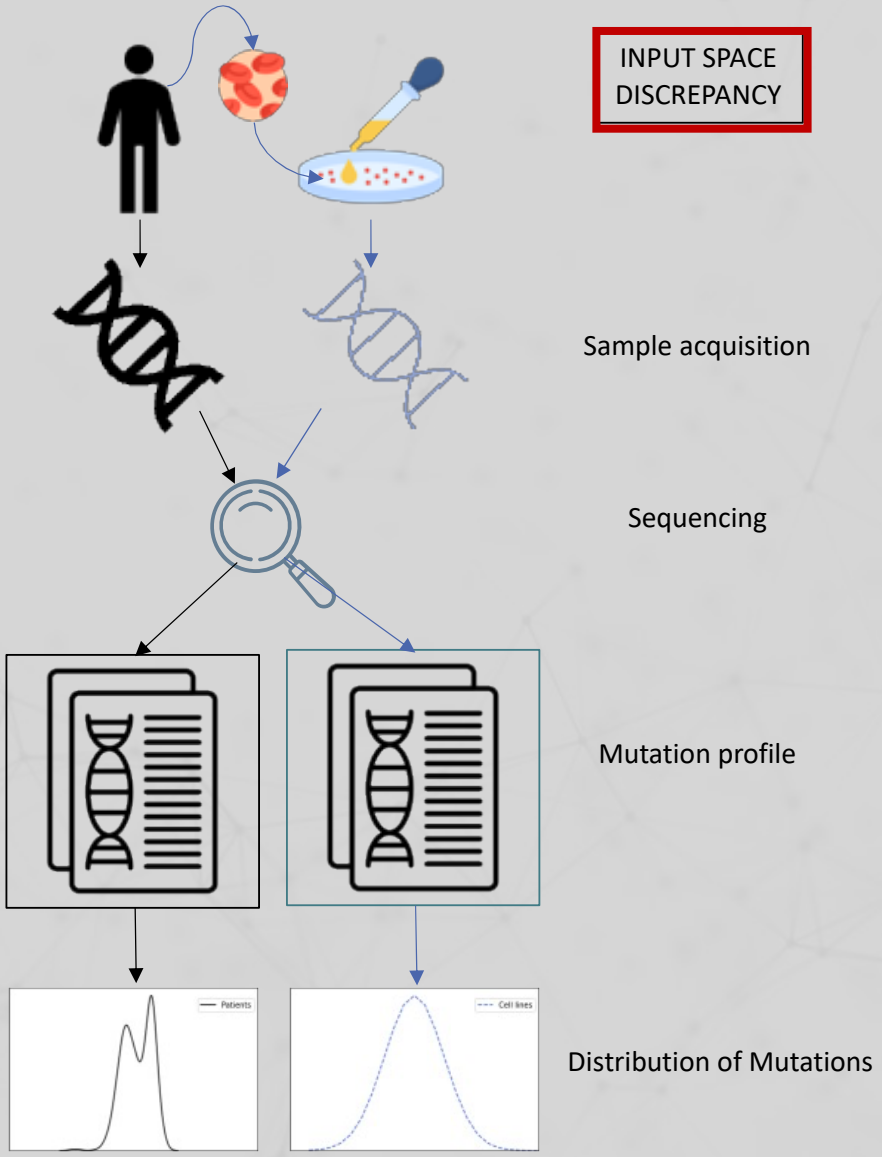
- Extract cancer cells and clone them in lab (living cells, continue growing)
  - Ensures each cell has same genomic data (X)
- Administer multiple drugs on cell lines, measure response Y
- Can a Drug Response Prediction model on cell lines $Y \sim f_d(X)$ work for patients?

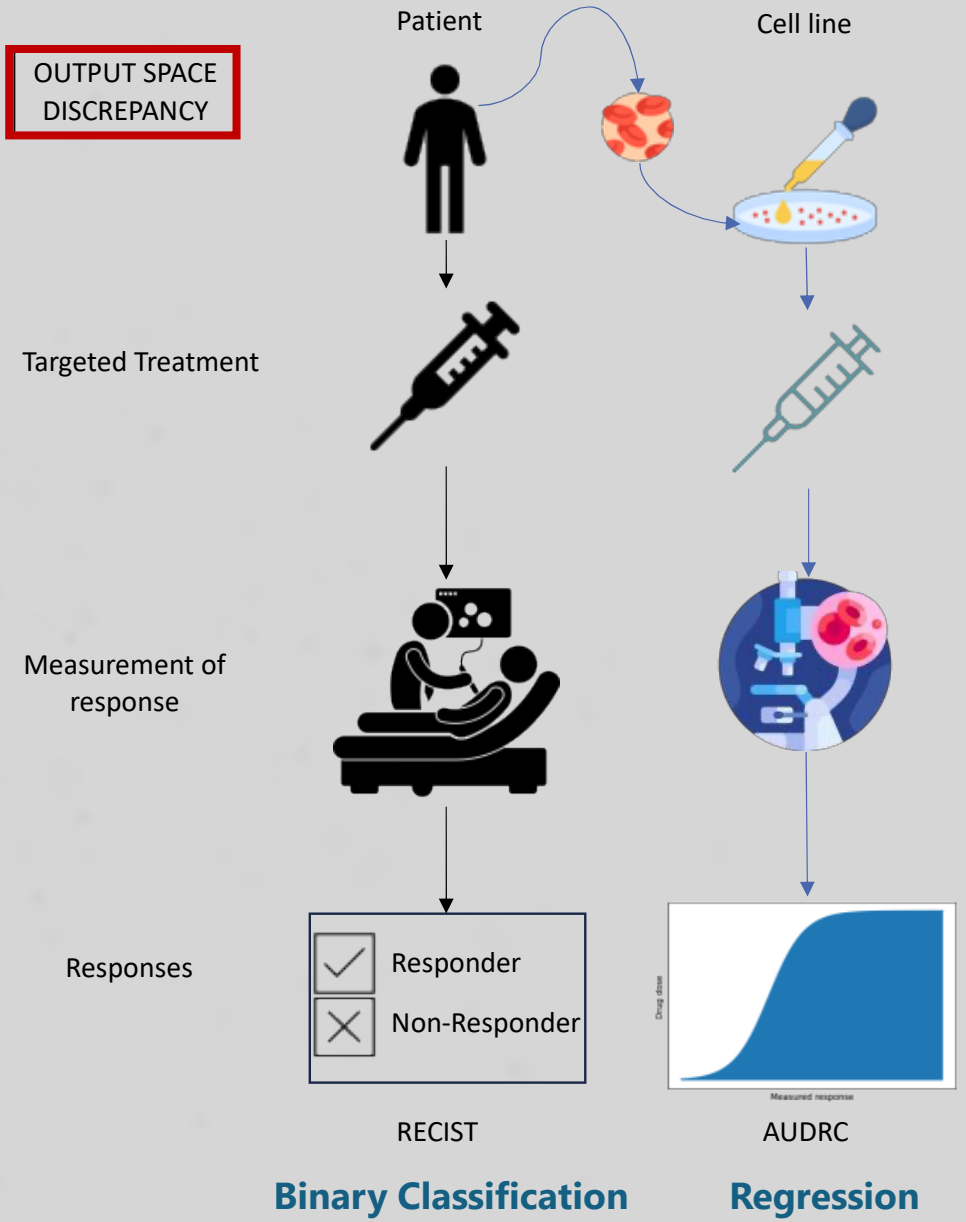
CELL LINES: A RELATED "DOMAIN"



- Extract cancer cells and clone them in lab (living cells, continue growing)
  - Ensures each cell has same genomic data (X)
- Administer multiple drugs on cell lines, measure response Y
- Can a Drug Response Prediction model on cell lines $Y \sim f_d(X)$ work for patients?

No: drug responses differ across patients and cell lines


INPUT SPACE DISCREPANCY
Sample acquisition
Sequencing
Mutation profile
Distribution of Mutations


OUTPUT SPACE DISCREPANCY
Patient
Cell line
Targeted Treatment
Measurement of response
Responses
Responder
Non-Responder
RECIST
AUDRC
Binary Classification
Regression

PROBLEM STATEMENT

- Given:

Domain	Genomic Profile	Drug Response	#samples
Cell Lines	X_c	$Y_c^d \in R$ (AADRC)	N_c labeled
Patients	X_t	$Y_t^d \in \{0,1\}$ (RECIST)	N_p labeled N_t unlabeled

$$N_p \ll N_c \ll N_t$$
$$P(X_c) \neq P(X_t)$$
$$f_c^d \neq f_t^d$$
where
$$Y_c^d \sim f_c^d(X_c), Y_t^d \sim f_t^d(X_t)$$

