



CLASSIFYING CLINICALLY ACTIONABLE GENETIC MUTATIONS

CHEONGYU CHYE – DSI 13 – CAPSTONE PROJECT

OUR PROBLEM STATEMENT

- Cancer is a top public health issue¹:
 - Globally, about 1 in 6 deaths is due to cancer
 - One in every 4-5 people in Singapore may develop cancer in their lifetime
- Today, clinical pathologists can perform genomic sequencing on a patient's tumour sample to determine if it carries mutations that could aid in treatment, or clinical trials²
- Once these genes and mutations have been identified, clinical pathologists then have to manually review a growing corpus of related biomedical literature to classify the mutations – this process is tedious and time consuming
- Our problem statement: to build a classifier that can help to automate this classification
- Metrics: balanced (weighted) accuracy and F1 scores, micro-average AUC
- Success measure: beat baseline accuracy (0.287) by $\geq 10\%$

Sources:

¹ - <https://www.nccs.com.sg/patient-care/cancer-types/cancer-statistics>, <https://www.who.int/news-room/fact-sheets/detail/cancers>

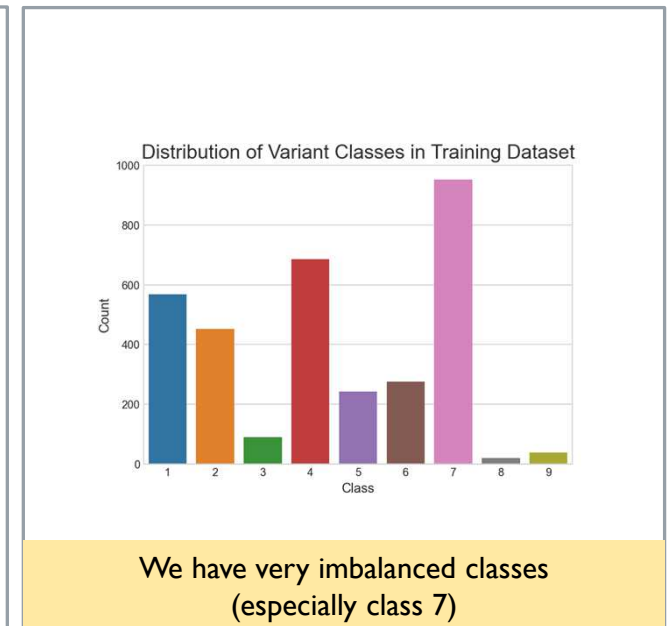
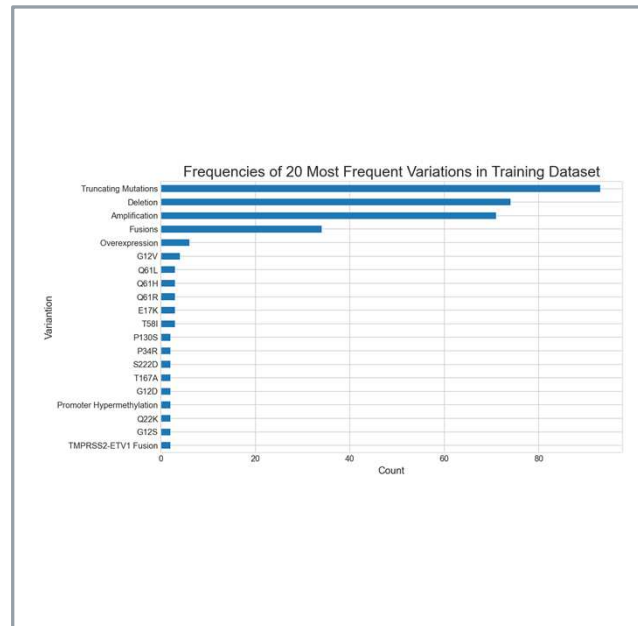
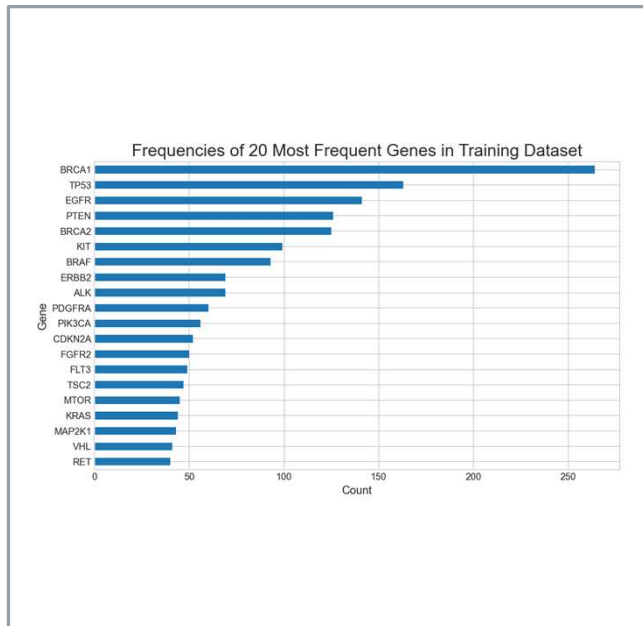
² - <https://www.mskcc.org/cancer-care/diagnosis-treatment/diagnosing/role-pathology>

