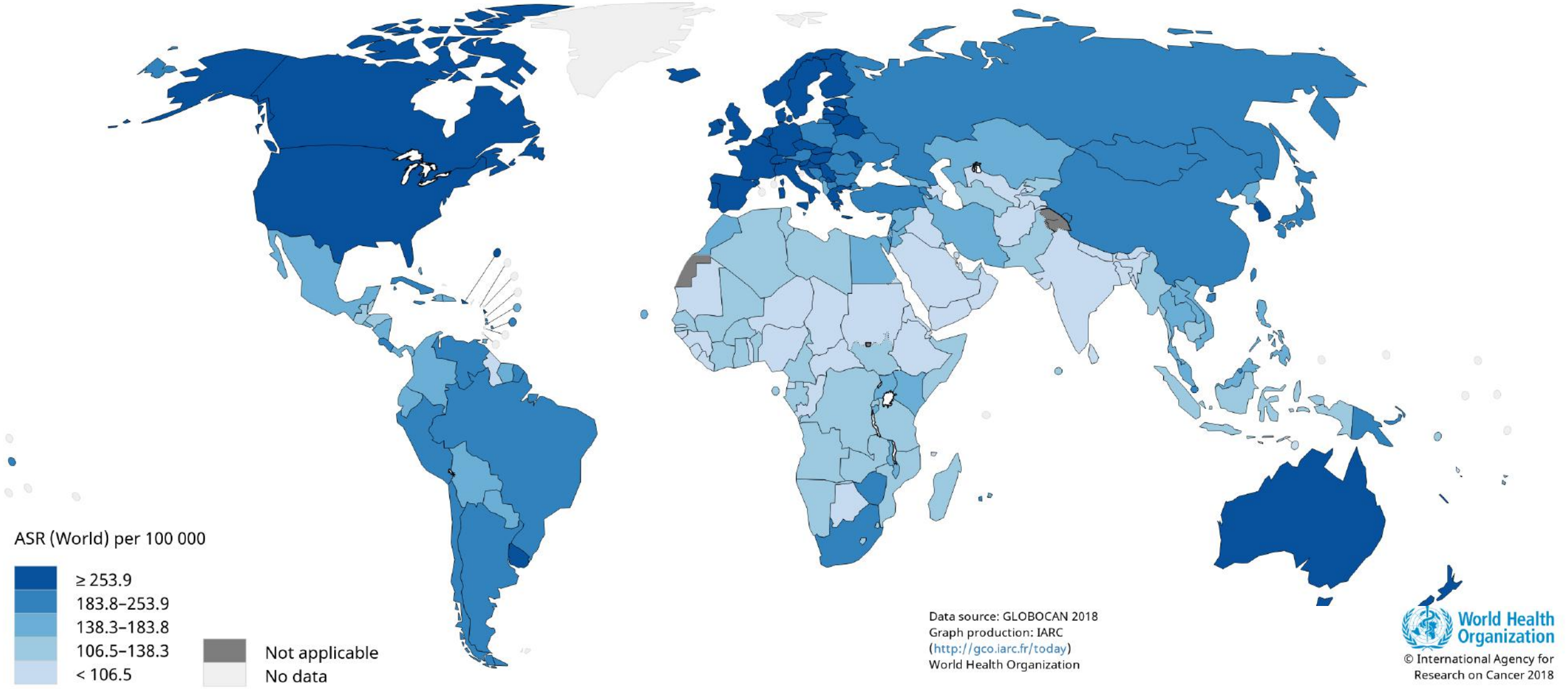
IBM **Research** | Africa

Addressing South Africa's Cancer Reporting Delay with Machine Learning

Waheeda Saib (PI)
Applied Research Scientist
IBM Research | Africa

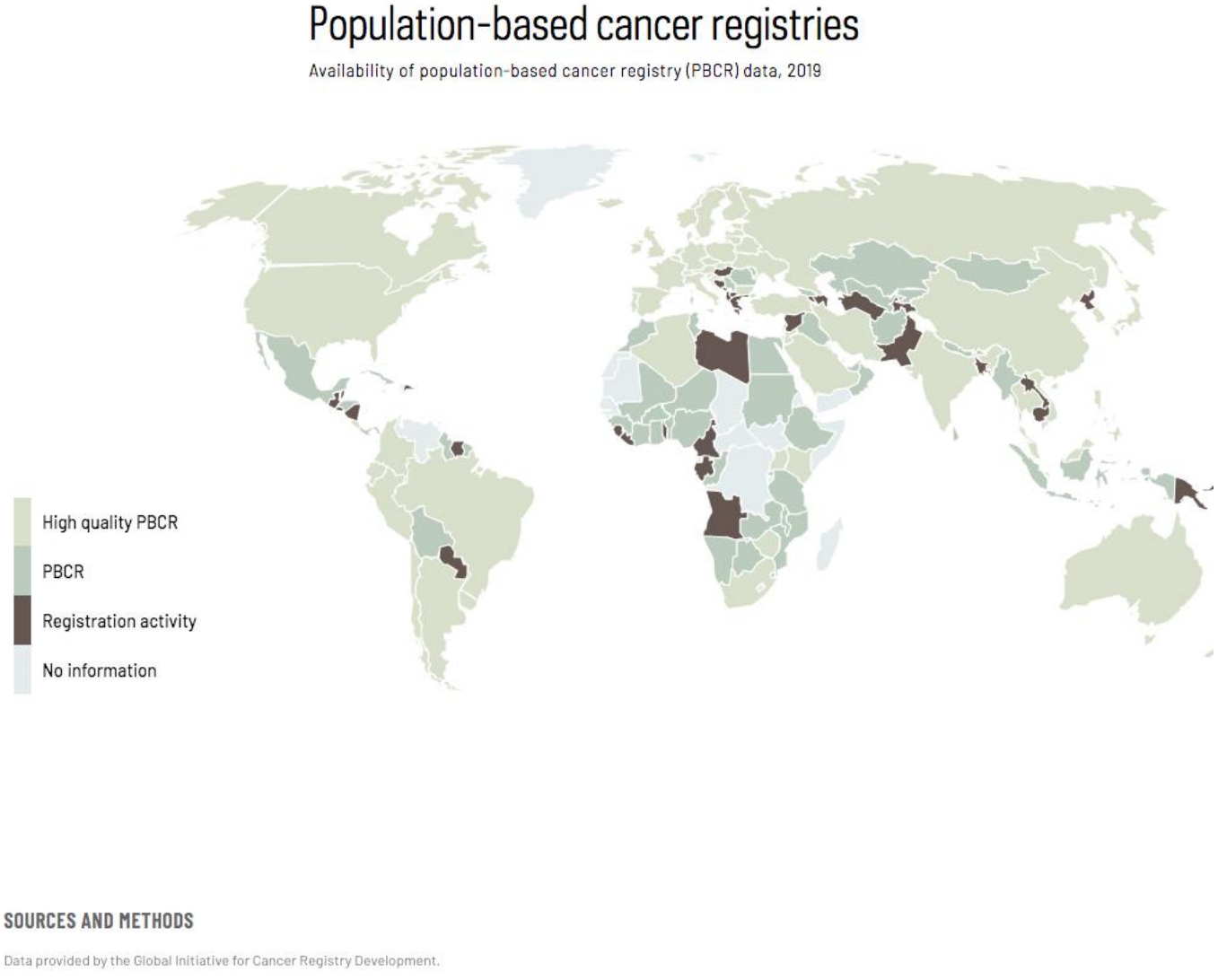
Cancer Burden



Cancer Surveillance

Cancer registries improve disease detection and treatment because they...

- Identify cancer trends and high-risk groups
- Help set priorities for health resources and programs
- Advance clinical and health research



Cancer Coder

Motivation

- Cancer registries report cancer statistics essential for healthcare resource and intervention planning.
- Manual processes result in considerable lag time in national cancer statistics reporting
- Cancer landscape in Africa is unclear

Problem

- Challenging for Health officials to understand the impact of cancer in the country and allocate resources accordingly.
- Real time statistics required to drive cancer policy