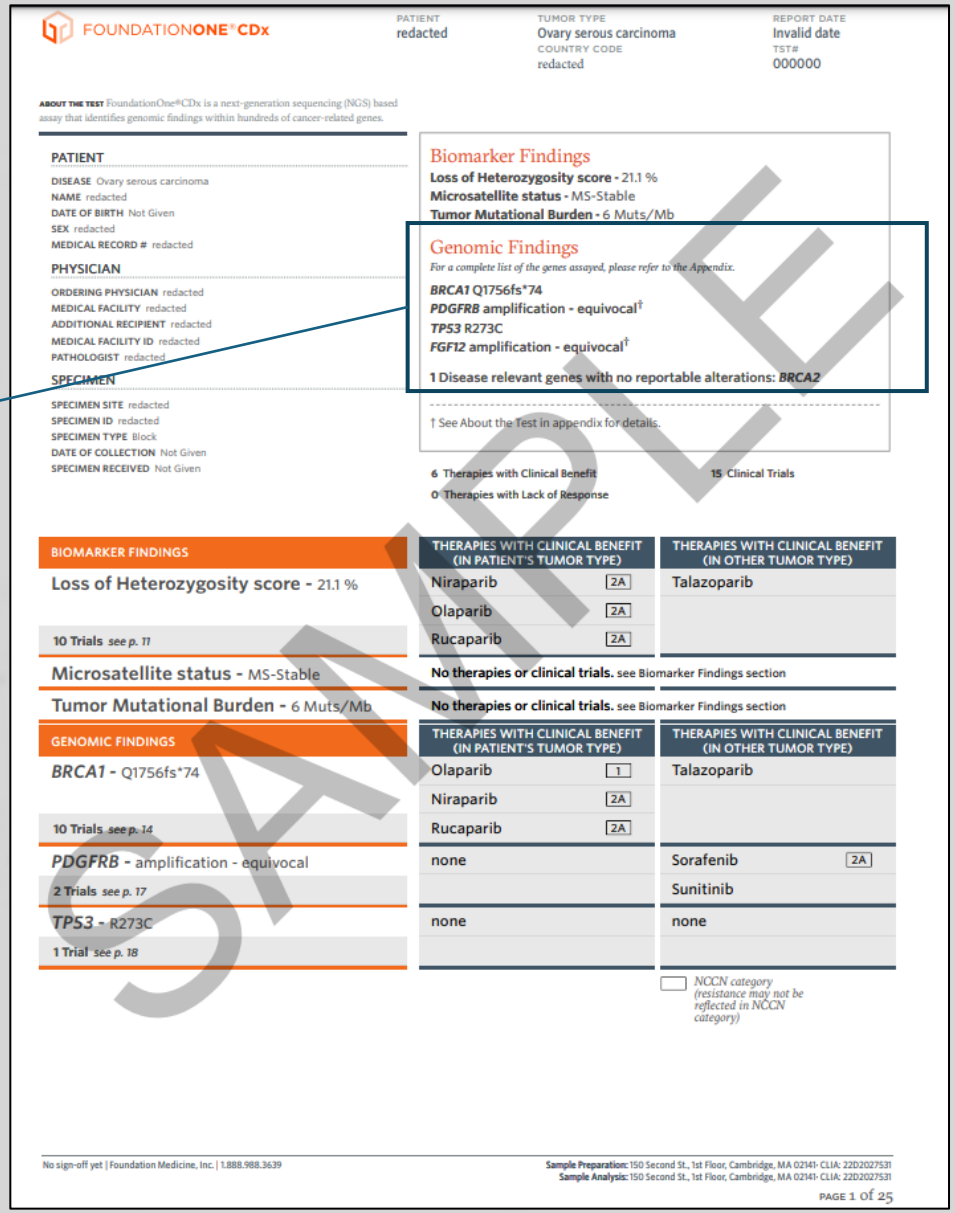
- Infer: Drug Response Prediction model

$$f_t^d: \; Y_t^d \sim f_t^d(X_t), \quad \forall \text{ drug } d \in \{d_1, d_2, \ldots d_n\}$$

One day....

- BRCA1 Q1756fs*74
- PDGFRB amplification
- TP53 R273C
- FGF12 amplification

Mutations identified



FOUNDATIONONE®CDx

PATIENT redacted | TUMOR TYPE Ovary serous carcinoma | COUNTRY CODE redacted | REPORT DATE Invalid date | TST# 000000

ABOUT THE TEST FoundationOne®CDx is a next-generation sequencing (NGS) based assay that identifies genomic findings within hundreds of cancer-related genes.

PATIENT
DISEASE Ovary serous carcinoma
NAME redacted
DATE OF BIRTH Not Given
SEX redacted
MEDICAL RECORD # redacted

PHYSICIAN
ORDERING PHYSICIAN redacted
MEDICAL FACILITY redacted
ADDITIONAL RECIPIENT redacted
MEDICAL FACILITY ID redacted
PATHOLOGIST redacted

SPECIMEN
SPECIMEN SITE redacted
SPECIMEN ID redacted
SPECIMEN TYPE Block
DATE OF COLLECTION Not Given
SPECIMEN RECEIVED Not Given

Biomarker Findings
Loss of Heterozygosity score - 21.1 %
Microsatellite status - MS-Stable
Tumor Mutational Burden - 6 Muts/Mb

Genomic Findings
For a complete list of the genes assayed, please refer to the Appendix.
BRCA1 Q1756fs*74
PDGFRB amplification - equivocal†
TP53 R273C
FGF12 amplification - equivocal†

1 Disease relevant genes with no reportable alterations: ***BRCA2***

† See About the Test in appendix for details.

6 Therapies with Clinical Benefit | 15 Clinical Trials
0 Therapies with Lack of Response

BIOMARKER FINDINGS	THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)	THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)
Loss of Heterozygosity score - 21.1 %	Niraparib 2A	Talazoparib
	Olaparib 2A	
10 Trials see p. 11	Rucaparib 2A	
Microsatellite status - MS-Stable	No therapies or clinical trials. see Biomarker Findings section	
Tumor Mutational Burden - 6 Muts/Mb	No therapies or clinical trials. see Biomarker Findings section	
GENOMIC FINDINGS	**THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)**	**THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)**
BRCA1 - Q1756fs*74	Olaparib 1	Talazoparib
	Niraparib 2A	
10 Trials see p. 14	Rucaparib 2A	
PDGFRB - amplification - equivocal	none	Sorafenib 2A
2 Trials see p. 17		Sunitinib
TP53 - R273C	none	none
1 Trial see p. 18		

NCCN category (resistance may not be reflected in NCCN category)

SAMPLE

No sign-off yet | Foundation Medicine, Inc. | 1.888.988.3639

Sample Preparation: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531
Sample Analysis: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531

PAGE 1 of 25

Can AI help us identify the right drug for this patient?

DO PREVIOUS METHODS WORK IN THE CLINIC?

- Input Data type
  - Gene expression data
    - Assumed as inputs in previous methods
    - Not measured in FDA approved clinical panels
  - Mutation data
    - Very sparse (typically a patient has ~10 mutations in panel out of ~million possible)
    - Previous methods do not perform well with such inputs
- Drug Repurposing
  - Use of a drug for one cancer in another cancer
  - Need predictions on drugs not in training set

Genes of interest:

G1	G2	G3	G4	G5	G6	G7	G8	G9	…
6	2.1	3	0	0	0	0	1	0	…

Count or real value: indicates activity level of gene

G1: R273C, G1: S1372L, G2: L145V …

In gene G1, at location 273 a mutation changed R to C in the protein

DRP Requirements

For clinical translation

- Training with mutations available in clinical sequencing reports
- Predict on drugs unseen during training
- Model varying length mutations
- Utilise all available auxiliary patient response information (PFS)

From prior DRP literature

- Handle input discrepancy
- Handle output discrepancy
- Model patient mutation heterogeneity

Model Design

Model varying length mutations	Genes and mutations to be tokenized	Use transformers on gene and mutation level tokens

Model Design

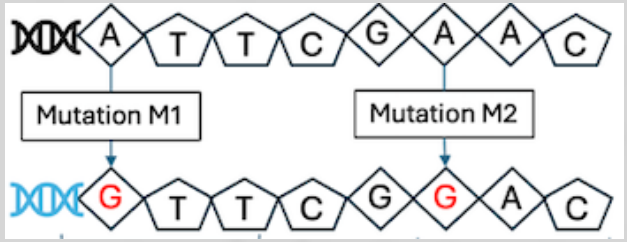
Model varying length mutations	Genes and mutations to be tokenized	Use transformers on gene and mutation level tokens
Use all patient response-related information	Model survival information (PFS)	Pretraining transformers to predict survival

Model Training

- Use all patient response-related information (PFS)
- Handles varying length mutation data
- Training on mutations

- Handles input discrepancy
- Handles output discrepancy
- Allows drug repurposing

CLINICAL DIAGNOSTIC REPORT



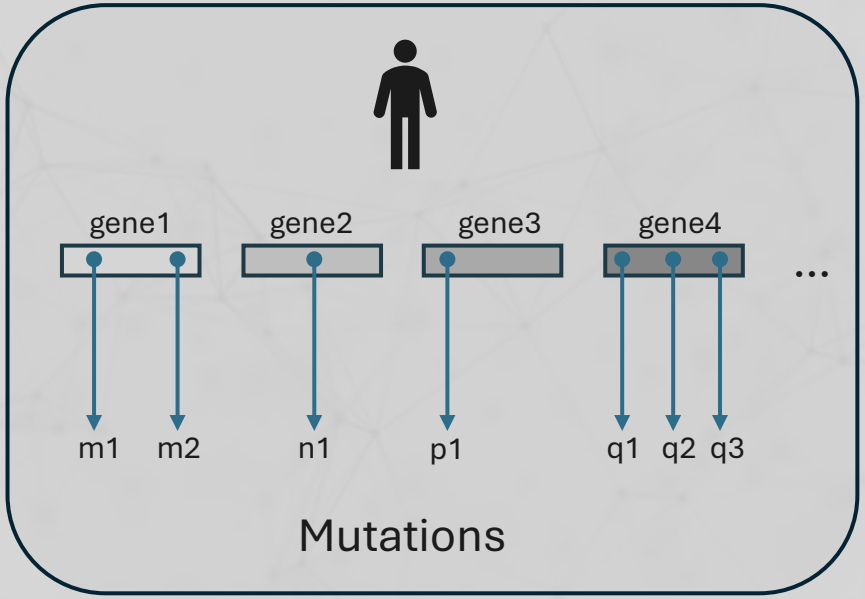
BRCA1 Q1756fs*74

PDGFRB amplification

TP53 R273C

FGF12 amplification

Mutations identified



FOUNDATIONONE®CDx

PATIENT redacted

TUMOR TYPE Ovary serous carcinoma
COUNTRY CODE redacted

REPORT DATE Invalid date
TST# 000000

ABOUT THE TEST FoundationOne®CDx is a next-generation sequencing (NGS) based assay that identifies genomic findings within hundreds of cancer-related genes.