THE CHALLENGE ...

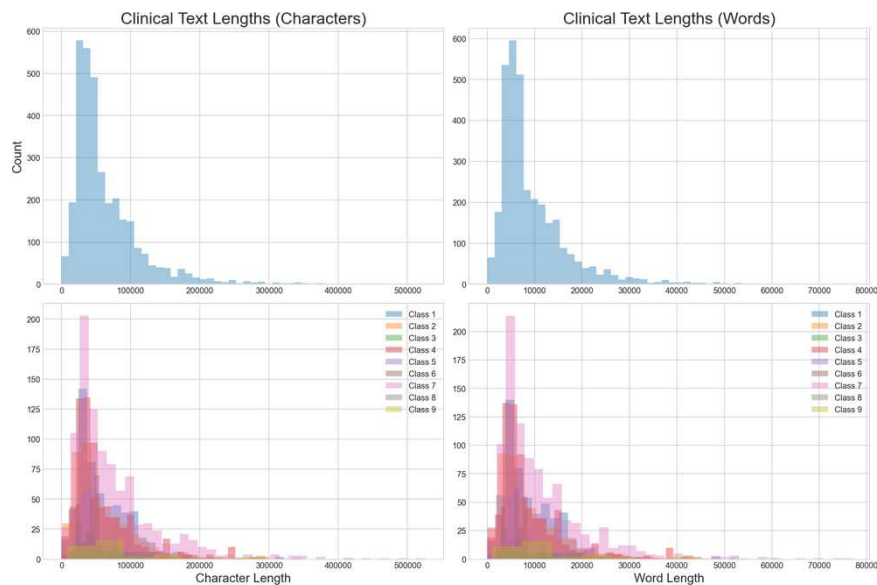
3,321 Training Samples	The incidence of breast cancer is increasing in China in recent decades, and familial breast cancer accounts for 5–10% of the total patients in Chinese women. Germline mutations in breast cancer susceptibility genes, BRCA1 (MIM# 113705) and BRCA2 (MIM# 600185), are responsible for only approximately 10% of Chinese breast cancer families [Liede and Narod, 2002]; ...	+	CHEK2	+	H371Y	=	9
	An unselected series of 310 colorectal carcinomas, stratified according to microsatellite instability (MSI) and DNA ploidy, was examined for mutations and/or promoter hypermethylation of five components of the WNT signaling cascade [APC, CTNNB1 (encoding β-catenin), AXIN2, TCF4, and WISP3] and three genes indirectly affecting this pathway [CDH1 (encoding E-cadherin), PTEN, and TP53]. APC and TP53 mutations were each present more often in microsatellite-stable (MSS) tumors ...	+	AXIN2	+	Q1537R	=	7
986 Test Samples	Mycosis fungoides and Sézary syndrome are primary cutaneous T cell malignancies derived from CD4+ skin-homing T cells 1, 2. Mycosis fungoides cases with limited skin involvement have a favorable prognosis; however, the median survival time for cases with cutaneous tumors and generalized erythroderma is approximately 4 years, and patients with Sézary syndrome fare even worse ...	+	WNT4	+	E216G	=	?
	Regulated progression through the cell cycle requires sequential expression of a family of proteins called cyclins. Upon their induction, cyclins form complexes with specific cyclin-dependent kinases (CDKs), creating active holoenzymes that phosphorylate target proteins that are required for cell-cycle progression. Induction of the proto-oncogene cyclin D1, and its binding to ...	+	GJB3	+	E183K	=	?

