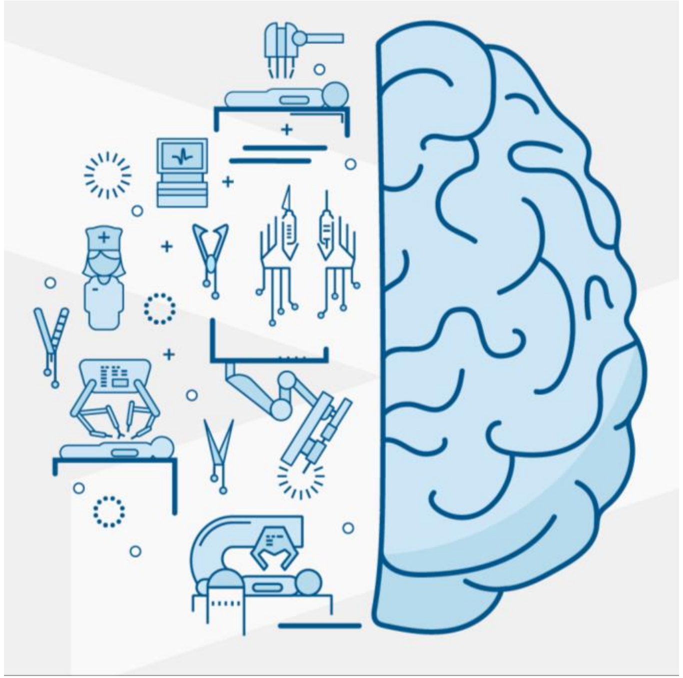


Transforming cancer registries with machine learning and neural networks

Adhari Abdullah AlZaabi
College of Medicine and Health Sciences
Sultan Qaboos University
adhari@squ.edu.om

Outline

Introduction

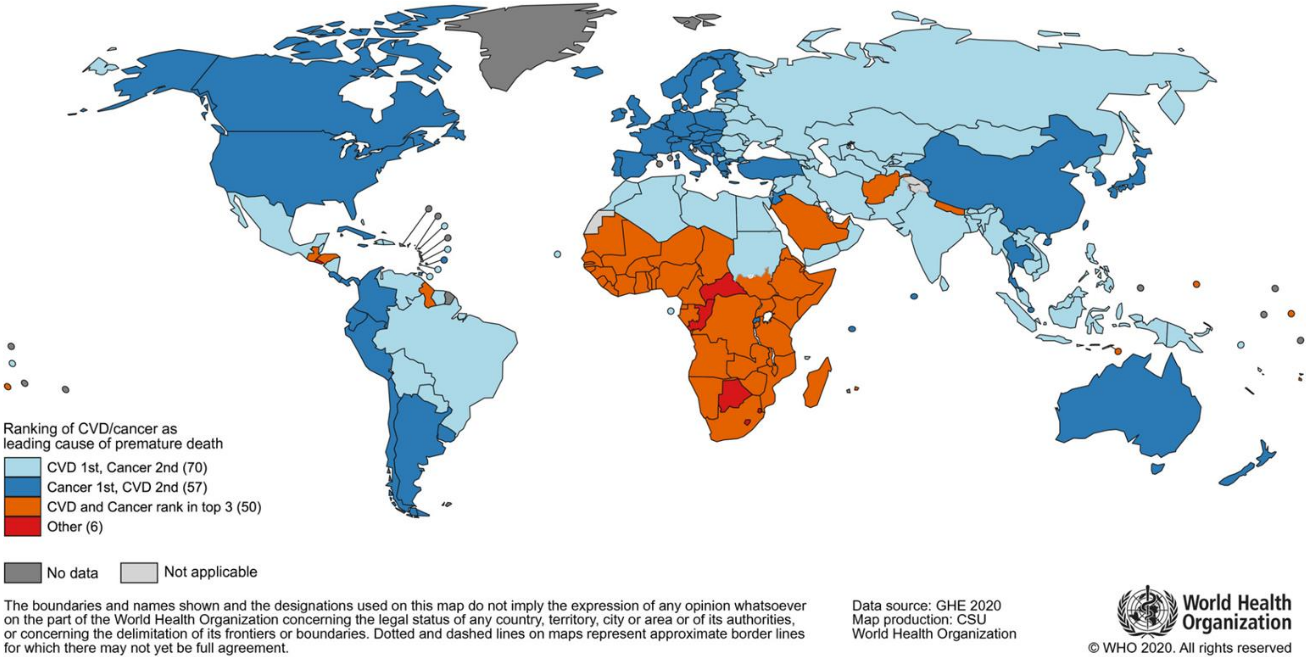
Methods

Results

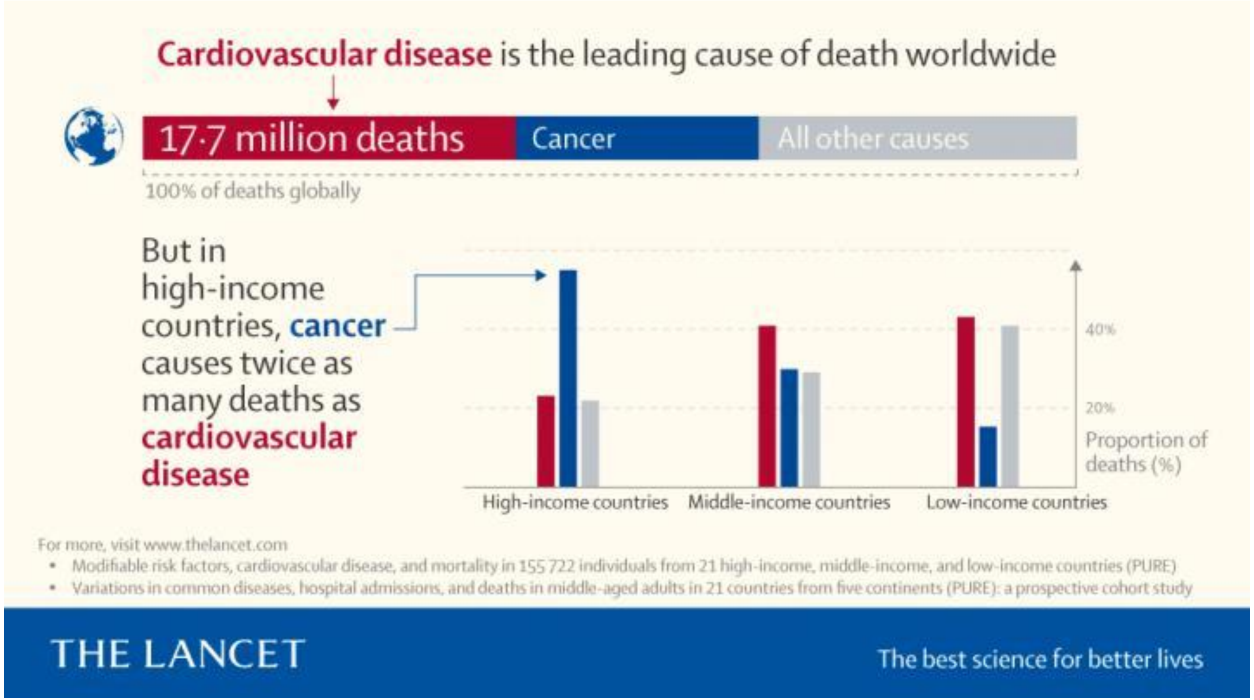
Conclusion

Please, consider this is work in progress !!!

Cancer Burden (leading cause of death)

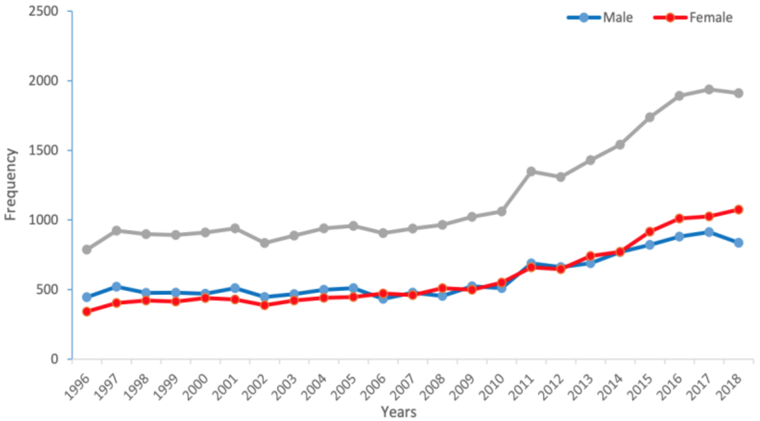


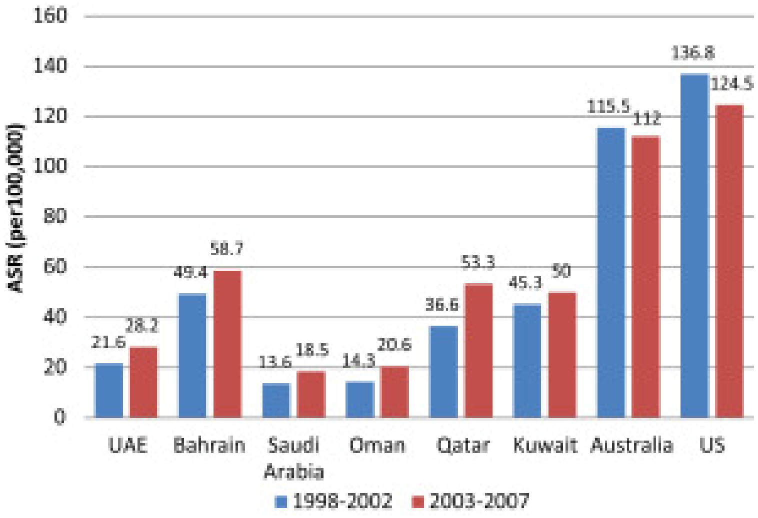
Cancer Burden (leading cause of death)