Cancer Coder

Research Objectives

- Use ML/DL to automatically assign topography and morphology codes to free text pathology reports
- Investigate how medical concept extraction can improve accuracy of doing automatic report coding
- Investigate how Graph Neural Networks can used to improve accuracy of automatic report coding

Near term Impact

- Increase speed and accuracy of cancer report labelling
- More consistent and less subjective coding can be performed
- Automated cancer classification system

Partners

External Partners

National Cancer Registry

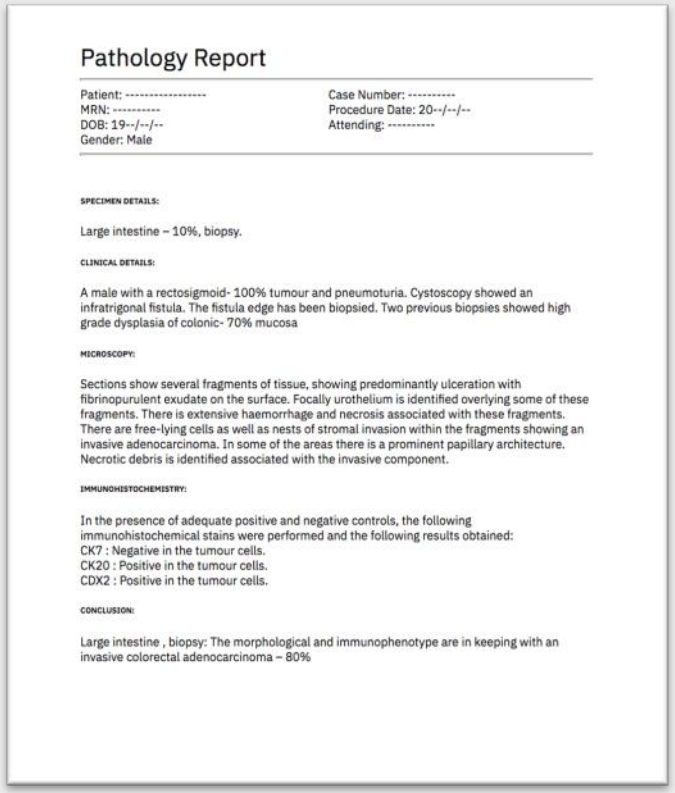
South Africa

Internal Partners

IBM Research Yorktown Lab

Data & Pre-Processing Pipeline

Pathology Report



Pathology Report

Patient: ----------
MRN: ----------
DOB: 19--/--/--
Gender: Male

Case Number: ----------
Procedure Date: 20--/--/--
Attending: ----------

SPECIMEN DETAILS:

Large intestine – 10%, biopsy.