Legend: Clinical Text Gene Variation Class

WHAT WE NOTICED – GENES, VARIANTS & CLASSES



WHAT WE NOTICED – CLINICAL TEXT

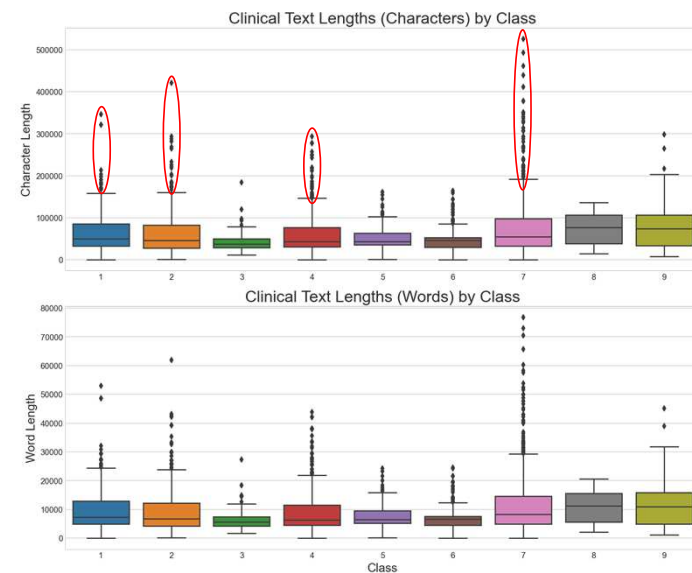


Average # of words

9,543

Maximum # of words

76,708

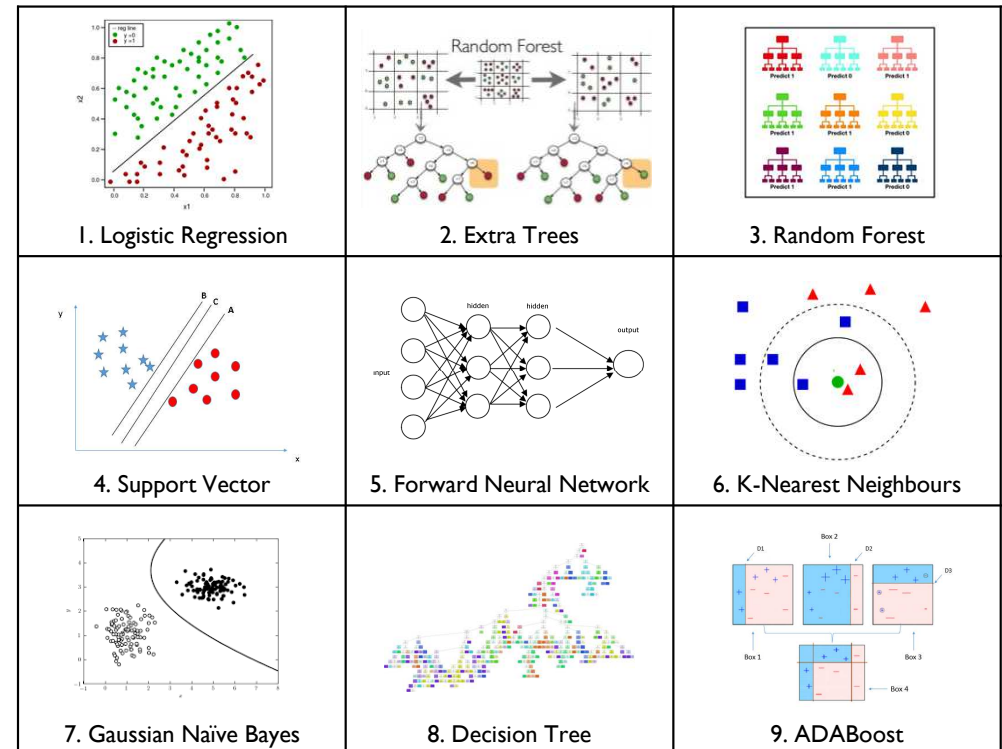


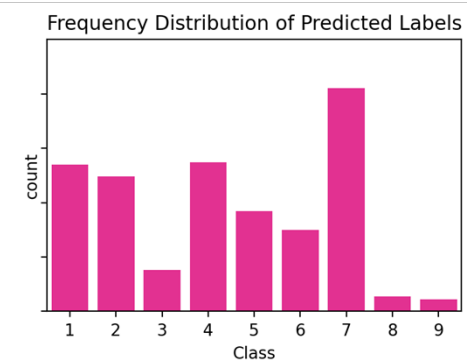
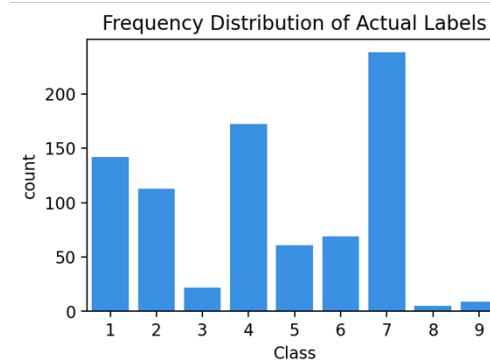
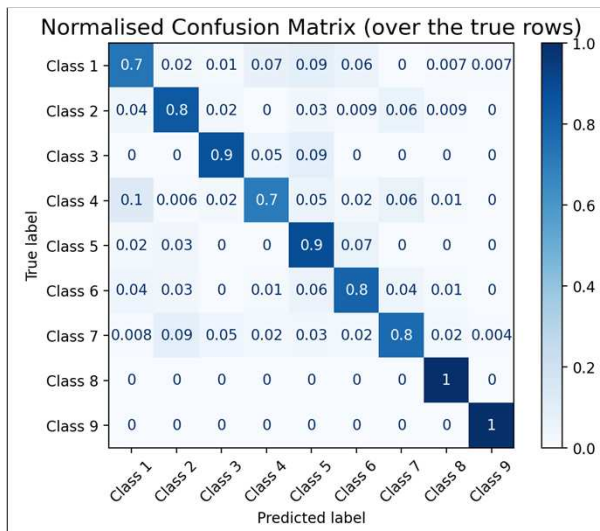
We decided not to remove any of the outliers given the importance of retaining as many words as possible for training our models

FINDING OUR BASELINE MODEL

- What we did to ...
 - Work with gene and variation data – we performed one-hot encoding to create a feature for each value
 - Address our imbalanced classes – we created more samples for selected classes (using ADASYN)
 - Make our model more generalisable – we ‘pivoted’ the data to reduce the number of features to just 1,800 (using PCA)
 - Pick the right model – we systematically trained and evaluated 9 different candidates
- The outcome – a baseline model consisting of a logistic regression classifier

Best Performing Classifiers (1=Best)