Oman and Gulf Cooperation Council cancer incidence

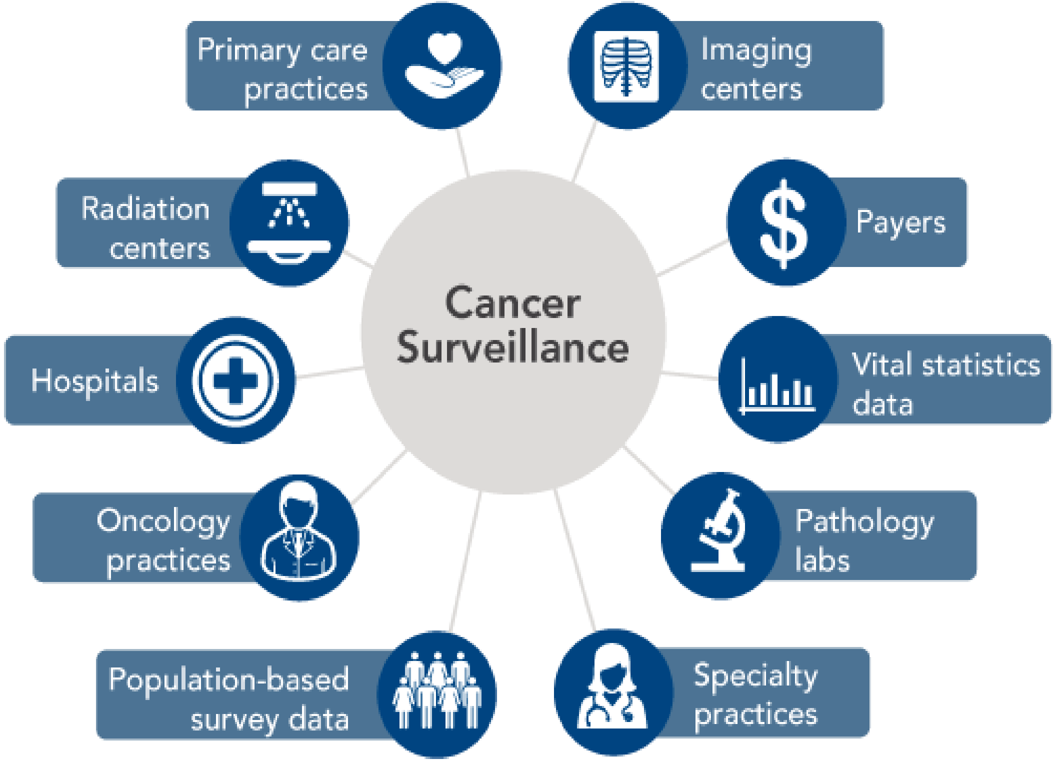
Figure 4: Trends of cancer cases 1996-2018

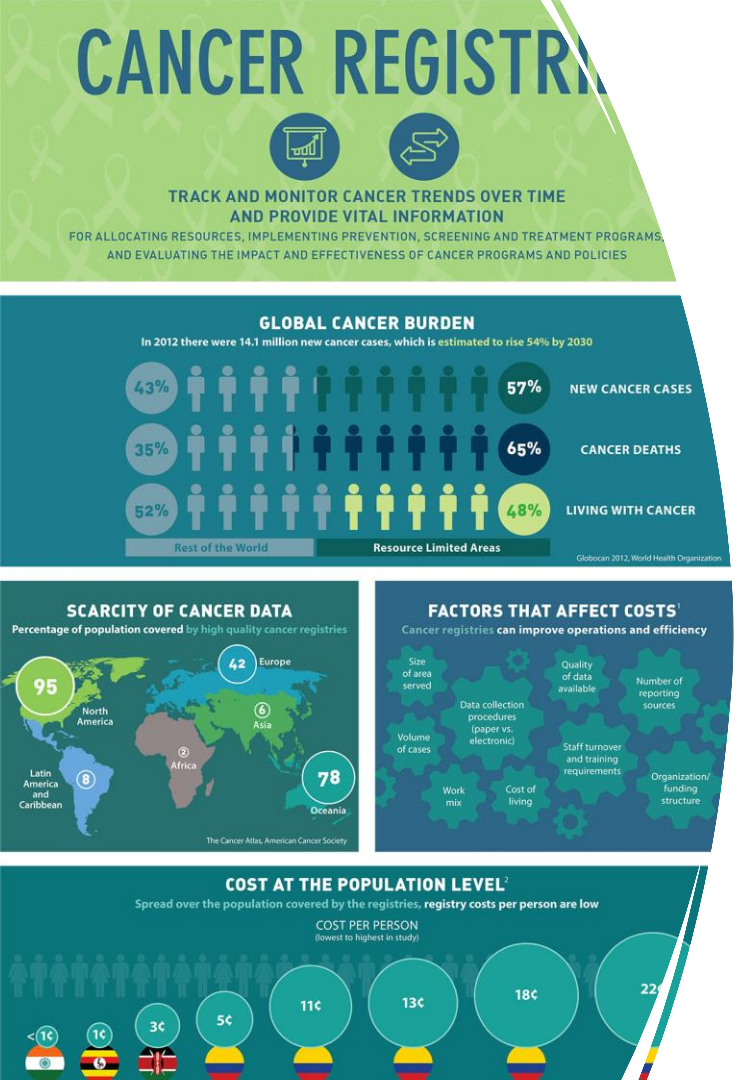




https://pmc.ncbi.nlm.nih.gov/articles/PMC11403302/

Cancer care continuum & surveillance





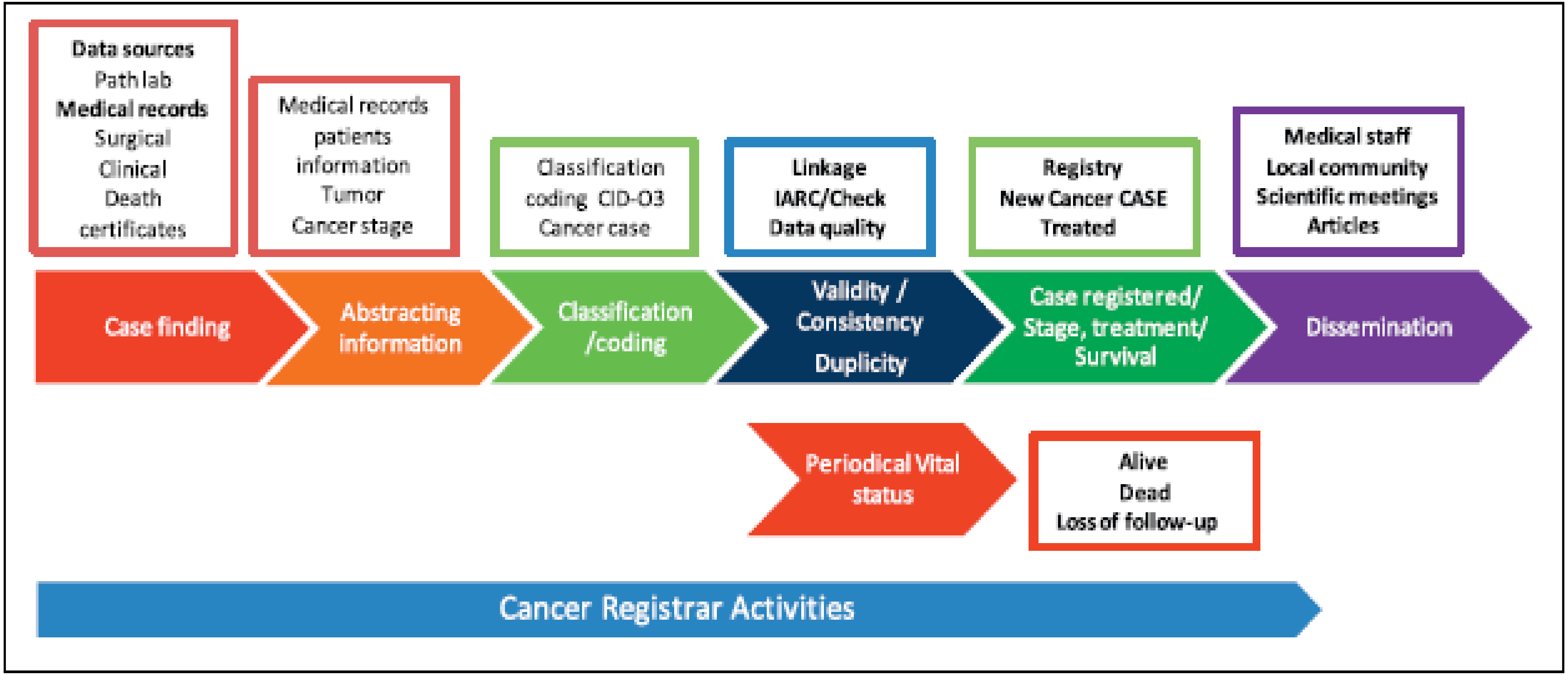
Cancer Registry

- Track trends <u>over time</u> (Incidence, mortality and survival)