PATIENT
DISEASE Ovary serous carcinoma
NAME redacted
DATE OF BIRTH Not Given
SEX redacted
MEDICAL RECORD # redacted

PHYSICIAN
ORDERING PHYSICIAN redacted
MEDICAL FACILITY redacted
ADDITIONAL RECIPIENT redacted
MEDICAL FACILITY ID redacted
PATHOLOGIST redacted

SPECIMEN
SPECIMEN SITE redacted
SPECIMEN ID redacted
SPECIMEN TYPE Block
DATE OF COLLECTION Not Given
SPECIMEN RECEIVED Not Given

Biomarker Findings
Loss of Heterozygosity score - 21.1 %
Microsatellite status - MS-Stable
Tumor Mutational Burden - 6 Muts/Mb

Genomic Findings
For a complete list of the genes assayed, please refer to the Appendix.
BRCA1 Q1756fs*74
PDGFRB amplification - equivocal†
TP53 R273C
FGF12 amplification - equivocal†

1 Disease relevant genes with no reportable alterations: ***BRCA2***

† See About the Test in appendix for details.

6 Therapies with Clinical Benefit
0 Therapies with Lack of Response
15 Clinical Trials

BIOMARKER FINDINGS	THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)	THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)
Loss of Heterozygosity score - 21.1 %	Niraparib 2A Olaparib 2A Rucaparib 2A	Talazoparib
10 Trials see p. 11		
Microsatellite status - MS-Stable	No therapies or clinical trials. see Biomarker Findings section	
Tumor Mutational Burden - 6 Muts/Mb	No therapies or clinical trials. see Biomarker Findings section	
GENOMIC FINDINGS	**THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)**	**THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)**
BRCA1 - Q1756fs*74	Olaparib 1 Niraparib 2A Rucaparib 2A	Talazoparib
10 Trials see p. 14		
PDGFRB - amplification - equivocal	none	Sorafenib 2A Sunitinib
2 Trials see p. 17		
TP53 - R273C	none	none
1 Trial see p. 18		

NCCN category (resistance may not be reflected in NCCN category)

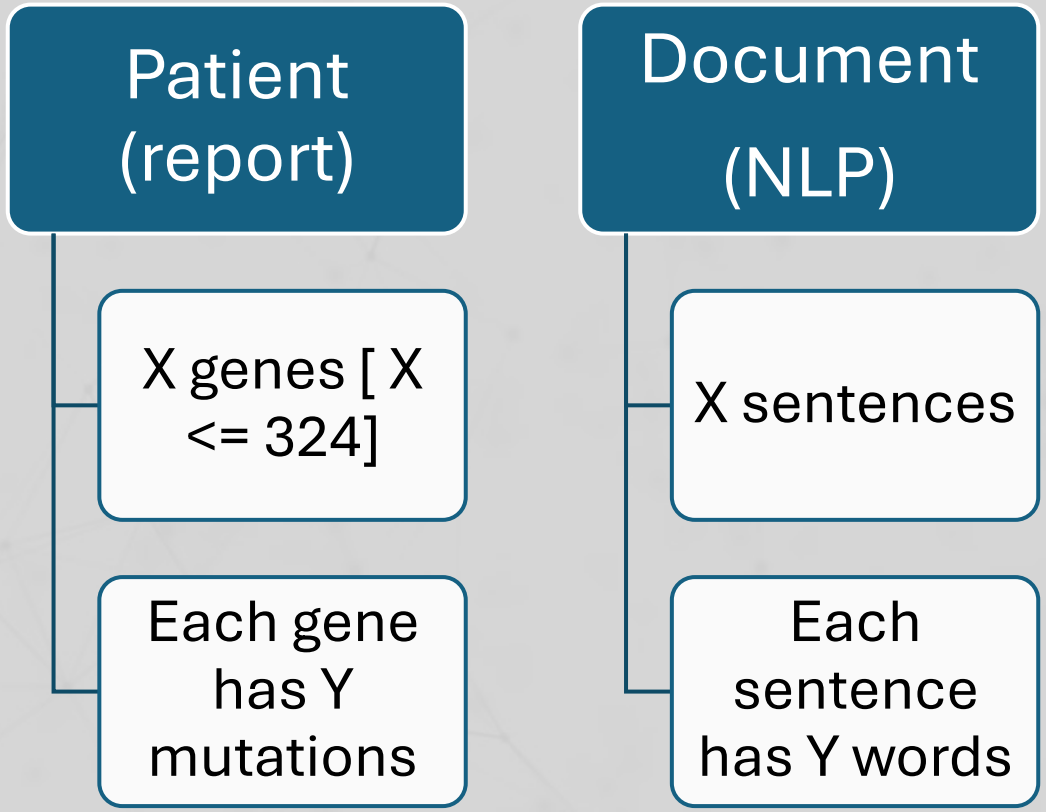
SAMPLE

No sign-off yet | Foundation Medicine, Inc. | 1.888.988.3639

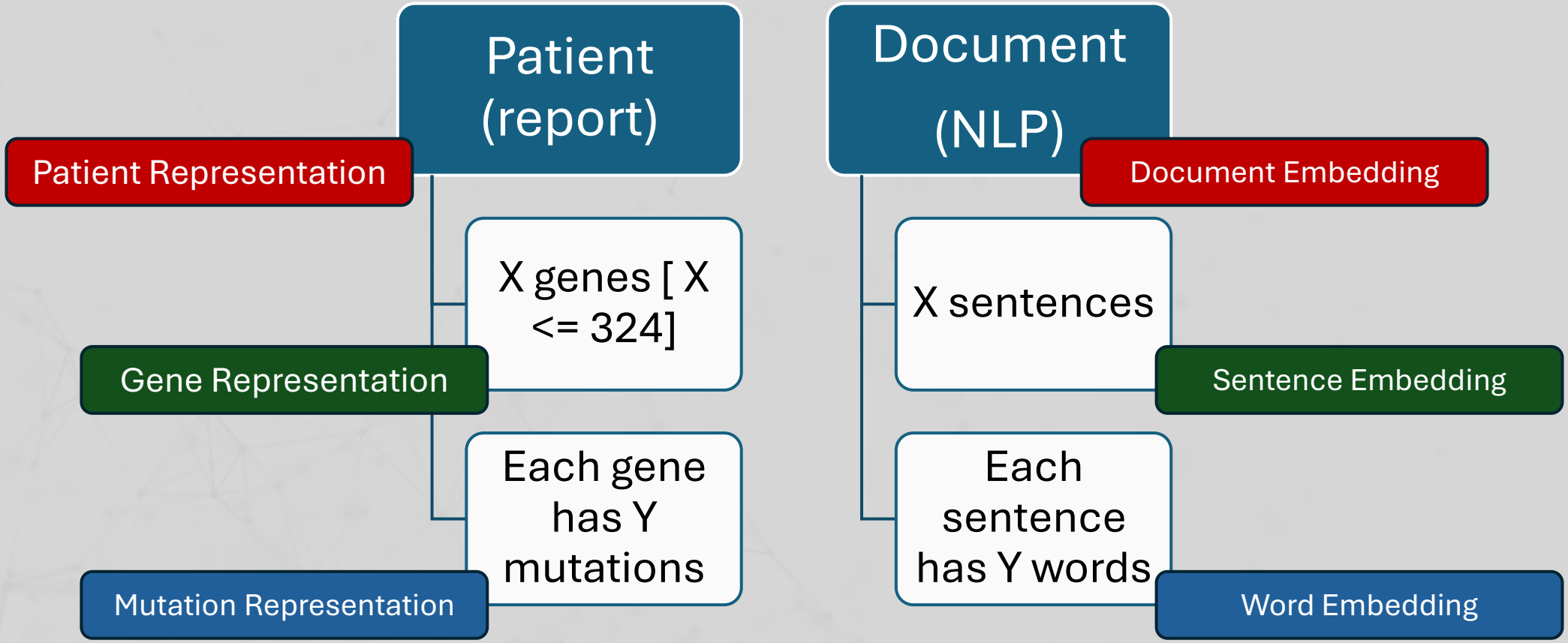
Sample Preparation: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531
Sample Analysis: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531