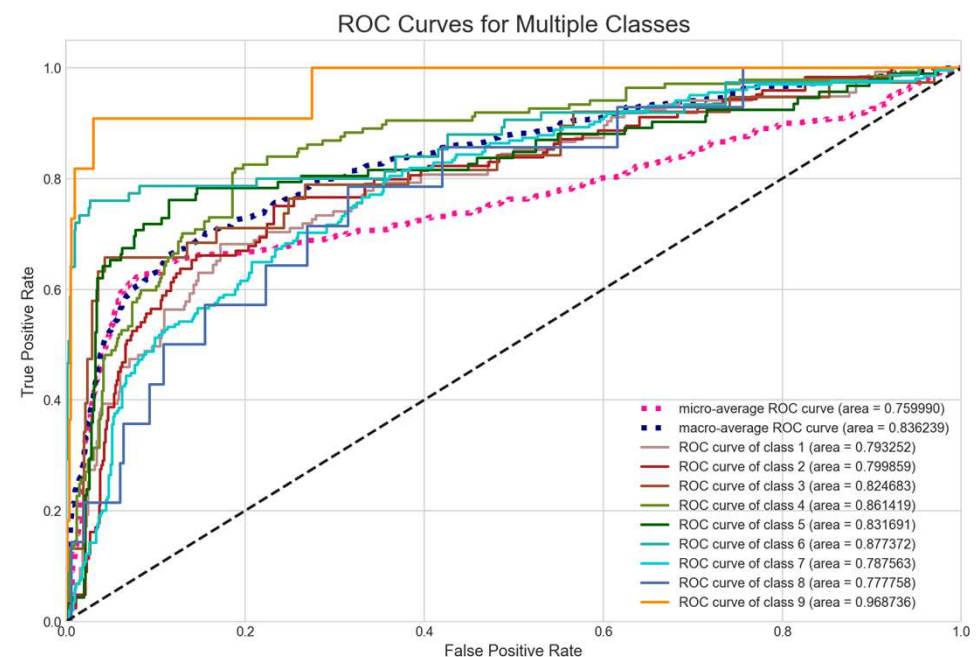




HOW DID OUR MODEL DO?

- Correctly predicts an average of ~54% of actual positives across all classes
- Has an average F1 score of ~62% for each class (weighted by the number of true instances of each class)
- Has an average AUC score of ~76% across all classes

8



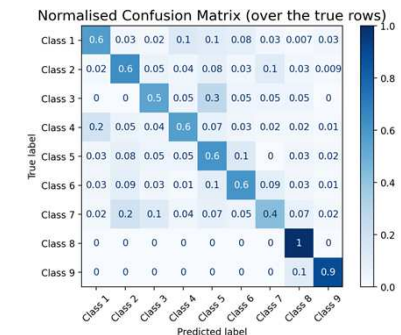
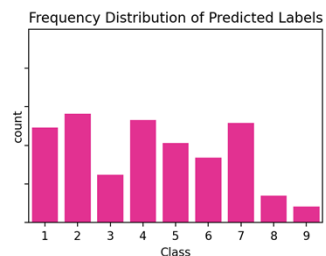
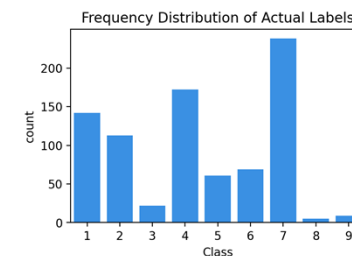
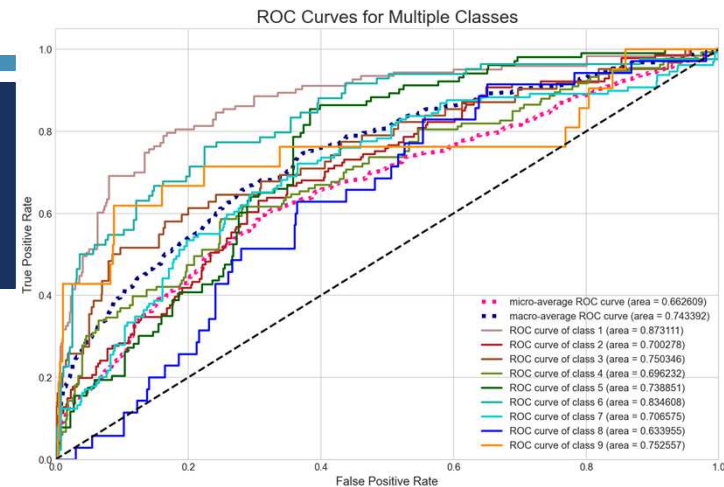
CAN WORD EMBEDDINGS IMPROVE OUR MODEL?