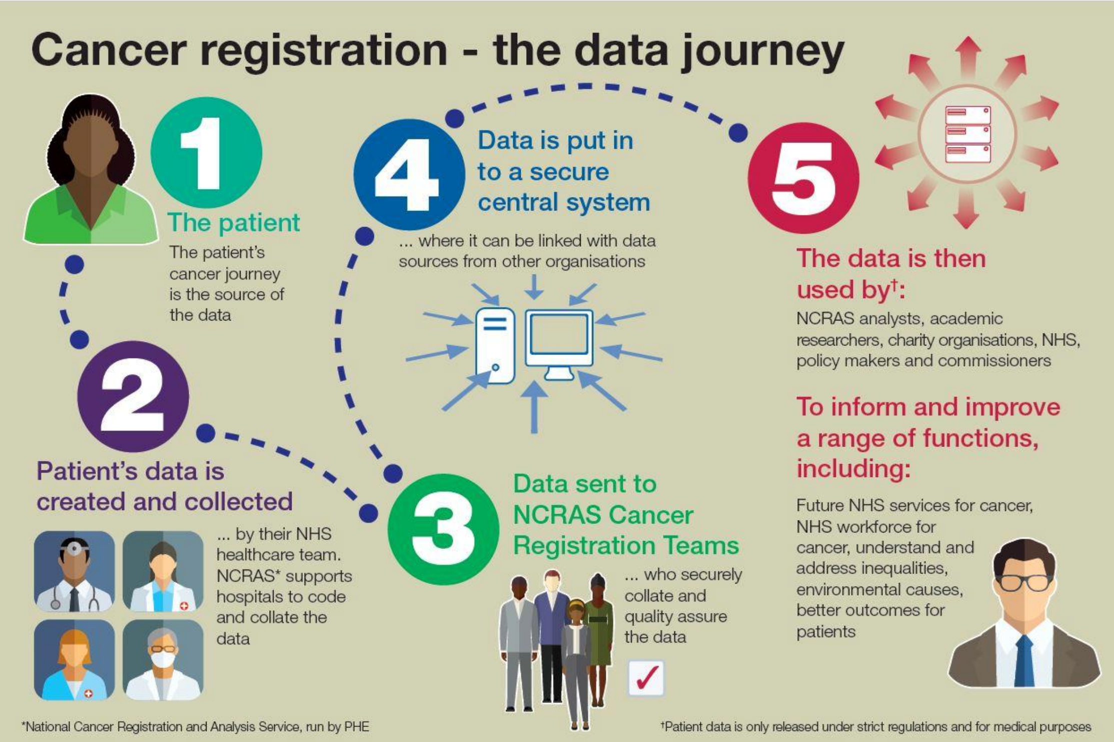
- Allocate resources, prevention, screening and treatment

- Evaluate effectiveness of cancer programs and policies

Cancer Registration



Manual Cancer Registration



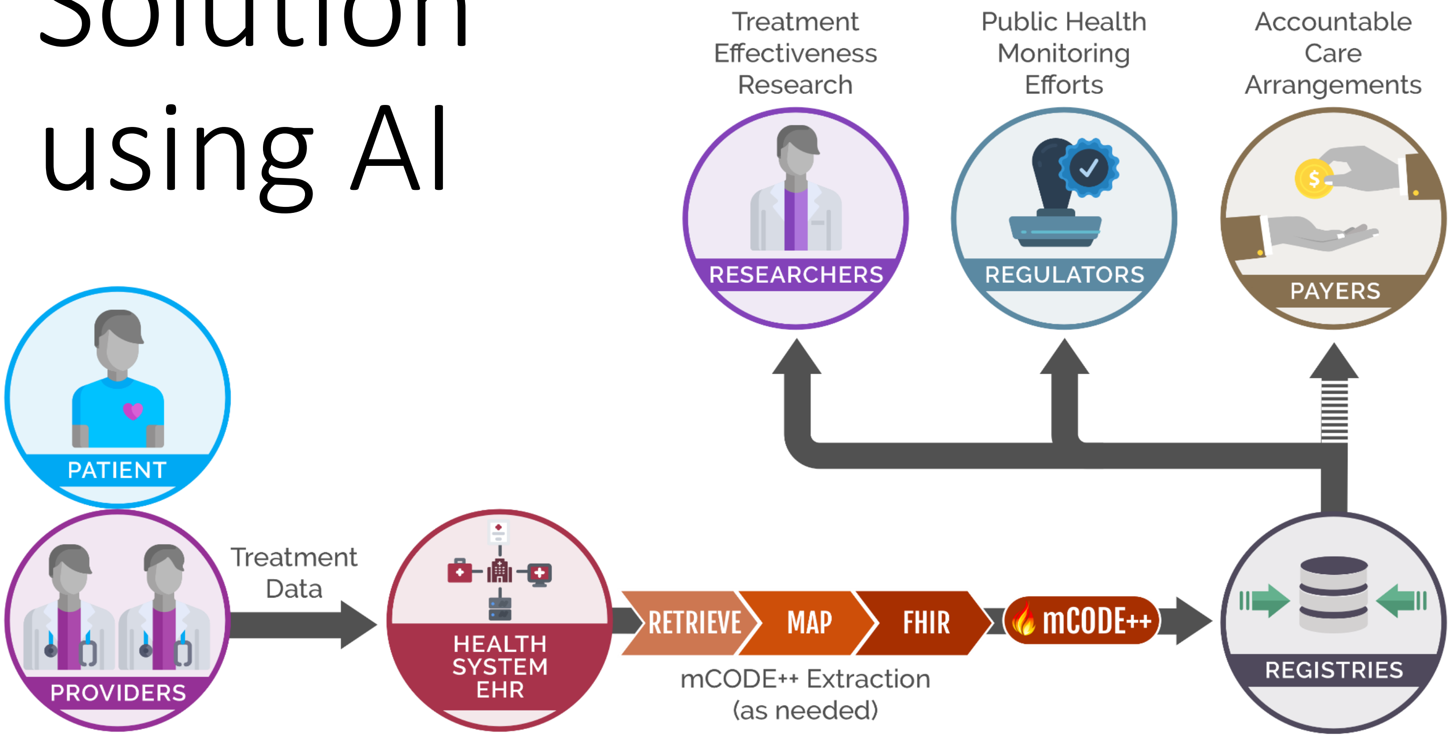
Challenges - Manual abstraction

- Expensive
- Prone to errors
- Affect quality, completeness,
- Accuracy and timeliness data
- Un-sustainable

Manual abstraction --→ delayed reporting

Cancer incidence reports are often not available until **<u>24 months</u>** or greater after a diagnosis

Solution using AI



NIH NATIONAL CANCER INSTITUTE
Computational Resources for Cancer Research

Search

About | Computational Resources | Publications | Connect & Learn | Contact

Modeling Outcomes Using Surveillance Data and Scalable Artificial Intelligence for Cancer (MOSSAIC)

Applies natural language processing (NLP) and deep learning algorithms to population-based cancer data