PAGE 1 of 25

DISCRETE SEQUENTIAL DATA



DISCRETE SEQUENTIAL DATA



OUR APPROACH

1. Capture functional effects of mutations

- Features indicating harmfulness of each mutation

2. Transformers for sequential inputs

- Tokenization at gene and mutation level

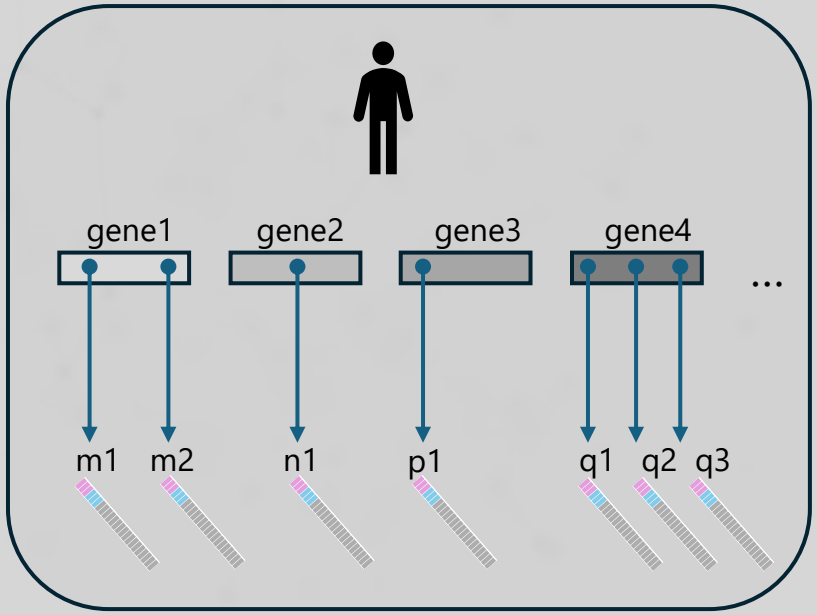
3. Survival data for Supervised Representation Learning

- Pretraining and joint multi-task learning

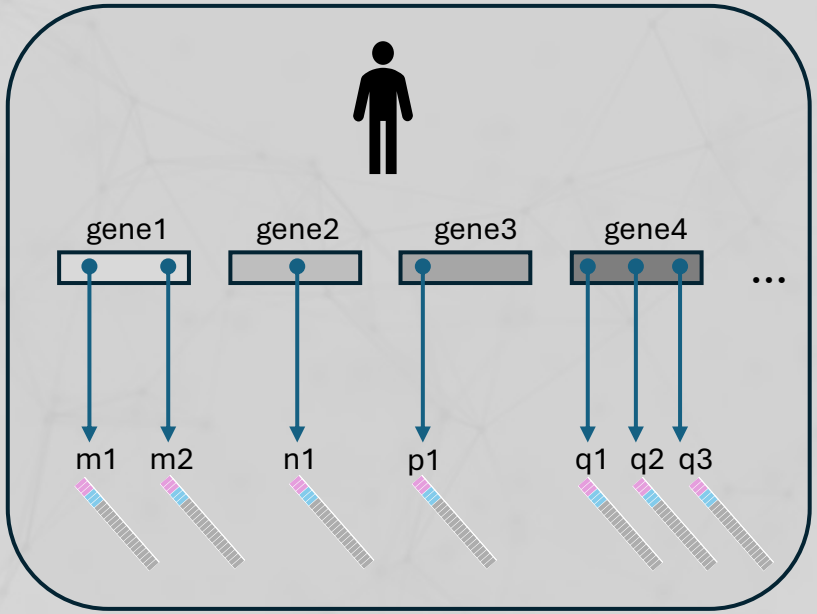
FUNCTIONAL ANNOTATIONS

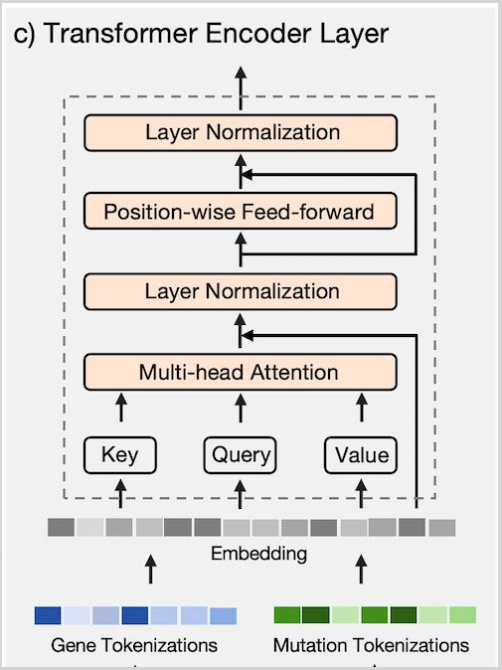
- **Clinvar:** 3 indicators pathogenic, benign, significance unknown
- **GPD:** 3 location indicators protein information unit, linker unit, non-coding unit
- **Annovar:** Predictions from 17 algorithms indicating deleteriousness (harmful/not)

23-dimensional feature vector per mutation

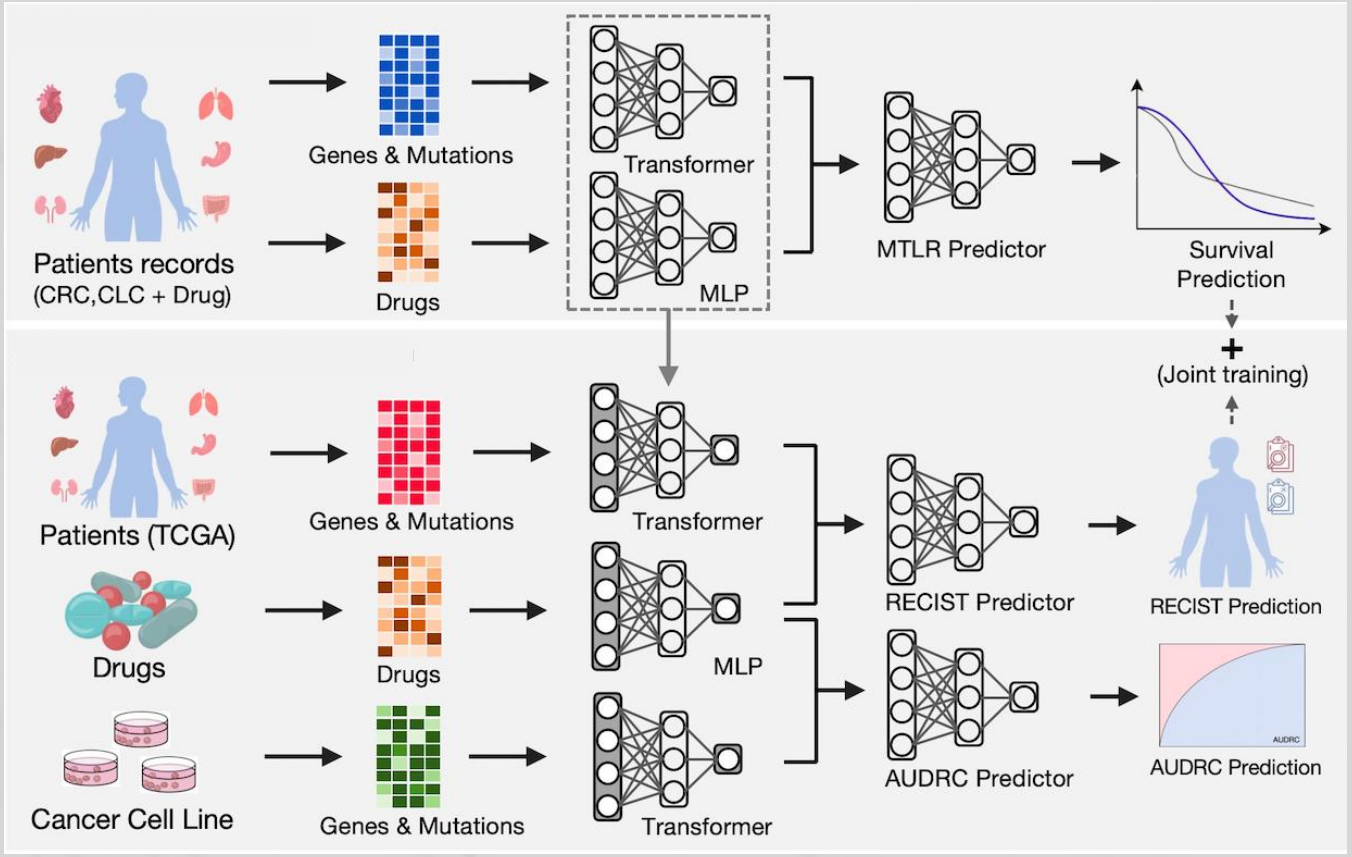


TRANSFORMERS FOR SEQUENTIAL INPUTS





MULTITASK LEARNING



Pretrain

- Train Neural MTLR
- Survival Data only

Joint Training

- Train Survival, Cell Line Response, Patient Response
- Survival, Cell Line, Patient Data

MULTITASK LOGISTIC REGRESSION

For each patient: $y = (y_1, y_2, y_3, y_4 \ \ldots y_m)$

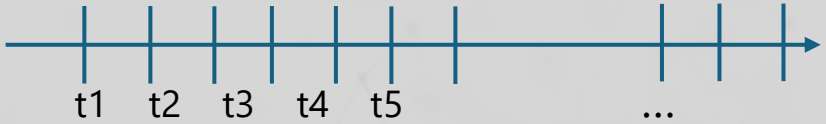
$y_i \in \{0,1\}$: survival status at time t_i

Survival time s:

$y_i = 0$ (no death) for all i with $t_i < s$

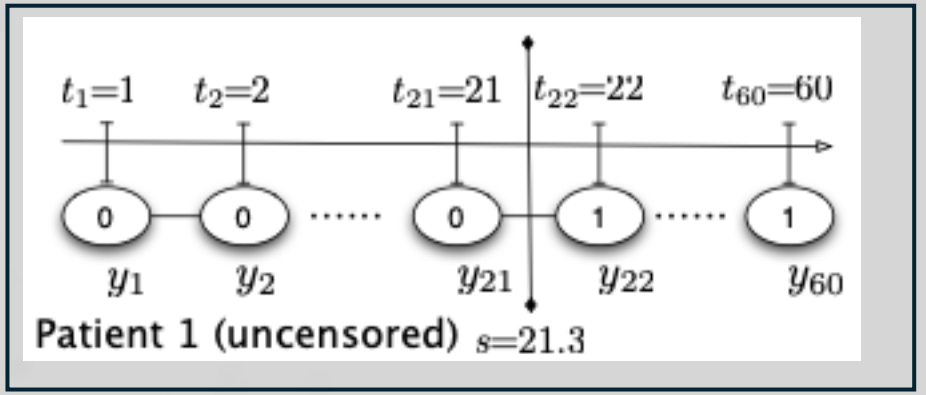
$y_i = 1$ (death) for all i with $t_i \geq s$

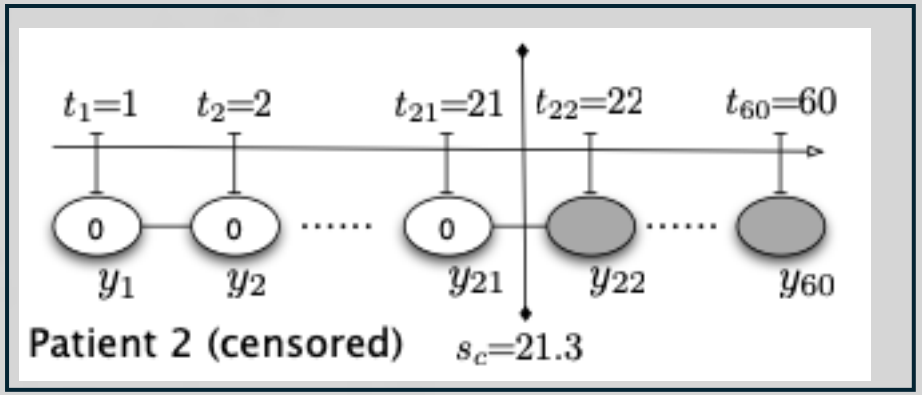
Survival Time:



$(y_1, y_2, y_3, y_4 \ \ldots y_m)$

Logistic Regression

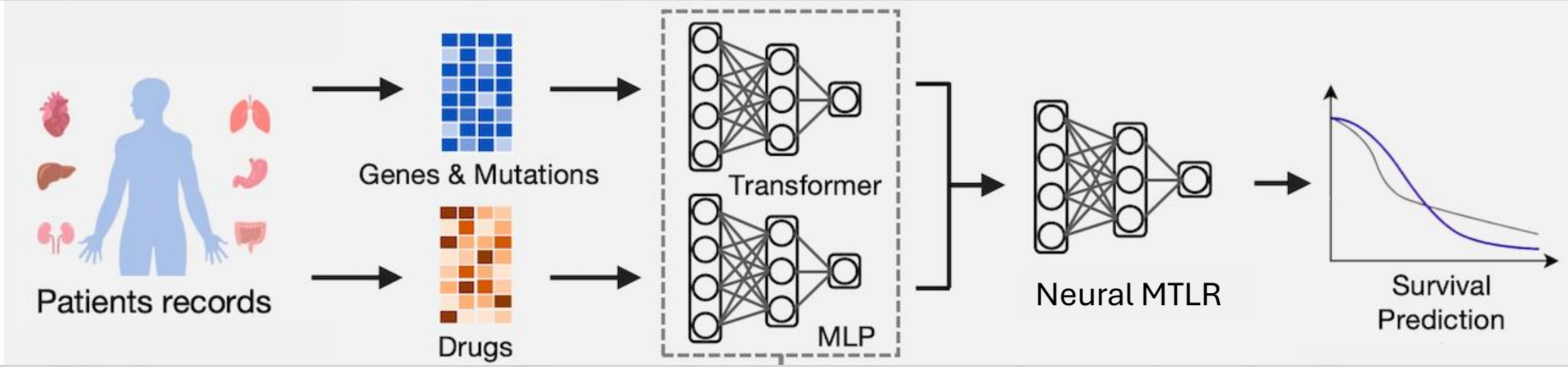




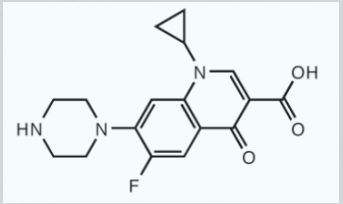
Marginalize over unobserved variables (EM)

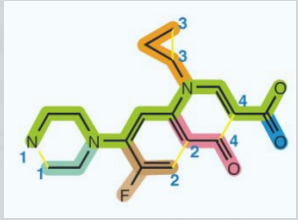
Yu, Chun-Nam, et al. "Learning patient-specific cancer survival distributions as a sequence of dependent regressors." *NIPS* 2011.

PRETRAINING WITH SURVIVAL PREDICTION



Drug Embedding





SMILES String

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

00110011000101000111100011

Morgan fingerprint

https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

EXPERIMENTS

Domain	Genomic Profile	Drug Response (RECIST)	Survival (PFS)	#(sample, drug) pairs	Drugs tested
Cell Lines	324 genes from FoundationOne panel Mutation sequences	689		3632	
Patients		470	2512	15732	drugs with RECIST labels in > 80 patients

- Evaluation only on patients: 3 random 80-20 splits of (sample, drug) pairs
- Binary classification task: RECIST category prediction
  - Metrics: AUROC and AUPRC

EXPERIMENTAL RESULTS

	5-Fluorouracil		Cisplatin		Paclitaxel		Overall	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
CODE-AE	61.77(±11.63)	88.02(±4.85)	38.09(±16.34)	71.88(±8.46)	44.83(±15.08)	74.68(±5.11)	47.15(±7.82)	73.74(±6.90)
TCRP	48.39(±11.96)	77.87(±9.49)	50.86(±16.75)	79.60(±10.73)	60.66(±24.22)	78.69(±16.73)	47.36(±3.26)	75.70(±4.03)
TUGDA	58.39(±16.94)	83.40(±7.69)	37.47(±4.36)	72.07(±3.75)	41.13(±18.28)	71.67(±1.03)	46.18(±3.18)	75.42(±2.22)
Velodrome	50.91(±19.54)	79.08(±15.04)	48.61(±9.73)	78.80(±7.97)	63.26(±26.45)	80.02(±16.34)	52.56(±4.93)	77.62(±0.48)
DruID	64.73(±8.73)	85.55(±6.43)	67.38(±10.63)	86.30(±7.35)	**63.43(±4.97)**	**82.55(±6.83)**	62.36(±3.60)	82.06(±5.02)
PREDICT-AI	**71.30(±3.87)**	**88.74(±2.82)**	**72.37(±10.10)**	**90.26** (±4.68)	62.19(±9.42)	81.08(±8.02)	**64.96(±4.50)**	**84.85(±4.02)**

EXPERIMENTAL RESULTS

	5-Fluorouracil		Cisplatin		Paclitaxel		Overall	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
CODE-AE	61.77(±11.63)	88.02(±4.85)	38.09(±16.34)	71.88(±8.46)	44.83(±15.08)	74.68(±5.11)	47.15(±7.82)	73.74(±6.90)
TCRP	48.39(±11.96)	77.87(±9.49)	50.86(±16.75)	79.60(±10.73)	60.66(±24.22)	78.69(±16.73)	47.36(±3.26)	75.70(±4.03)
TUGDA	58.39(±16.94)	83.40(±7.69)	37.47(±4.36)	72.07(±3.75)	41.13(±18.28)	71.67(±1.03)	46.18(±3.18)	75.42(±2.22)
Velodrome	50.91(±19.54)	79.08(±15.04)	48.61(±9.73)	78.80(±7.97)	63.26(±26.45)	80.02(±16.34)	52.56(±4.93)	77.62(±0.48)
DruID	64.73(±8.73)	85.55(±6.43)	67.38(±10.63)	86.30(±7.35)	**63.43(±4.97)**	**82.55(±6.83)**	62.36(±3.60)	82.06(±5.02)
PREDICT-AI	**71.30(±3.87)**	**88.74(±2.82)**	**72.37(±10.10)**	**90.26** (±4.68)	62.19(±9.42)	81.08(±8.02)	**64.96(±4.50)**	**84.85(±4.02)**

More details/results: https://arxiv.org/abs/2402.10551

Aishwarya Jayagopal, Hansheng Xue, Ziyang He, Robert J Walsh, Krishna Kumar H,
David SP Tan, Tuan Z Tan, Jason J Pitt, Anand D Jeyasekharan, Vaibhav Rajan

Personalised Drug Identifier for Cancer Treatment with Transformers using Auxiliary Information

KDD 2024

One Last Requirement...