- The problem: we have too many features to begin with (~76k) because we are creating a feature for every unique word in all the training text
- The idea: use word embeddings that need fewer features (in our case we use 100-dimensional embeddings)
- The hope: we can train a model faster and achieve higher accuracy

unselected series colorectal carcinoma
stratify accord microsatellite instability msi
dna ploidy examine promoter
hypermethylation five component wnt signal
cascade apc cttnb encode catenin axin tcf
wisp three gene indirectly affect pathway
cdh encode e cadherin pten tp apc tp
present often microsatellite stable ms



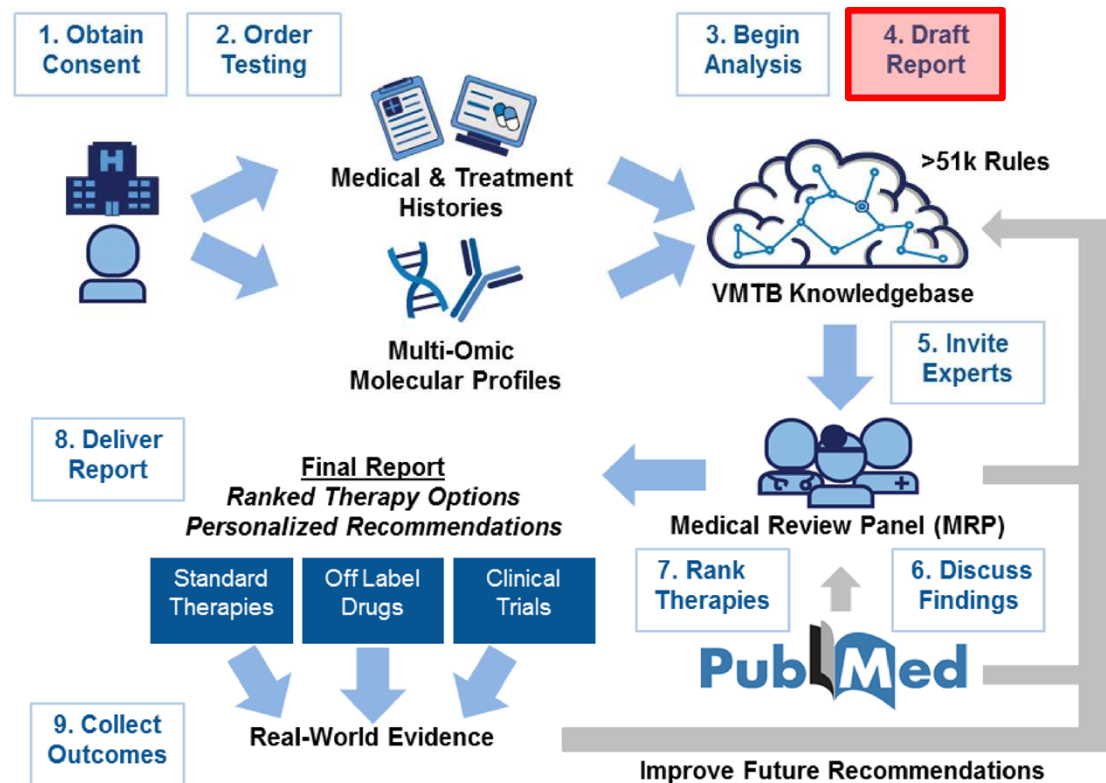
Text Instance	100-dimensional word embeddings									
	1	2	3	...	99	100				
1	1.624678	-0.914357	-0.074222	...	0.997731	-0.449048				
2	1.724697	-0.940365	-0.251819	...	1.002514	-0.585488				
3	1.692799	-0.932071	-0.164691	...	0.950496	-0.45507				
4	1.70435	-0.86725	-0.223373	...	1.00146	-0.539933				
5	1.677125	-0.908731	-0.161353	...	0.99578	-0.496753				
...				