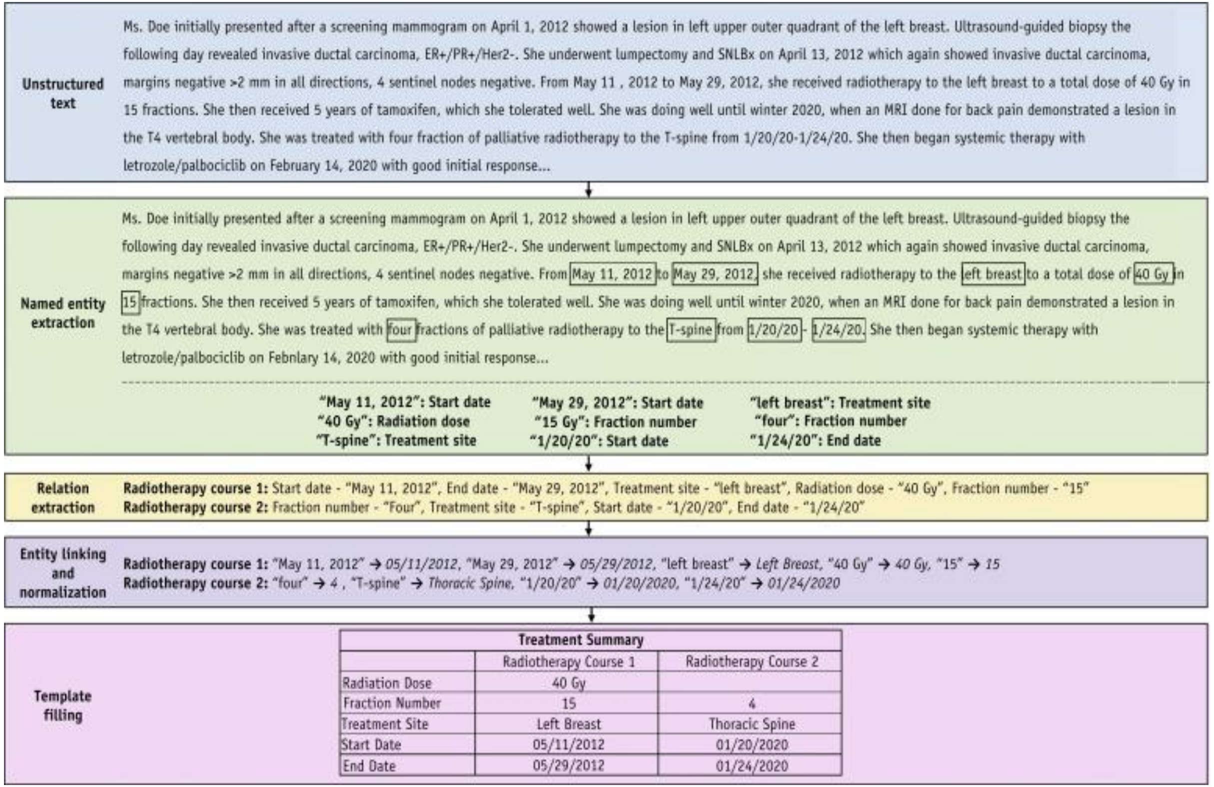
To develop scalable NLP tools for deep text comprehension of unstructured clinical text

To enable automated and accurate capture of reportable cancer surveillance data elements

World Summit AI
جامعة السلطان قابوس
Sultan Qaboos University

Solutions - Automate Data collection using ML & NLP

Unstructured data



Treatment Summary		
	Radiotherapy Course 1	Radiotherapy Course 2
Radiation Dose	40 Gy	
Fraction Number	15	4
Treatment Site	Left Breast	Thoracic Spine
Start Date	05/11/2012	01/20/2020
End Date	05/29/2012	01/24/2020

Clinical text context is important

Present: default category
Patient had a stroke

Absent: problem does not exist in the patient
History inconsistent with stroke

Possible: uncertainty expressed
We are unable to determine whether she has leukemia

Conditional: patient experiences the problem only under certain conditions
Patient reports shortness of breath upon climbing stairs

Hypothetical: medical problems the patient may develop
If you experience wheezing or shortness of breath

Corresponds to
SHARPn conditiona

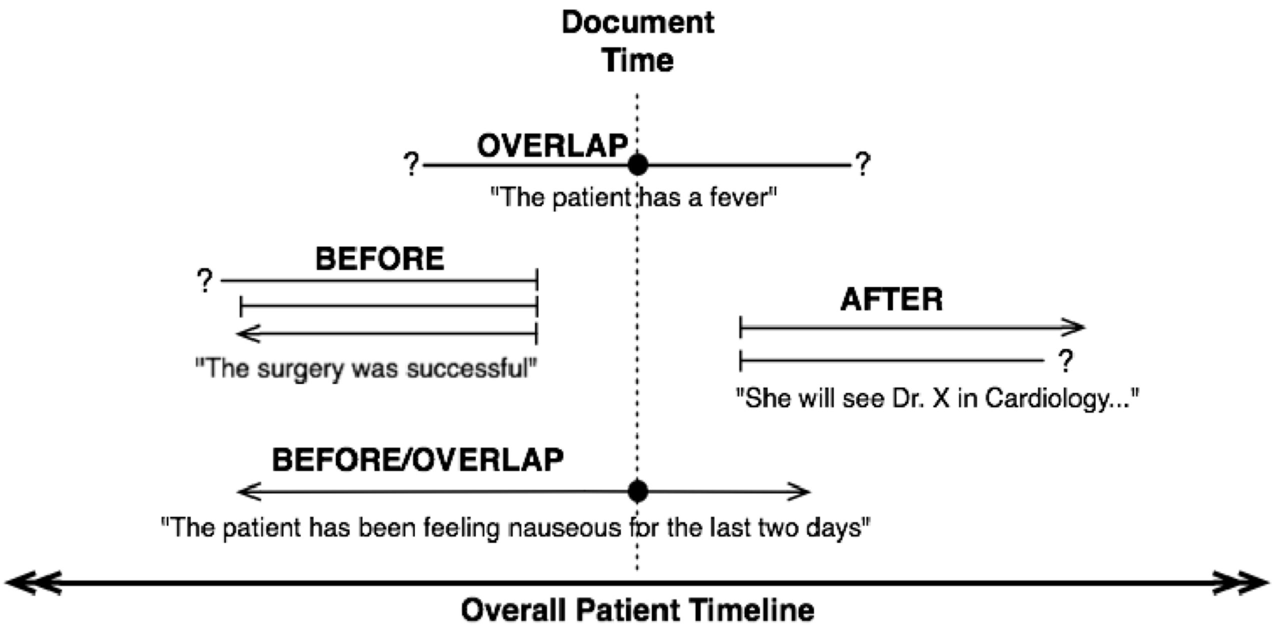
Not Patient: problem associated with someone who is not the patient
Family history of prostate cancer

The Challenge: Text Mentions versus Clinical Facts

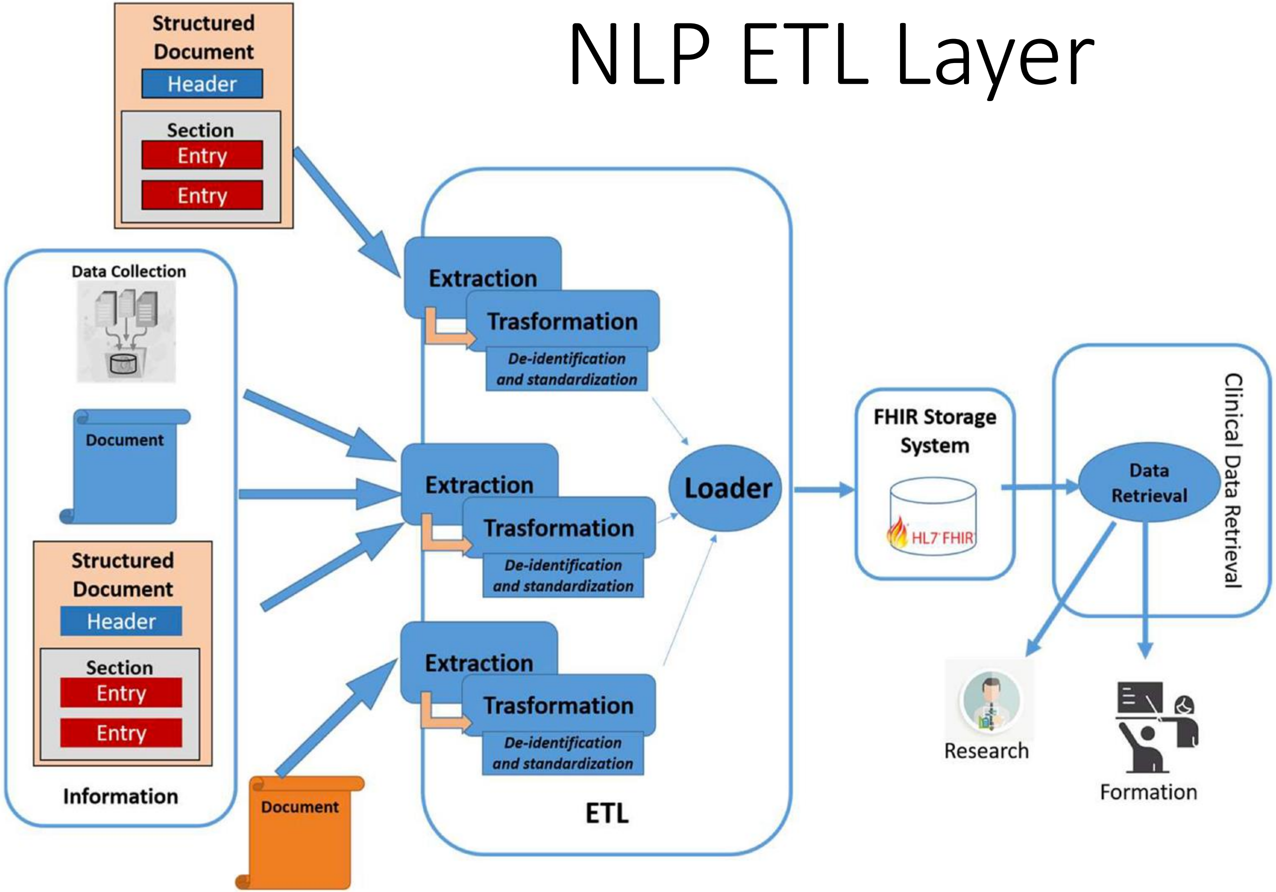
- **Negation**: event has not occurred or entity does not exist
  She had no fever yesterday.
- **Uncertainty:** a measure of doubt
  The symptoms are not inconsistent with renal failure.
- **Conditional**: could exist or occur under certain circumstances
  The patient should come back to the ED if any rash occurs.
- **Subject**: person the observation is on; experiencer
  Mother had lung cancer.
- ***Generic***: no clear subject/experience
  E. coli is sensitive to Cipro but enter

fever	no
renal infarction	uncertain
rash	conditional
lung cancer	family member
Cipro	generic
…	