CLINICAL DETAILS:

A male with a rectosigmoid- 100% tumour and pneumoturia. Cystoscopy showed an infratrigonal fistula. The fistula edge has been biopsied. Two previous biopsies showed high grade dysplasia of colonic- 70% mucosa

MICROSCOPY:

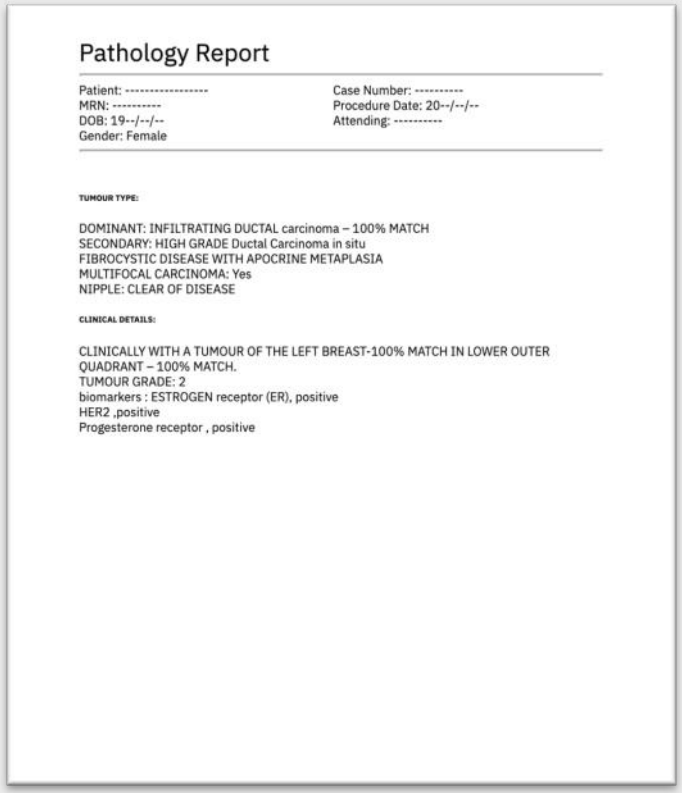
Sections show several fragments of tissue, showing predominantly ulceration with fibrinopurulent exudate on the surface. Focally urothelium is identified overlying some of these fragments. There is extensive haemorrhage and necrosis associated with these fragments. There are free-lying cells as well as nests of stromal invasion within the fragments showing an invasive adenocarcinoma. In some of the areas there is a prominent papillary architecture. Necrotic debris is identified associated with the invasive component.

IMMUNOHISTOCHEMISTRY:

In the presence of adequate positive and negative controls, the following immunohistochemical stains were performed and the following results obtained:
CK7 : Negative in the tumour cells.
CK20 : Positive in the tumour cells.
CDX2 : Positive in the tumour cells.

CONCLUSION:

Large intestine , biopsy: The morphological and immunophenotype are in keeping with an invasive colorectal adenocarcinoma – 80%



Pathology Report

Patient: ----------
MRN: ----------
DOB: 19--/--/--
Gender: Female

Case Number: ----------
Procedure Date: 20--/--/--
Attending: ----------

TUMOUR TYPE:

DOMINANT: INFILTRATING DUCTAL carcinoma – 100% MATCH
SECONDARY: HIGH GRADE Ductal Carcinoma in situ
FIBROCYSTIC DISEASE WITH APOCRINE METAPLASIA
MULTIFOCAL CARCINOMA: Yes
NIPPLE: CLEAR OF DISEASE

CLINICAL DETAILS:

CLINICALLY WITH A TUMOUR OF THE LEFT BREAST-100% MATCH IN LOWER OUTER QUADRANT – 100% MATCH.
TUMOUR GRADE: 2
biomarkers : ESTROGEN receptor (ER), positive
HER2 ,positive
Progesterone receptor , positive

Challenges

- Structured and unstructured text reports
- Multilingual reports
- Duplicate reports, sometimes with different labels
- Highly imbalanced classes

Data Pre-Processing

All path. reports

634K

Duplicate removal

513K

Filter short reports (<20 words)