- The outcome: the best-trained alternative model had **lower** scores than our baseline model
- The conclusion: our baseline model is still a better choice

HOW OUR MODEL CAN BE USED IN FUTURE

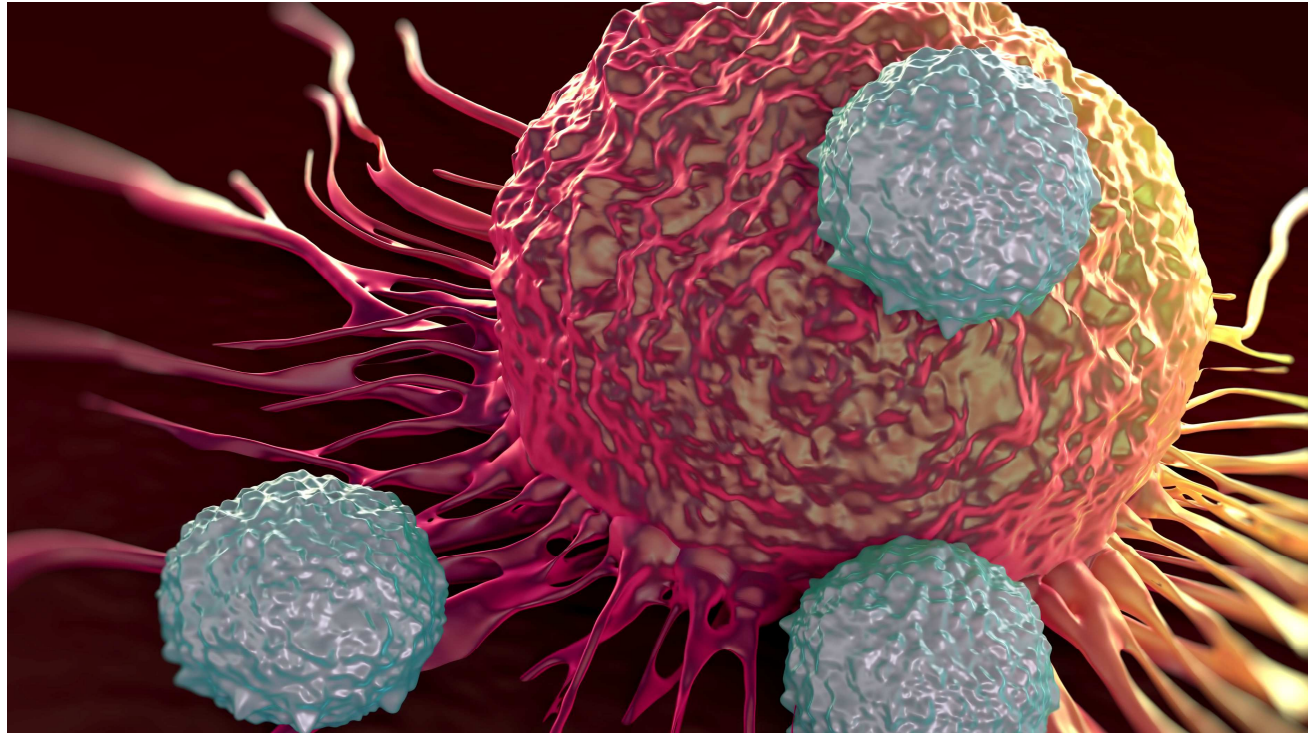
- Model predictions can be incorporated into reports that a molecular tumour board can use to make decisions on treatment



Source: JAMIA Open, Volume 2, Issue 4, December 2019, Pages 505–515, <https://doi.org/10.1093/jamiaopen/ooz045>

LIMITATIONS & RISKS

- Limitations
 - Numbers, dashes and Greek characters are currently being removed during pre-processing – some of these may be important
 - E.g. “E-cadherin” → “e cadherin”, “ β -catenin” → “catenin”, “TP53” → “tp”



THANK YOU

YUCHYE@GMAIL.COM



ADDITIONAL REFERENCE SLIDES



OUR APPROACH – IN DETAIL