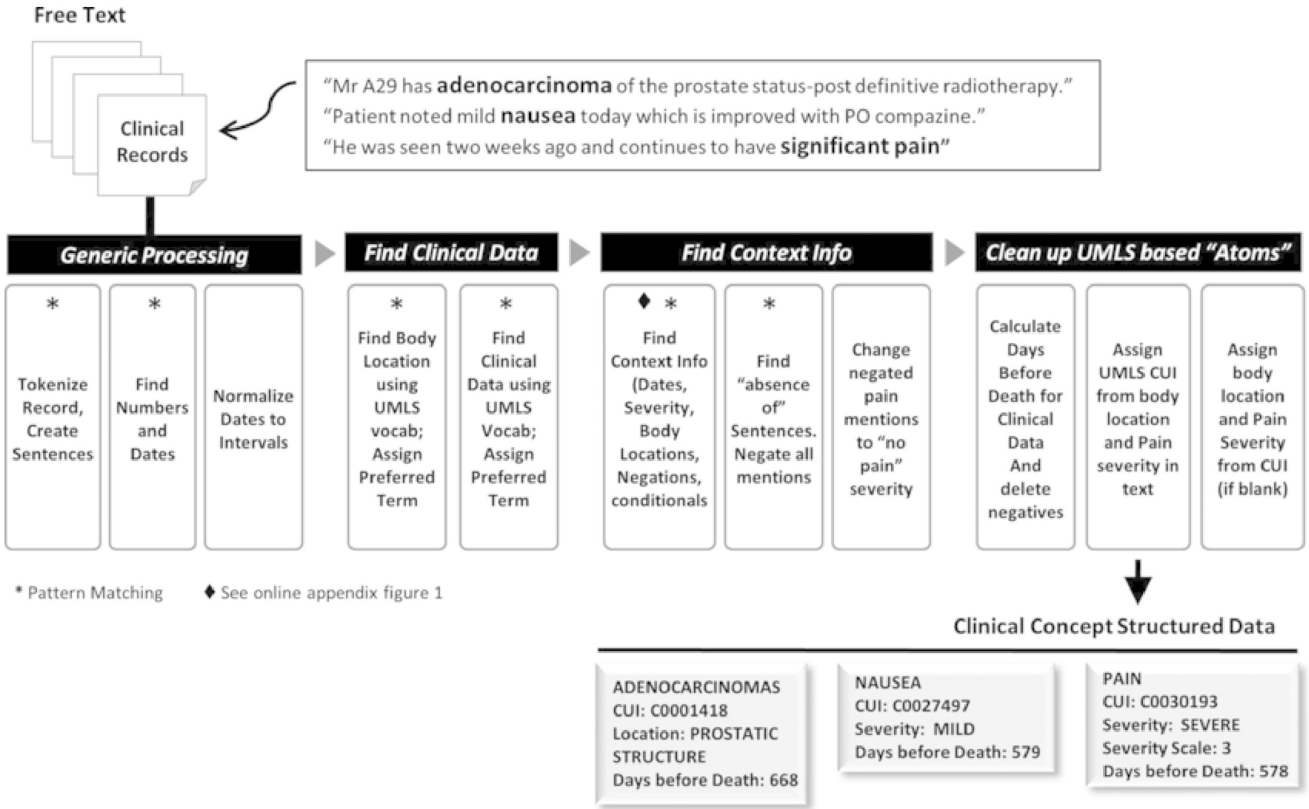
Clinical text is temporal



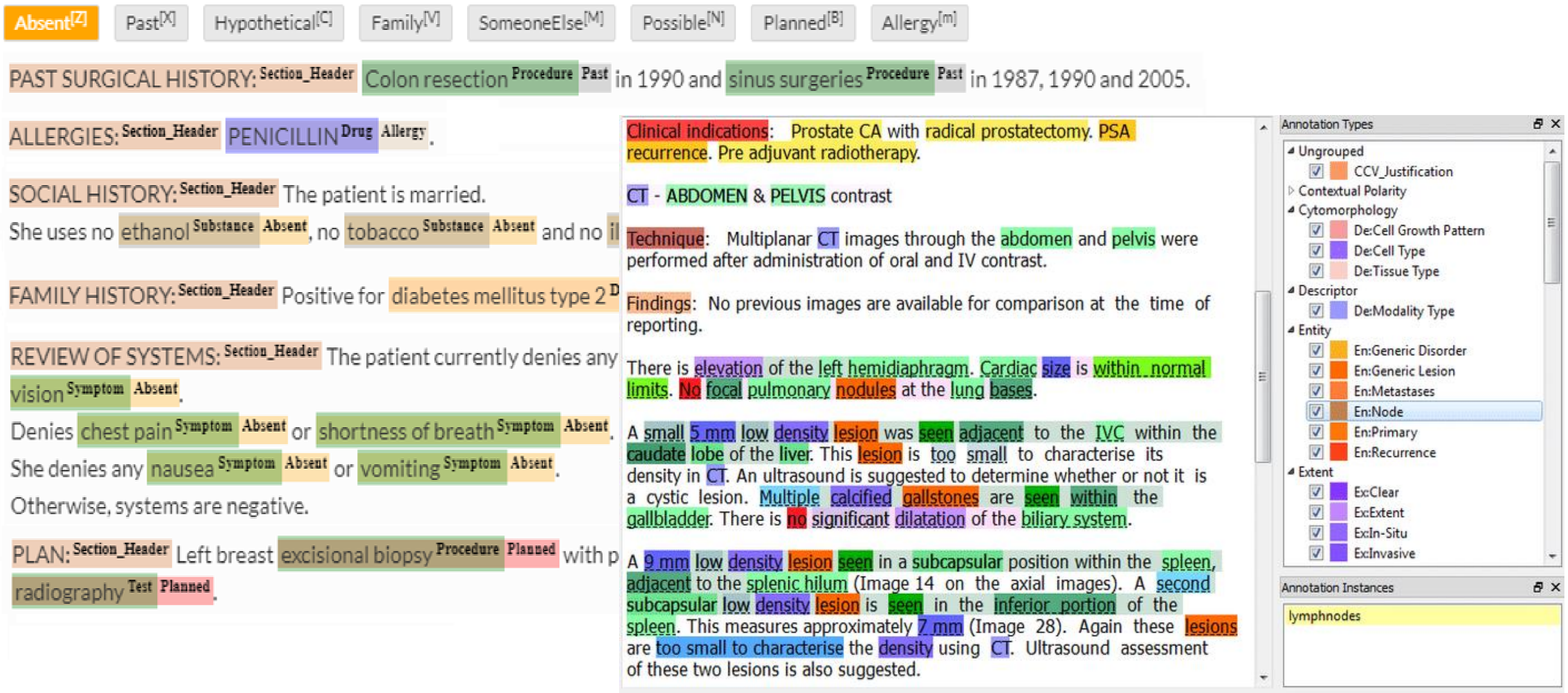
NLP ETL Layer



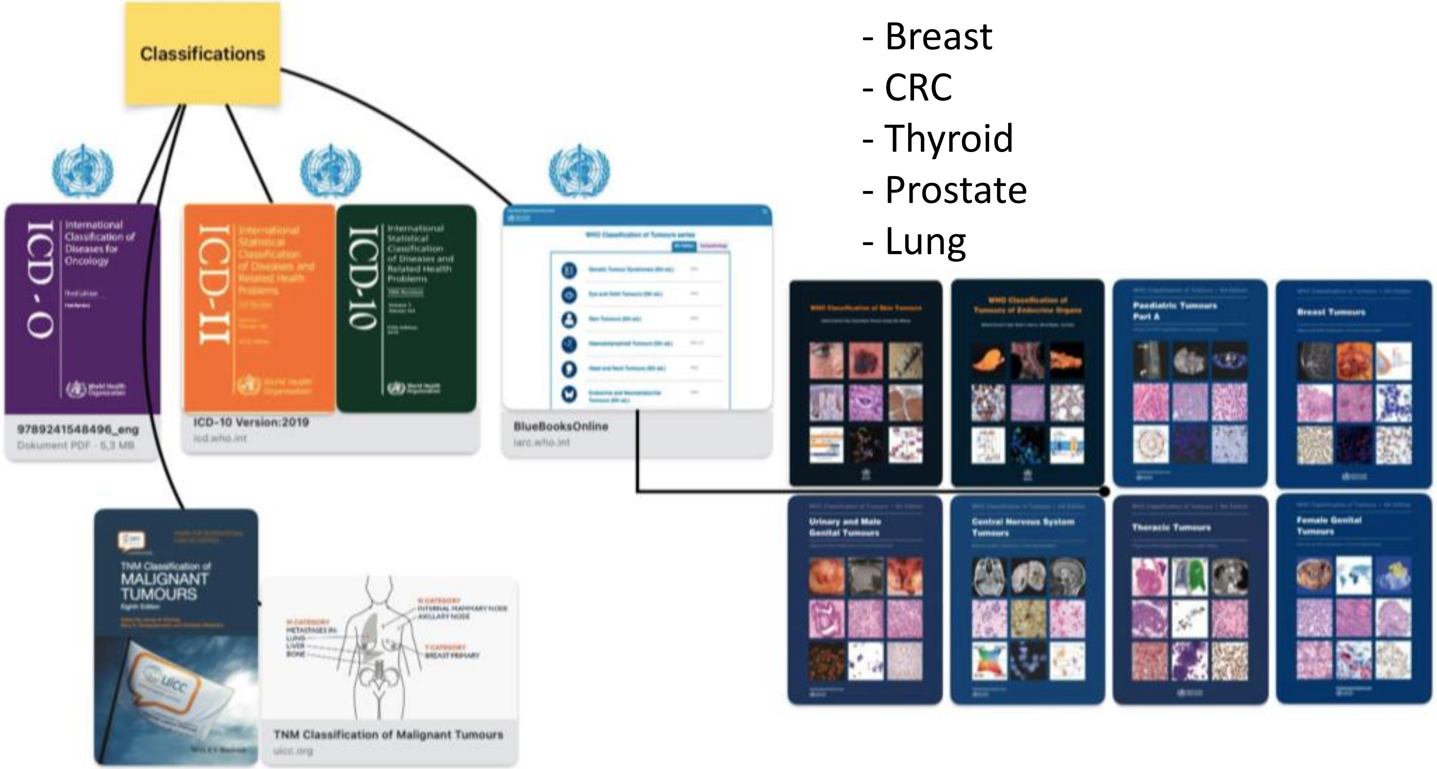
Methods - NLP Pipeline



Methods - Annotation or labeling (Gold standard)



Choosing standard dataset: classifications



- Breast
- CRC
- Thyroid
- Prostate
- Lung

Objective 1

Consoloditate TNM staging from

Clinical text

Clinical text to TNM staging

“TUMOR INVADES INTO BUT NOT THROUGH VISCERAL PLEURA” => stage T2

“8 LYMPH NODES NEGATIVE FOR TUMOR” => stage N0

Dataset

	Lung	Colon	Prostate	Total
Training	1365	1228	1540	4133
Validation	194	178	221	593
Testing	394	354	441	1189
Total	1953	1760	2202	5915

Results (TNM document-level)

Table 5: Evaluation with the test set.

Evaluation Method	System	TNM mentions			Pathological/Clinical		
		Precision	Recall	F1-measure	Precision	Recall	F1-measure
Strict match	REGEX	0.890	0.884	0.887	0.370	0.368	0.369
	CRF	0.923	0.845	0.882	0.810	0.742	0.774
	REGEX-CRF	0.890	0.884	0.887	0.779	0.774	0.777
Partial match	REGEX	0.961	0.955	0.958	0.386	0.384	0.385
	CRF	0.989	0.906	0.946	0.873	0.800	0.835
	REGEX-CRF	0.961	0.955	0.958	0.841	0.835	0.838

World Summit AI

جامعة السلطان قابوس
Sultan Qaboos University

Results (Patient-level)

Classifier	Site	TNM	Agreement (%)
Baseline	(All)	(All)	2358/3567 (66.1%)
Baseline	(All)	M	871/1189 (73.3%)
Baseline	(All)	N	779/1189 (65.5%)
Baseline	(All)	T	708/1189 (59.5%)
Baseline	Colon	(All)	810/1062 (76.3%)
Baseline	Colon	M	280/354 (79.1%)
Baseline	Colon	N	297/354 (83.9%)
Baseline	Colon	T	233/354 (65.8%)
Baseline	Lung	(All)	593/1182 (50.2%)
Baseline	Lung	M	222/394 (56.3%)
Baseline	Lung	N	196/394 (49.7%)
Baseline	Lung	T	175/394 (44.4%)
Baseline	Prostate	(All)	955/1323 (72.2%)
Baseline	Prostate	M	369/441 (83.7%)
Baseline	Prostate	N	286/441 (64.9%)
Baseline	Prostate	T	300/441 (68.0%)
Linear SVM	(All)	(All)	2958/3567 (82.9%)
Linear SVM	(All)	M	1138/1189 (95.7%)
Linear SVM	(All)	N	920/1189 (77.4%)
Linear SVM	(All)	T	900/1189 (75.7%)
Linear SVM	Colon	(All)	960/1062 (90.4%)
Linear SVM	Colon	M	341/354 (96.3%)
Linear SVM	Colon	N	323/354 (91.2%)
Linear SVM	Colon	T	296/354 (83.6%)
Linear SVM	Lung	(All)	888/1182 (75.1%)
Linear SVM	Lung	M	370/394 (93.9%)
Linear SVM	Lung	N	238/394 (60.4%)
Linear SVM	Lung	T	280/394 (71.1%)
Linear SVM	Prostate	(All)	1110/1323 (83.9%)
Linear SVM	Prostate	M	427/441 (96.8%)
Linear SVM	Prostate	N	359/441 (81.4%)