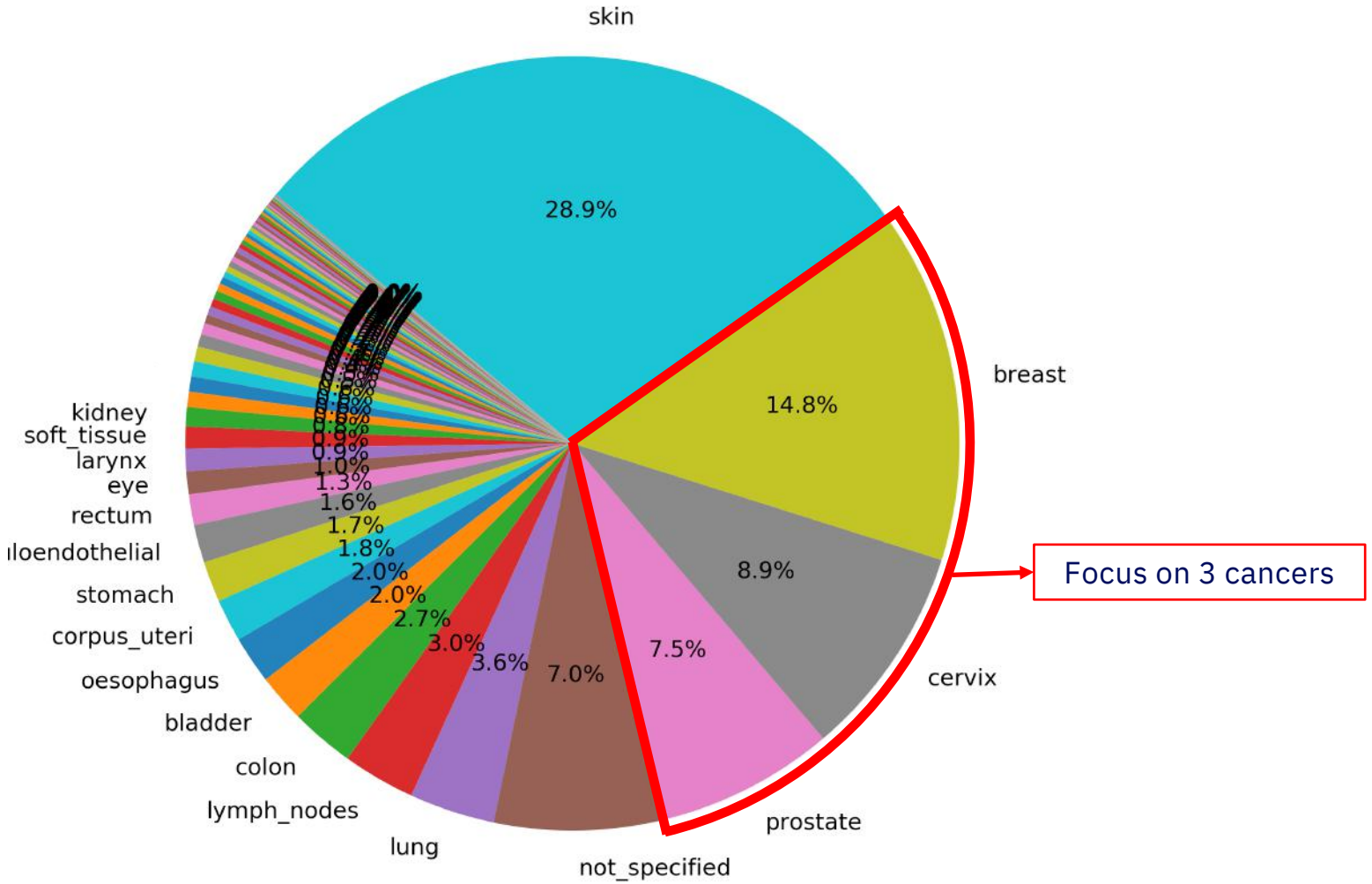
480K

Language detection & filtering

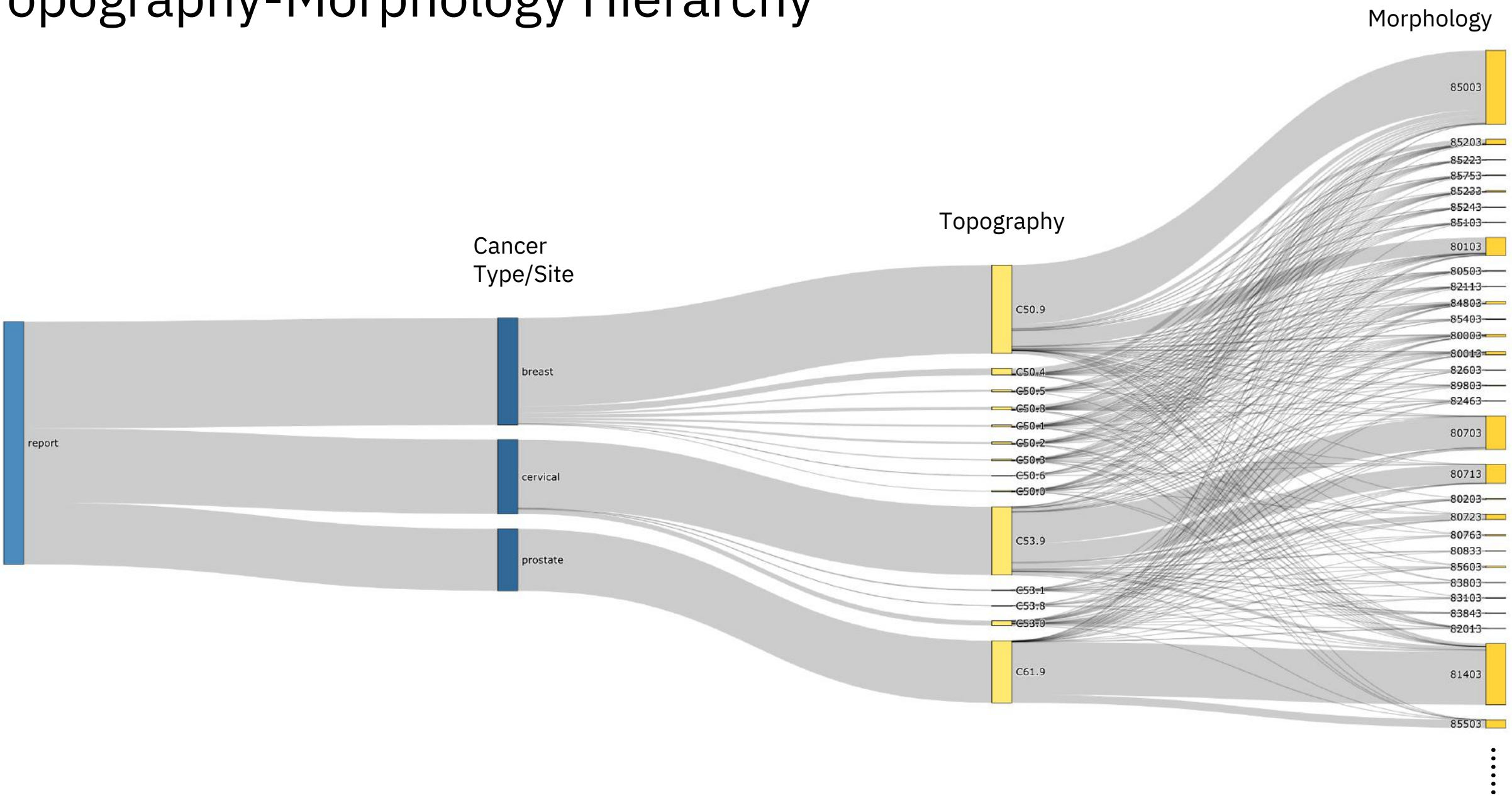
410K

Train (313K) *2012-2015*	**Valid (52K)** *Jan-Jun 2016*	**Test (45K)** *Jul-Dec 2016*