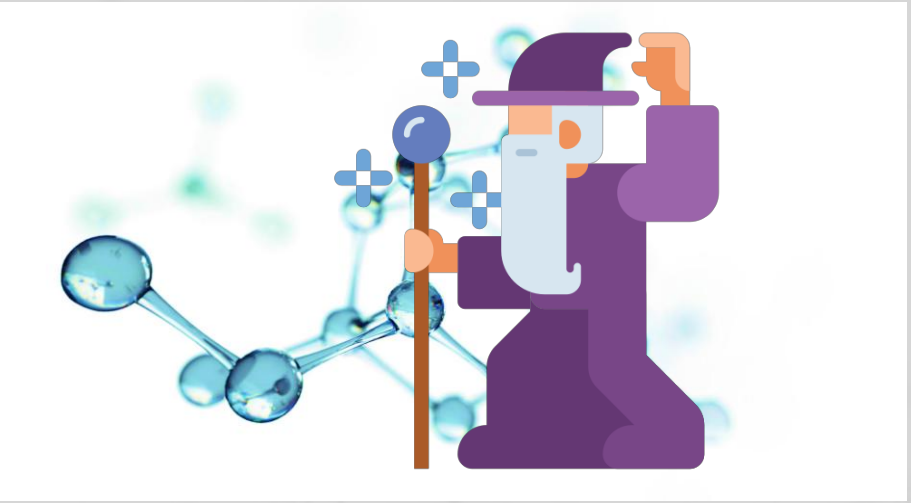
DRP Requirements

For clinical translation

- Training with mutations available in clinical sequencing reports
- Predict on drugs unseen during training
- Model varying length mutations
- Utilise all available auxiliary patient response information (PFS)

From prior DRP literature

- Handle input discrepancy
- Handle output discrepancy
- Model patient mutation heterogeneity



GANDALF: Generative AtteNtion based Data Augmentation and predictive modeLing Framework

ICLR 2025

Jayagopal, A., Zhang, Y., Walsh, R.J., Tan, T.Z., Jeyasekharan, A.D. and Rajan, V., 2025. GANDALF: Generative AttentioN based Data Augmentation and predictive modeLing Framework for personalized cancer treatment.

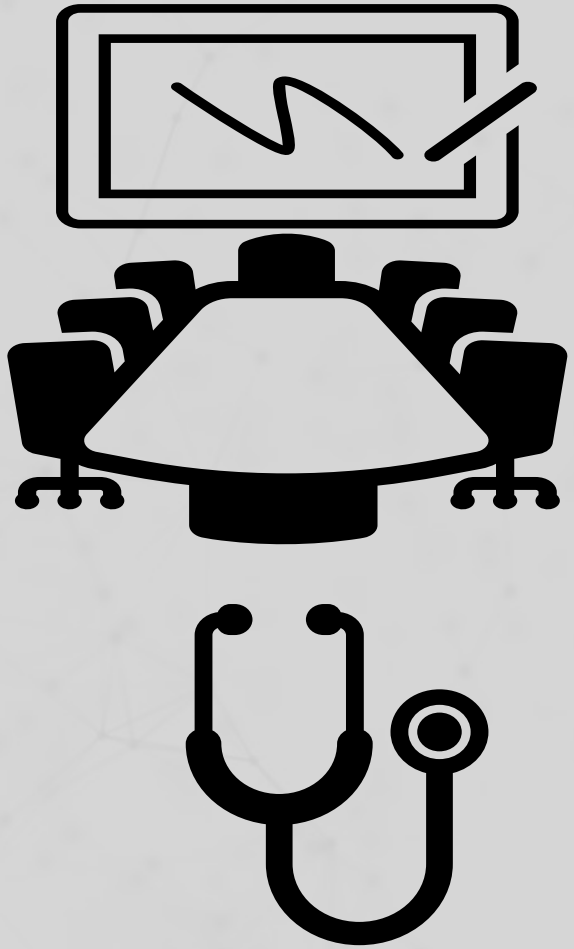
Model Design

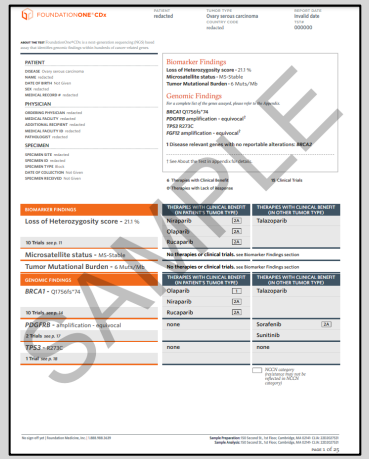
Requirement	Considerations	Design Choice
Model patient mutation heterogeneity	Generate more patient-like data, from cell line profiles	Generate samples from cell lines, using diffusion models

Treatment Recommendation System

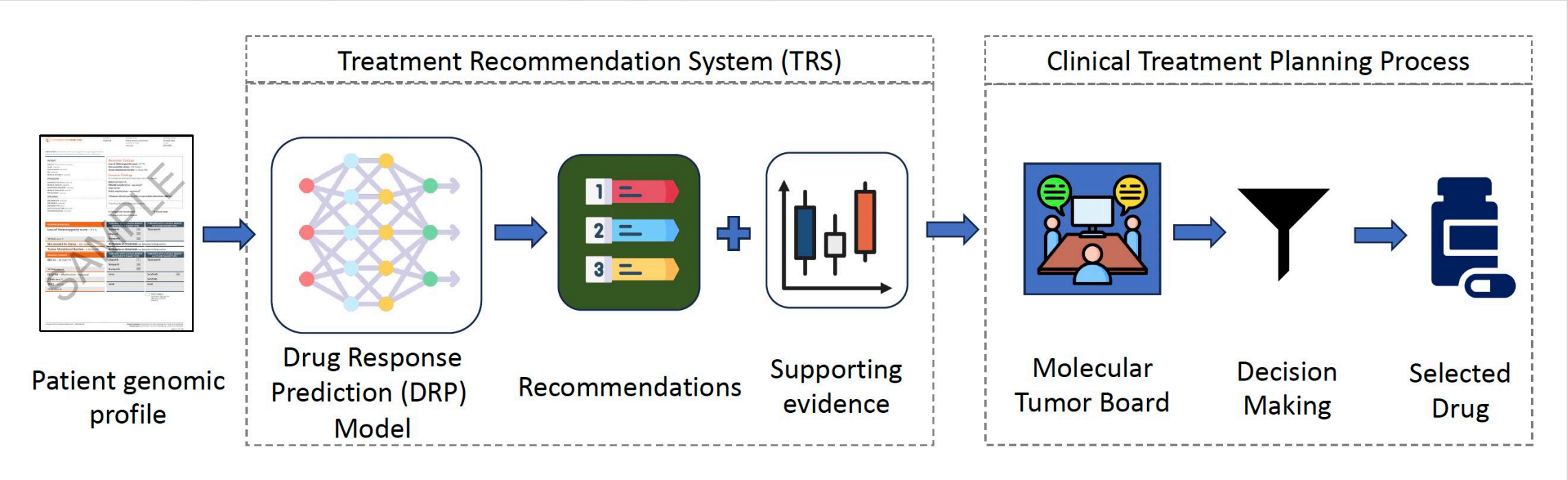
MOLECULAR TUMOR BOARD (MTB)

- Treatment planning for complex cancer cases is increasingly done by a Molecular Tumor Board
- Several expert clinicians collectively decide on the most suitable treatment