Linear SVM	Prostate	T	324/441 (73.5%)

Study 1 conclusions

- Consolidation of M stage accuracy = (93%-98%)
- Consolidation of T and N different for each site
    - Colon accuracy: 80-90%
    - Prostate accuracy: 70-80%
    - Lung accuracy: 60-70%

- Colon staging criteria is easier

- 24% of lung cases un-staged due to missing information

Proceedings — AMIA Joint Summits on Translational Science

AMIA Jt Summits Transl Sci Proc. 2018 May 18;2018:16–25.

Automated Extraction and Classification of Cancer Stage Mentions fromUnstructured Text Fields in a Central Cancer Registry

Abdulrahman K AAlAbdulsalam [1], Jennifer H Garvin [1,3], Andrew Redd [2], Marjorie E Carter [3], Carol Sweeny [3], Stephane M Meystre [4]

Author information Article notes Copyright and License informatio

PMCID: PMC5961766 PMID: 29888032

Abstract

Cancer stage is one of the most important prognostic para
subtypes. The American Joint Com-mittee on Cancer (AJC
each cancer type based on tumor characteristics (T), lymph

Machine Learning to Automate Cancer Stage Consolidation in a Central Cancer Registry

Abdulrahman AAlAbdulsalam[a,*], Jennifer H. Garvin, MBA, PhD[a], Andrew Redd, PhD[b], Kimberly Herget[c], Marjorie E. Carter, MS[c], Carol Sweeny, PhD[c], Stephane M. Meystre[d]

[a] *Biomedical Informatics, University of Utah, Salt Lake City, UT*
[b] *Epidemiology, University of Utah, Salt Lake City, UT*
[c] *Utah Cancer Registry, University of Utah, Salt Lake City, UT*
[d] *Medical University of South Carolina, Charleston, SC*

ABSTRACT

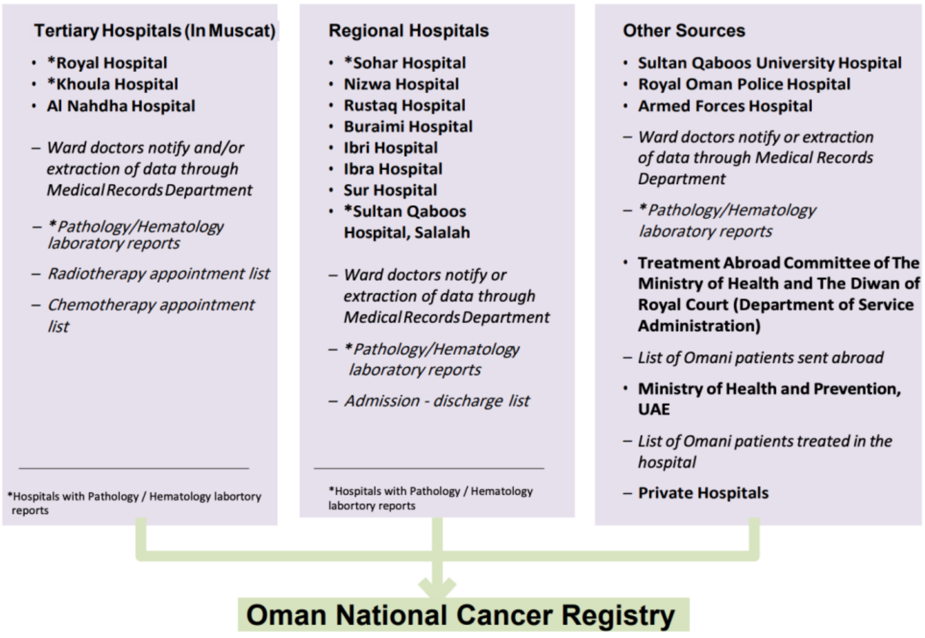
Background Consolidating cancer stage from multiple records is one of the primary tasks performed by central cancer registries. Team of certified tumor registrars (CTR) conduct the consolidation by manually reviewing records received from multiple sources for each newly diagnosed cancer case. The large volume of cases handled by central registries and the complexity of staging guidelines make staging one of the barriers to reducing the time delay between diagnosis and reporting for national surveillance data.
Objective Implement and evaluate Natural Language Processing (NLP) and Machine Learning algorithms to automate cancer stage consolidation.
Materials and Methods Records collected at the Utah Cancer Registry (UCR) for patients with colon, lung, or prostate cancers were used for this study. UCR receives multiple