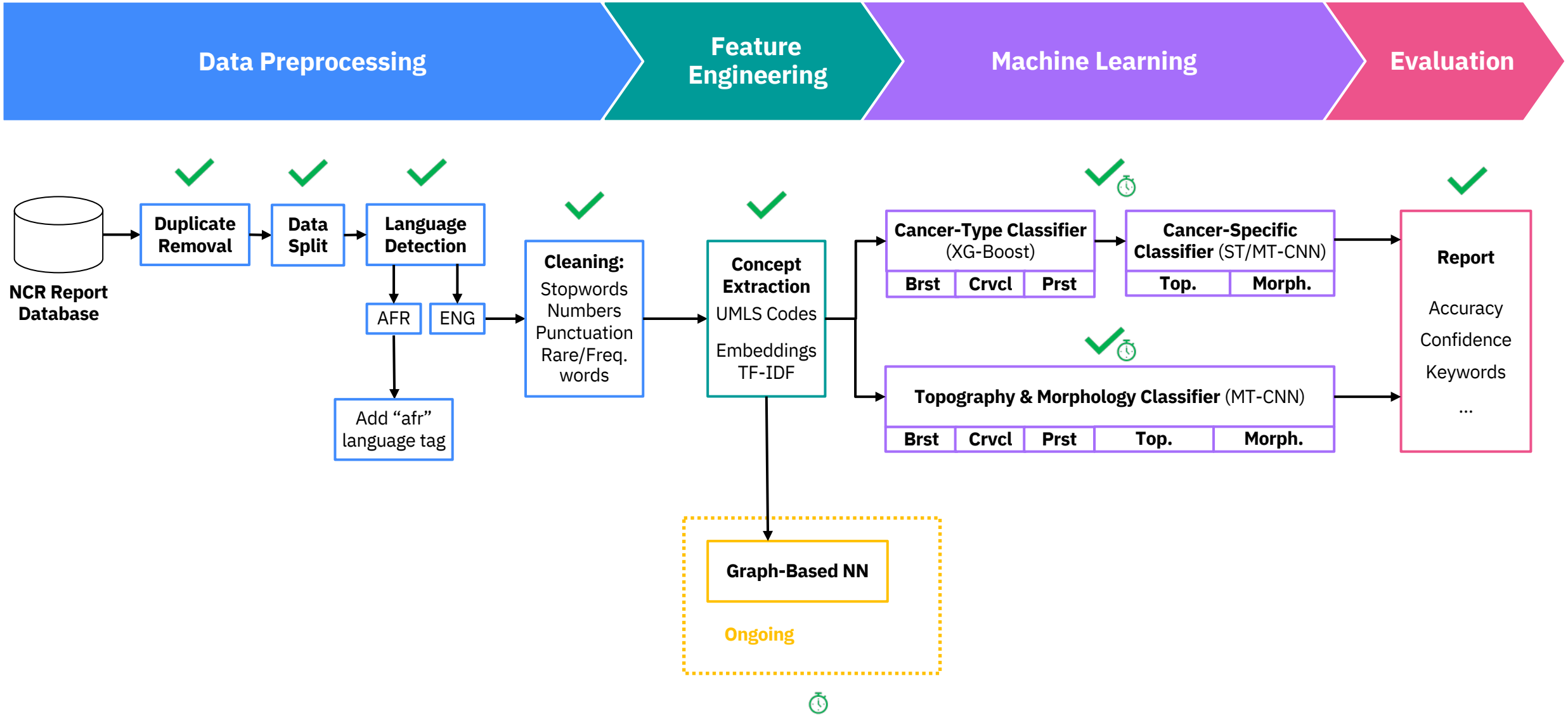
Cancer-type Breakdown



Topography-Morphology Hierarchy



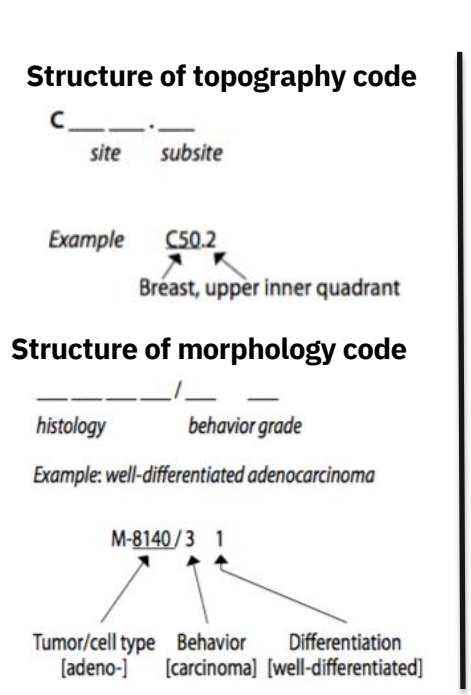
Cancer Coder Pipeline

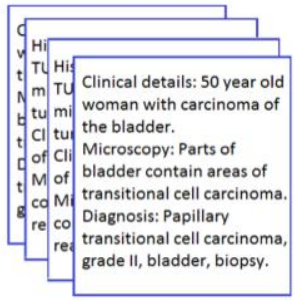


ML for Cancer-Site Classification and Topography & Morphology Coding

Hierarchical Classification for ICD-O Classification

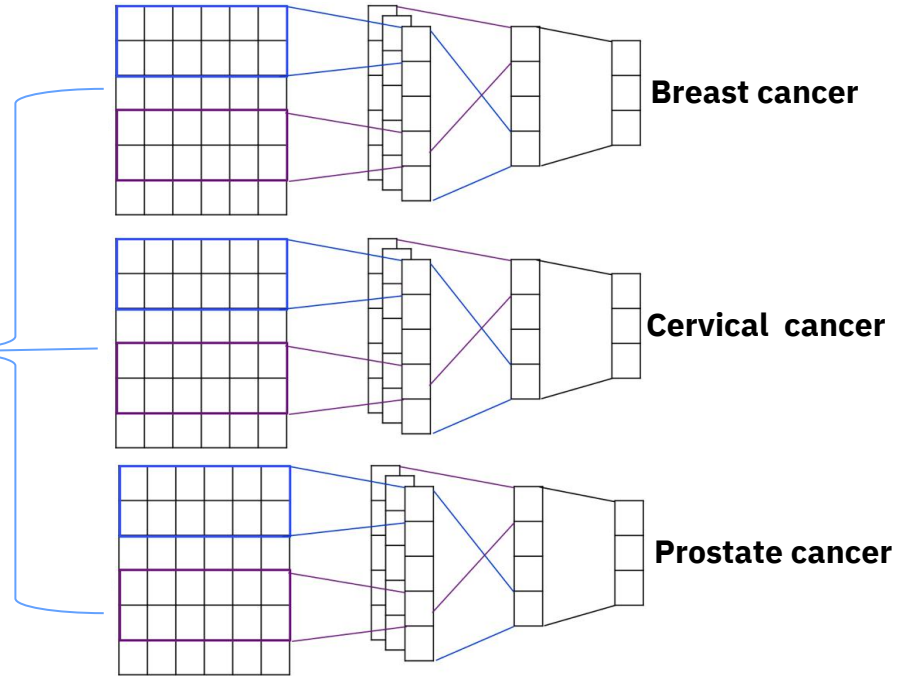
Top down Multi-task Convolutional Neural Network Ensemble for ICD-0 classification





	R1	R2	R3	R4
Carcinoma	1.386		2.773	
Grade		0.863	0.287	0.575
Bladder	2.773		2.079	
Tumour	0.863	0.575		1.438

Tf–idf (term frequency–inverse document frequency)



Motivation

- ICD-0 is hierarchical in nature
- Hierarchical classification has not been explored for ICD-0 classification

Related work

- Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. Alawad, M et al.,(2020)
- Deep learning for automated extraction of primary sites from cancer pathology reports. Qiu, J et al., (2018).

Method

- All models employ Text Filtering TF-IDF and word embeddings
- XGBoost model first level classifier identifying cancer type
- Specialized Multi-task CNN models are second level classifiers identifying primary tumour site and cell origin

Cancer Specific Model Exploration

Cancer Type	ICD-O Classes	Model Variant	Classification	F1 Micro	F1 Macro	Accuracy
Breast	8	Multi-task Multiclass CNN	Topography	0.86	0.43	0.86
Breast	8	Multi-task Multiclass CNN	Morphology	0.87	0.61	0.87
Cervical	3	Multi-task Multiclass CNN	Topography	0.95	0.65	0.95
Cervical	8	Multi-task Multiclass CNN	Morphology	0.88	0.83	0.88
Prostate	3	Multiclass CNN	Morphology	0.93	0.73	0.93

Next steps

Feature Engineering

Incorporate medical domain knowledge

Pilot Study

Future Impact

- Provide real time cancer statistics
- Provide clinically relevant cancer coding and statistics platform that is not currently available
- Improved healthcare resource and intervention planning


More than 100 000 South Africans are diagnosed with cancer every year
SUPERVISOR
CODER