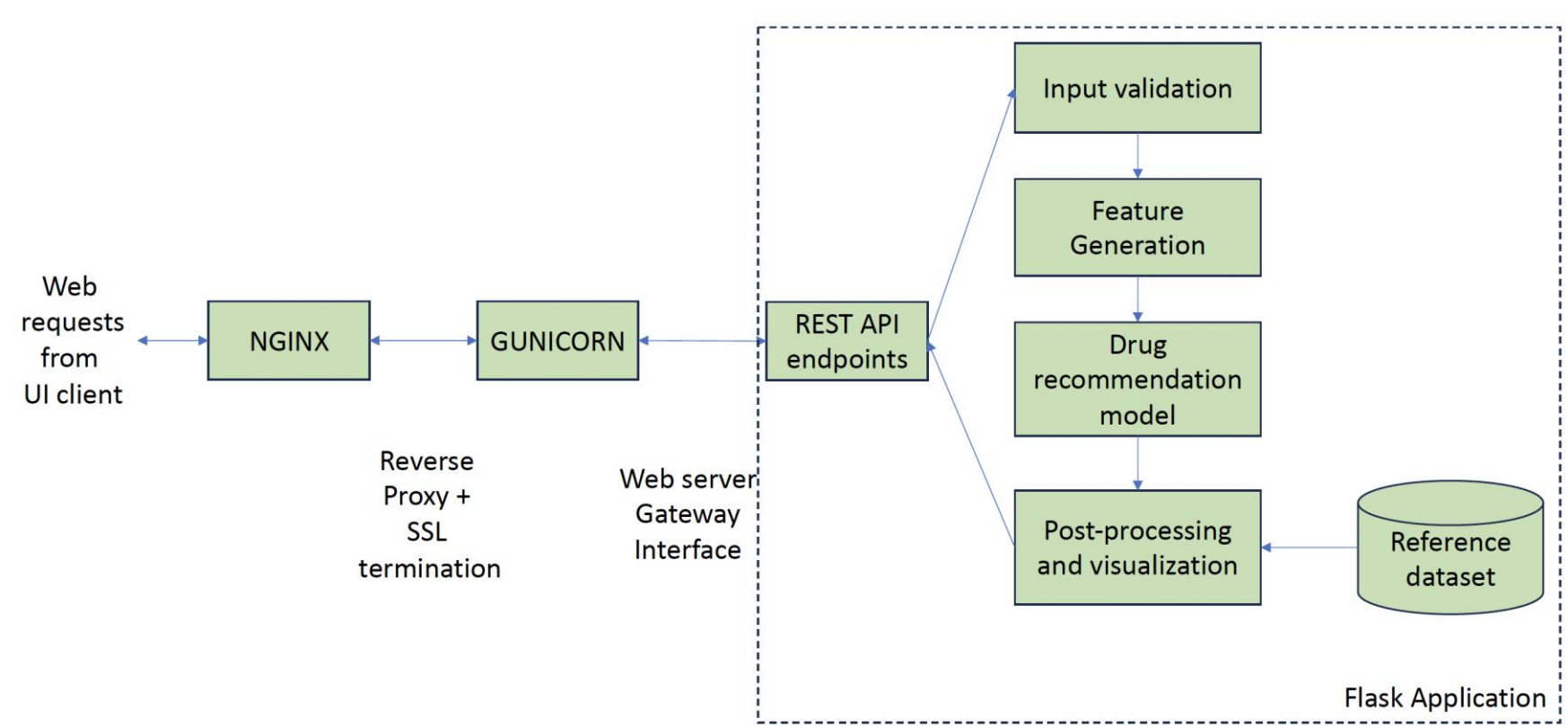




AI-IN-THE-LOOP MTB



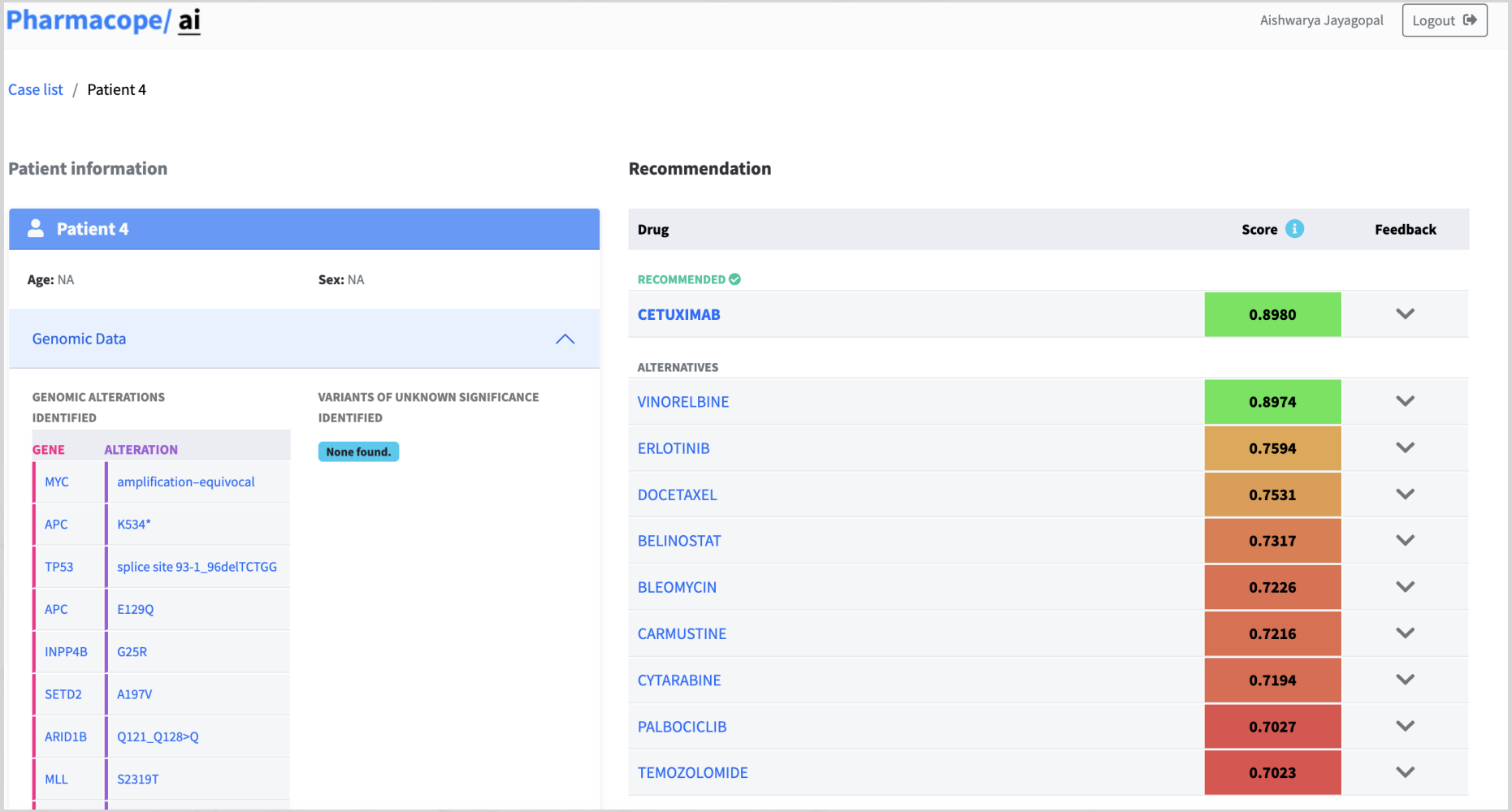
TREATMENT RECOMMENDATION SYSTEM @ NUH, SINGAPORE



Deployed on Amazon LightSail instance

https://pharmacope.ai/

TREATMENT RECOMMENDATION SYSTEM @ NUH, SINGAPORE



https://pharmacope.ai/

SUPPORTING EVIDENCE: WHY?

- Clinical decision of prescribing a specific drug
  - Clinical guidelines
  - Evidence of efficacy from:
    1. Previous clinical trials
    2. Mechanism of action

Mostly unknown

Completely or partially unknown

- Multiple incomplete or indirect sources of evidence needed
  - Even to evaluate a DRP model in a clinical trial

SUPPORTING EVIDENCE

Model Explainability

- XAI algorithms to highlight genes most important for prediction

Auxiliary Drug Databases

- Gene-drug associations, clinical trial information (mostly limited to effects of single gene mutations)

Drug-level validation across patients

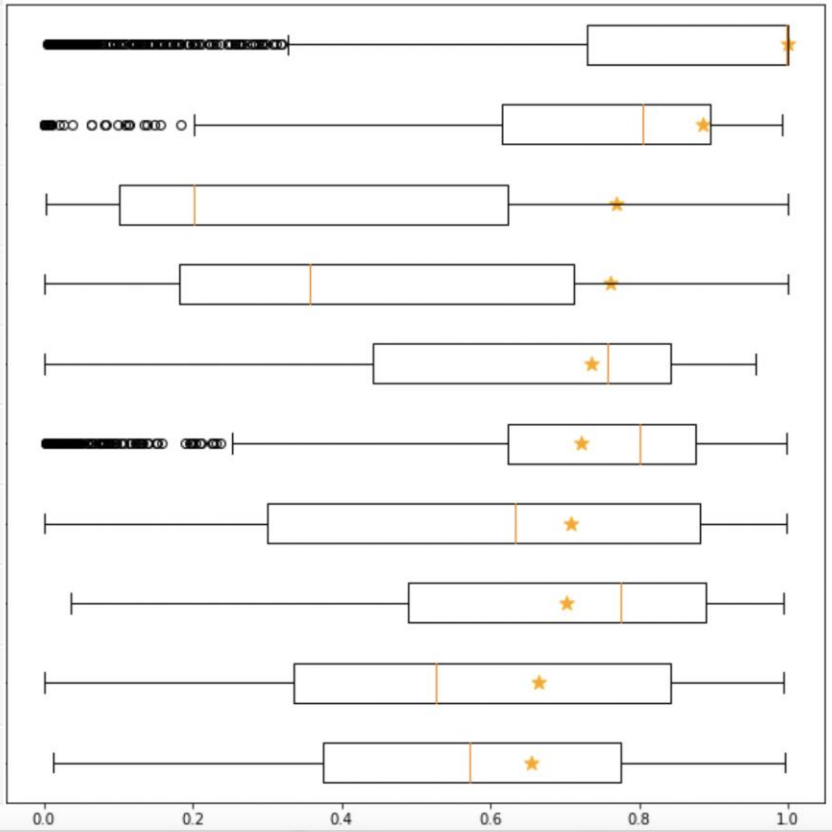
- Difference in the prediction (for the input patient) to the predictions on a reference set

Patient-level validation across drugs

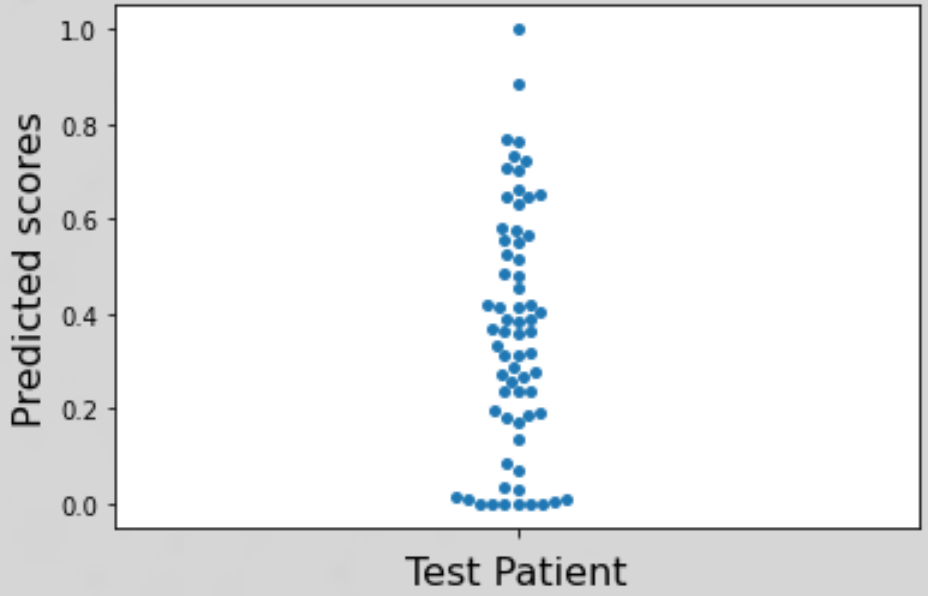
- Distribution of predictions for all drugs