Objective 2

Extraction of cancer registry data from un-structured pathology report

Oman Cancer Registry Data

Figure 3: Data flow into the national cancer registry



SAMPLE OF OMAN NATIONAL CANCER REGISTRY FORM

To : **Directorate General of Health Affairs**
Department of Non-Communicable Disease
Surveillance & Control (DNCD)

DGHA, Muscat
Tel.: 24696187, Fax : 24695480
E-mail : dep-ncd@moh.gov.om

1. Cancer Registry No. (**To be filled by the Cancer Registry**)
2. Date of Registration (**To be filled by the Cancer Registry**) D D M M Y Y Y Y
3. Patient Hospital File No.
4. Hospital Name
5. Department of
6. Civil ID.
7. First Name
8. Father's Name
9. Grandfather's Name
10. Tribe Name
11. Sex — 1=M, 2=F, 9=Unknown
12. Marital Status — 1=Single, 2=Married, 3=Divorced, 4=Widowed, 9=Unknown
13. Age or D. of Birth — D D M M Y Y Y Y
14. Nationality — 1= Omani, 2= Non-Omani (Specify) ...., 9= Unknown
15. Country of Birth ....
16. Religion — 1= Muslim, 2= Christian, 3= Hindu, 4= Jewish, 5= Others, 9= Not Known
17. Ethnic Group — 1= Arab, 2= Asian, 3= Caucasian, 4= Other, 9= Not Known
18. Occupation ....

Patient's Address

19. Telephone — Land Line / Mobile
20. Other contact's Tel. No.
21. Wilayat : ....
22. Village : ....
23. Date of First Diagnosis — D D M M Y Y Y Y
24. If Pt had Biopsy give Lab. Biopsy Specimen No.
25. Primary Site of Cancer...................
26. Histological type of Cancer ....

ICD-0-3 Code
(ICD-0-3 Code will be filled by the Central Registry)

27. Laterality
    - 0=Not a Paired Site, Unknown or ill-defined Site
    - 1=Right : Origin of Primary
    - 2=Left : Orgin of Primary
    - 3=Bilateral Involvement
    - 9=Paired Site, Laterality Unknown
28. Extent of Disease
    - 0=In Situ
    - 1=Localized
    - 2=Regional by Direct Extension
    - 3=Regional to Lymph Nodes
    - 4=Regional, (both 2,3)
    - 5=Regional, (not specified)
    - 7=Distant Metastasis/Systemic Disease
    - 8=Not applicable (e.g. Leukemias)
    - 9=Unstaged, Unknown or Unspecified
29. Histologic Grading
    - 1=Grade I Well differentiated/Differentiated, NOS
    - 2=Grade II Moderately differentiated/Moderately well differentiated/ Intermediate differentiation.
    - 3=Grade III Poorly differentiated
    - 4=Grade IV Undifferentiated anaplastic
    - 5=T - Cell
    - 6=B. Cell
    - 7=Null Cell
    - 8=Killer Cell
    - 9=Grade or differentiation not determined, not stated or not applicable
    - 10= Other's ....

TNM Classification (Clinical)

30. T	Tx	To	Tis	Tmic	T1	T2	T3	T4
	1	2	3	4	5	6	7	8

31. N	Nx	N0	N1	N2	N3
	1	2	3	4	5

32. M	Mx	M0	M1
	1	2	3

33.M Not applicable

34. Clinical (Stage)	I	Ia	Ib	Ic	II	IIa	IIb	IIc	III	IIIa	IIIb	IIIc	IV	IVa	IVb	IVc	IVd	99= Un-Known, Un-Specified, Not Applicable
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	99

35. Most valid basis of Diagnosis

Non-Microscopic
- 0=Death certificate only
- 1=Clinical only
- 2=Clinical investigation (eg. X-ray, Isotopes)
- 3=Exploration surgery but without Histology
- 4=Specific Biochemical and/or immunological tests

Microscopic
- 5=Cytological/Haematology
- 6=Histology of primary
- 7=Histology of Metastasis
- 8=Autopsy
- 9=Not known

36. Treatment
    - 1=Surgery 1 Y 2 N
    - 2=Radiotherapy 1 Y 2 N
    - 3=Chemotherapy 1 Y 2 N
    - 4=Hormonal Therapy 1 Y 2 N
    - 5=Immunotherapy 1 Y 2 N
    - 6=Bone Marrow Transplant 1 Y 2 N
    - 7=Neoadjuvent 1 Y 2 N
    - 8=Supportive/palliative 1 Y 2 N
    - 9=Other Treatment 1 Y 2 N
    - 10=Not Known 1 Y 2 N
37. Date of Death — D D M M Y Y Y Y
38. Cause of Death — 1=Cancer or Cancer related, 3=Unrelated to Cancer, 9=Not known
39. Date of Last Contact — D D M M Y Y Y Y
39. Source of information — 1 = Medical File, 2 = Death Certificate, 3 = Other, 9 = Unknown

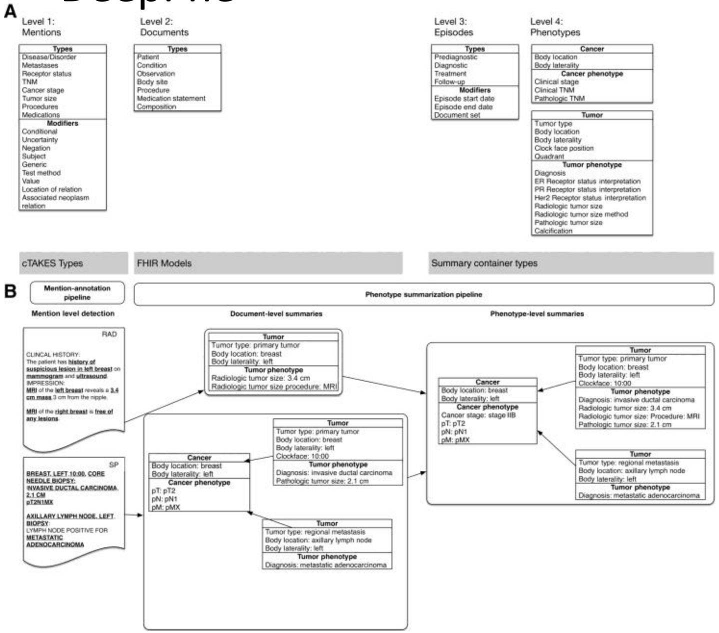
Doctor's Name :
Doctor's Designation :
Doctor's Signature :
Departmemt :
Date :

Tumor Registrar :
(To be Filled by the Cancer Registry)
Coding :...................
Entry : ....

1. Send White copy to NCD Section Fax :24695480 2. Keep Pink Copy in Patient's Case Notes (File) 3. Send Blue Copy to Medical Records Dept.

Methods - NLP

DeepPhe

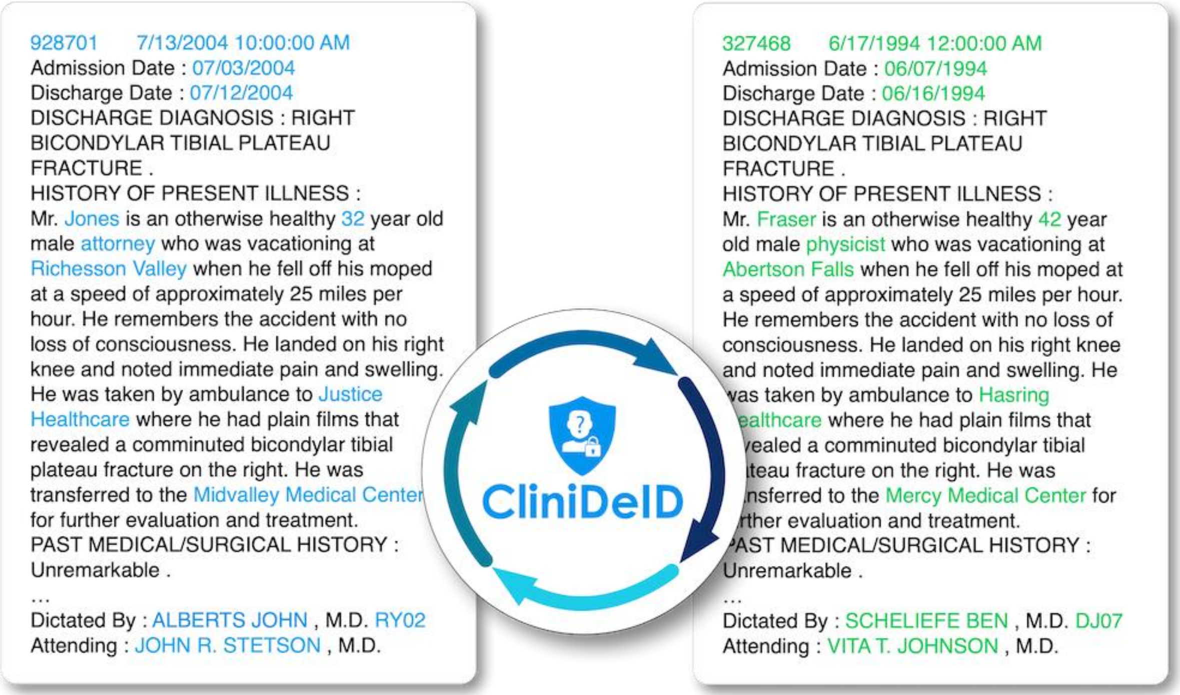


REGEX

Regular expression rules to match directly text mentions of

1. cancer primary site
2. Histology
3. Grade
4. pathological TNM stage
5. summary stage