Cancer Guidelines Navigator, Future Applications & IBM Service Corps

American Cancer Society®

IBM

NCCN National Comprehensive Cancer Network®

AFRICAN CANCER COALITION

CLINTON HEALTH ACCESS INITIATIVE

ALLIED AGAINST CANCER™

Goal

Help improve access to high-quality cancer care and treatment in Sub-Saharan Africa

Our Impact

25K

African health workers trained in assessment and basic pain management according to the World Health Organization guidelines

34

NCCN Harmonized Guidelines™ for Sub-Saharan Africa, which cover 85% of cancer incidence in Sub-Saharan Africa

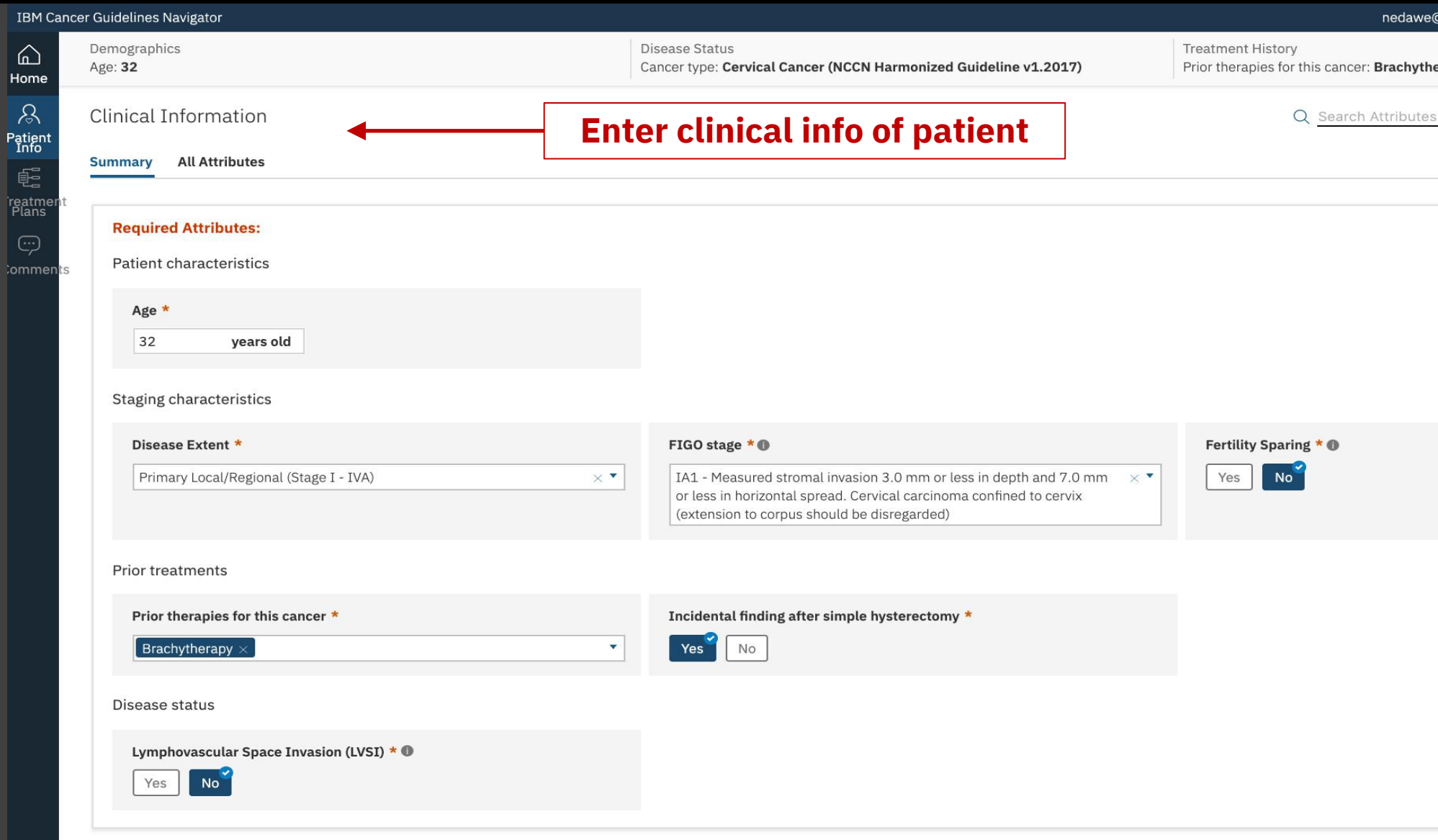
42%

Cancer cases in the region covered by market access agreements reached for 16 SRA-approved cancer medicines