SUPPORTING EVIDENCE

Drug	Score
VINORELBINE	0.9999
OSIMERTINIB	0.8863
FULVESTRANT	0.7687
BORTEZOMIB	0.7617
BELINOSTAT	0.7350
CYTARABINE	0.7219
ALPELISIB	0.7087
DACTINOMYCIN	0.7016
VENETOCLAX	0.6646
METHOTREXATE	0.6546

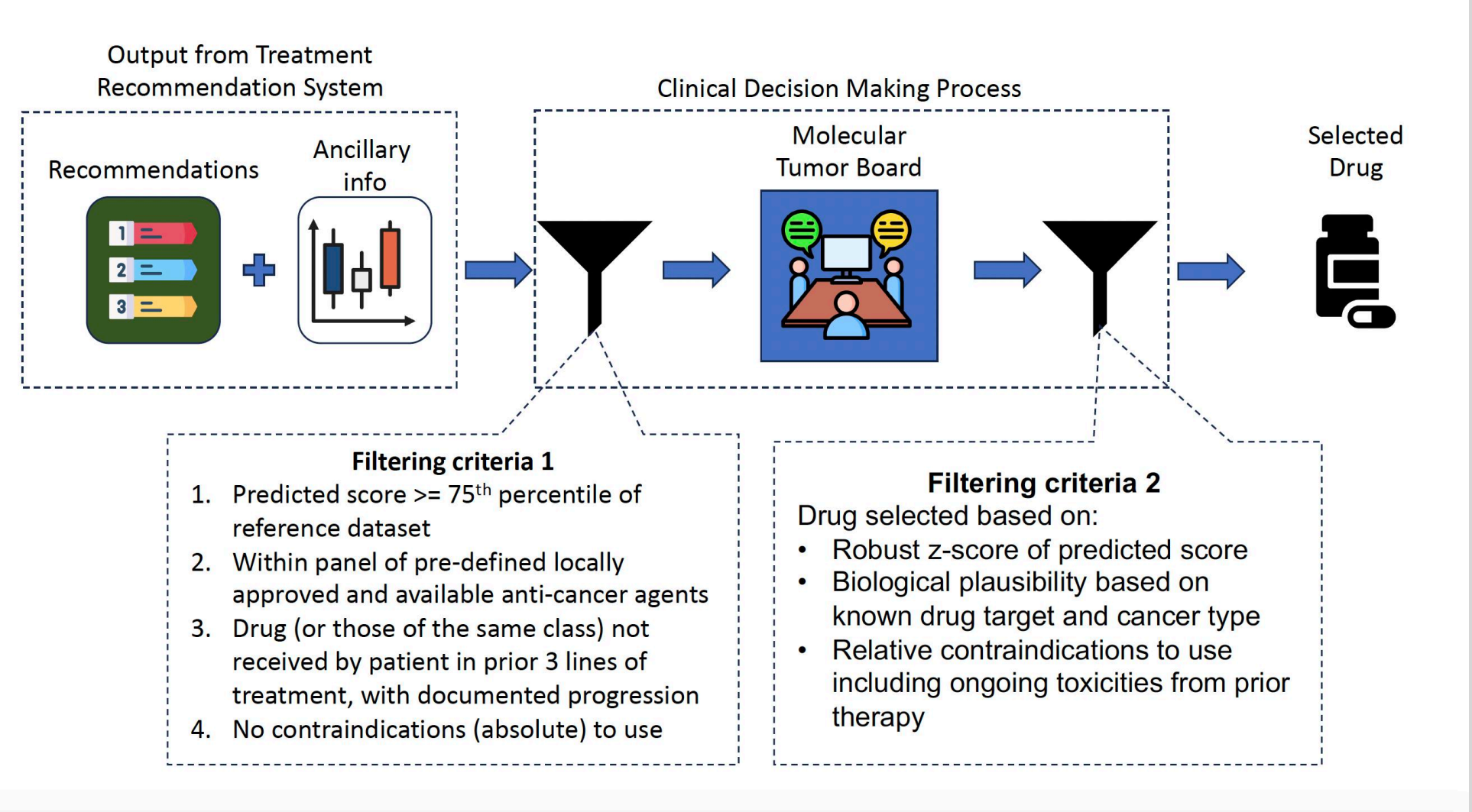


Difference in the prediction (for the input patient) to the predictions on a reference set



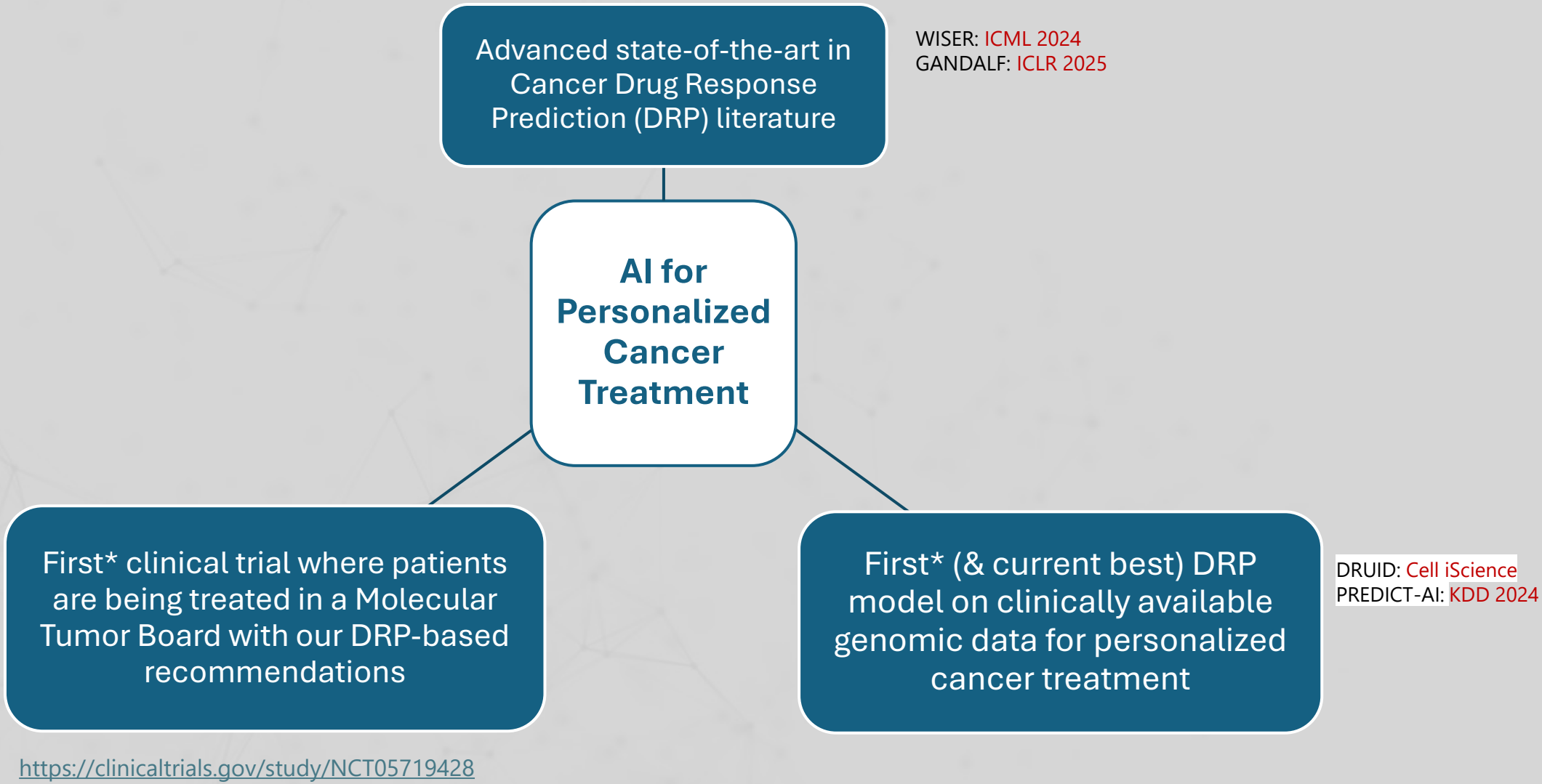
Distribution of predictions for all drugs

Clinical trial



https://clinicaltrials.gov/study/NCT05719428

CONCLUSION



* To the best of our knowledge

TEAM

Aishwarya Jayagopal
Hansheng Xue
Ziyang He
Krishna Kumar
Ragunathan Mariappan
Debabrata Mahapatra
Vaibhav Rajan

ML Team at School of Computing, NUS

Tuan Zea Pitt
Jason J Pitt

Bioinformatics Team at Cancer Science Institute, NUS

Renie Ravin

Primal Rave Software Labs, India

Robert J Walsh
Patrick Williams Jaynes
Diana Lim
David Shao Peng Tan
Anand D Jeyasekharan

Clinical Team at National University Hospital

Principal Investigators: Vaibhav Rajan and Anand D Jeyasekharan

FUTURE DIRECTIONS

Techniques

- LLMs
- Gene representations from Knowledge Graphs

Model Improvements

- Drug combinations
- Additional inputs – clinical data, genomic rearrangements
- Temporality

Decision Support Systems

- Supporting Evidence from External Knowledge Bases
- Integrate supporting evidence in ranking recommendations
- Improved AI-in-the-loop system for collective decision making in MTBs

Thanks to Aishwarya Jayagopal for most of the slides