https://www.alliedagainstcancer.org/

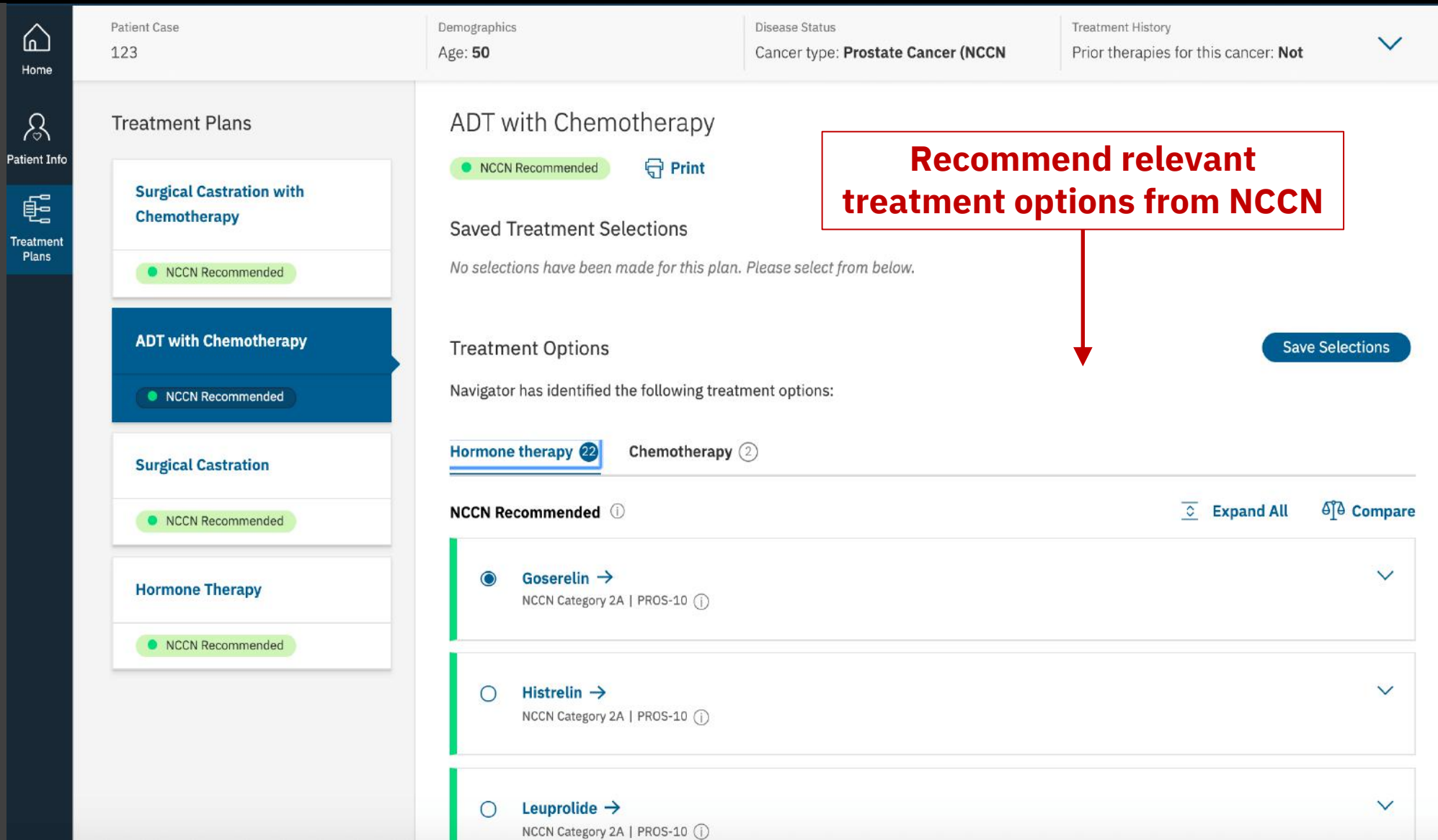
IBM Cancer Guidelines Navigator: Reference system for NCCN Harmonized Guidelines

An online tool that provides African oncology professionals with interactive access to the NCCN Harmonized Guidelines for Sub-Saharan Africa



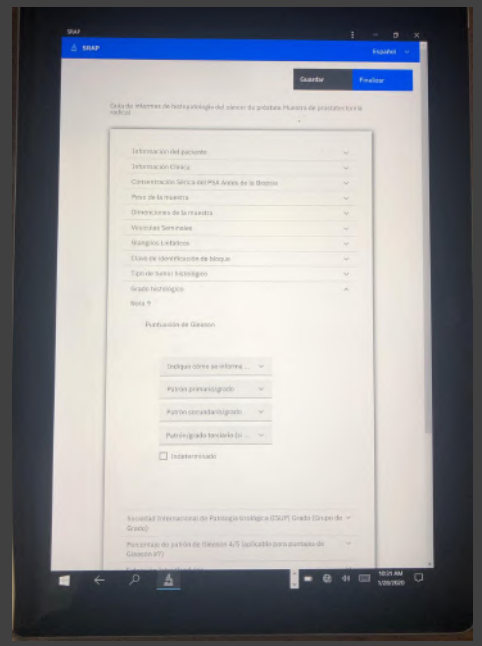
IBM Cancer Guidelines Navigator: Reference system for NCCN Harmonized Guidelines

An online tool that provides African oncology professionals with interactive access to the NCCN Harmonized Guidelines for Sub-Saharan Africa



IBM Service Corps Paraguay

- We built a structured reporting tool for anatomical pathology reports (SRAP)
- New law in Paraguay: all hospitals to submit cancer reports to the Cancer Registry in structured form (paper or electronic)
- Cancer Registry planning to hire human coders because pathologists aren't expected to enter ICD-10 codes
- Ministry of Health and National Cancer Institute (INCAN)
- SRAP + Cancer Coder would be an impactful solution for Paraguay



Thank you

Waheeda Saib
Applied Research Scientist
WSaib@za.ibm.com


IBM