Methods - Clinical Text De-identification



Results

		Primary Site		Histological type of Cancer		Laterality		Grade		T		N		M		Summary (Stage)	
		Royal	SQUH	Royal	SQUH	Royal	SQUH	Royal	SQUH	Royal	SQUH	Royal	SQUH	Royal	SQUH	Royal	SQUH
REGEX	Prec.	1	1	0.87	0.86	0.83	0.76	0.52	0.65	0.54	0.59	0.59	0.61	0.57	0.6	0.46	0.32
	Recall	0.97	0.99	0.69	0.82	0.69	0.83	0.52	0.57	0.25	0.51	0.26	0.53	0.17	0.51	0.63	0.41
	F1	0.99	1	0.76	0.84	0.75	0.79	0.51	0.53	0.18	0.45	0.2	0.47	0.22	0.53	0.53	0.36
DeepPhe	Prec.	1	1	0.89	0.85	0.82	0.74	0.54	0.34	0.33	0.38	0.69	0.56	0.61	0.54	0.5	0.31
	Recall	0.83	0.3	0.47	0.12	0.62	0.25	0.47	0.14	0.13	0.09	0.14	0.08	0.1	0.07	0.67	0.55
	F1	0.91	0.47	0.61	0.21	0.7	0.37	0.47	0.16	0.13	0.14	0.14	0.13	0.13	0.12	0.55	0.4

Clinical text ambiguity

- TX abbrev. for treatment

DISCUSSED PALLIATIVE TX W/ CARBO/TAXL . . .
NEW LUNG CANCER F/U & TX . . .

- Alpha-numeric terms: T0012-9071, N13-129

- MRI and biomarker references:
“SUBTLE AREA OF FOCAL T2 SIGNAL LOSS”
“weakly postitive for WT1

- Some errors with partial matching of “M1” in middle of words:
AIM140.6 AIM111.1

World Summit AI

جامعة السلطان قابوس
Sultan Qaboos University

2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)

Extraction of Breast Cancer Information from Clinical Record for Cancer Registry using Natural Language Processing

Adhari Abdullah Alzaabi, MD, PhD[1] and Abdulrahman AAlAbdulsalam, PhD[2]*
[1] *College of Health Science and Medicine*

Abstract—National cancer registries r… abstraction of free-text clinical records t… information about cancer diagnosis, stage, p… treatment. Many prior studies have demonstr… of natural language processing (NLP) base… learning to extract information from fr… records for a variety of purposes (diagnosis, … discovery, clinical trial matching, …, etc.). … this study experimental results of applying … information from the records of breast can… the cancer registry in Oman.

Keywords—Clinical information extra… Language Processing, Structured Data, El… Records

INTRODUCTION

Cancer registries are important resource f… of cancer disease in the population and … vital for research and decision-maki… disease control and prevention [1]. Howe…

Machine Learning for Healthcare 2023 – Clinical Abstract, Software, and Demo Track

Natural Language Processing for Automated Extraction of Breast Cancer Information for the Cancer Registry

Adhari Abdullah Alzaabi, MD, PhD[1] and Abdulrahman AAlAbdulsalam, PhD[2]
[1] *College of Health Science and Medicine, Sultan Qaboos University* [2] *Department of Computer Science, Sultan Qaboos University*

Background. National cancer registries rely on manual abstraction of free-text clinical records to collect vital information about cancer diagnosis, stage, progression and treatment [1]. Many prior studies have demonstrated the ability of natural language processing (NLP) based on machine learning to extract information from free-text clinical records for a variety of purposes (diagnosis, adverse events discovery, clinical trial matching, …, etc.) [2]. We present in this study experimental results of applying NLP to extract information from the records of breast cancer patients for the cancer registry in Oman.

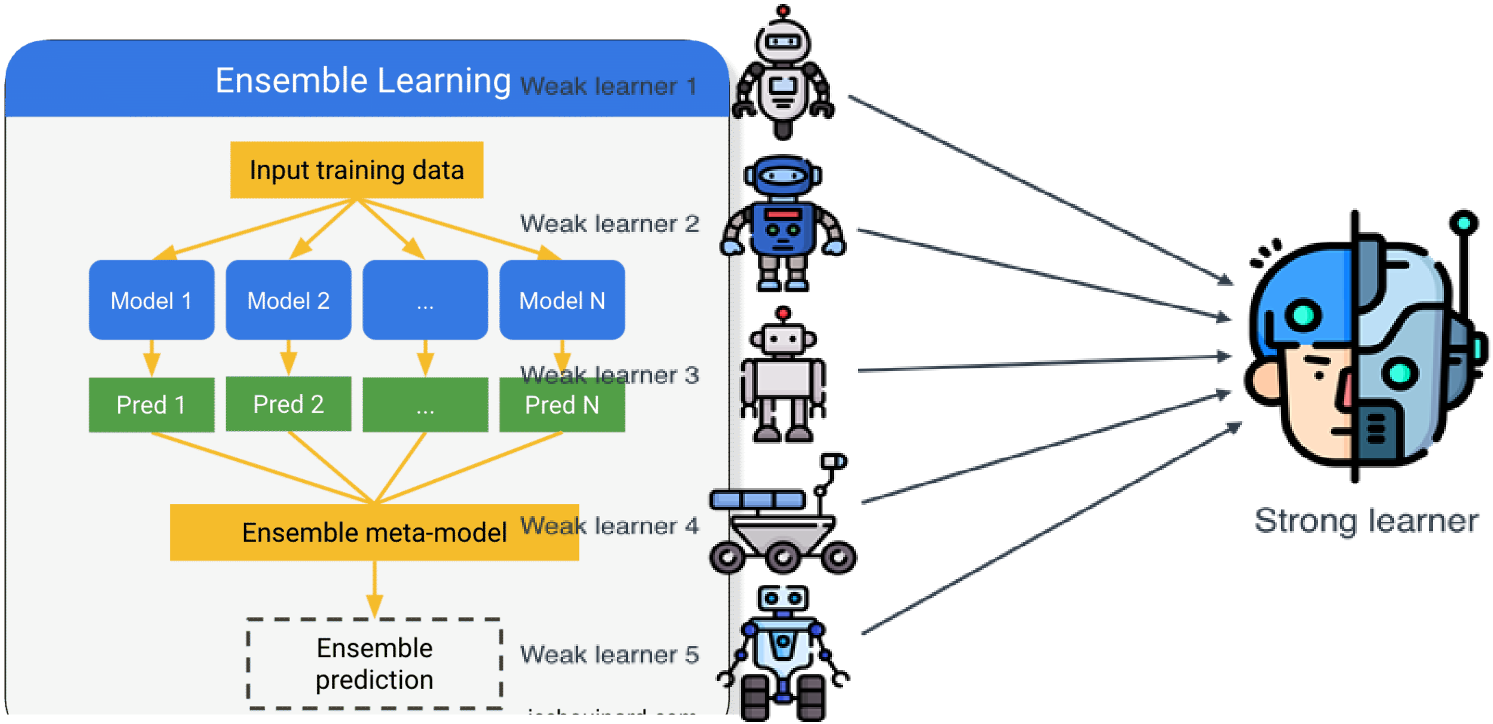
Methods. After obtaining ethical approval from two local institutions (Sultan Qaboos University Hospital and Royal Hospital), the clinical records (pathology, oncology and surgical notes) were collected for 1152 patients (462 from SQUH and 690 from Royal) who have been diagnosed with breast cancer in the years 2013 to 2018. Manually abstracted data within the cancer registry databases for the same patients were extracted to serve as the gold standard to evaluate the NLP approaches. We experimented with two approaches for information extraction from free-text clinical records: 1) using the readily available **DeepPhe** system [3], and 2) rule-based regular expression matching approach (**REGEX**). The precision (positive predictive value), recall (sensitivity) and F1 metrics were used to report the performance of each approach.

Limitations

- **The unavailability of training data:** Cancer reporting requires robust annotated training data that accurately represents the problem space.

- **Frequent changes in coding standards**
- **Clinical guidelines: complicated and overlap**

Methods - ensemble learning



?

Can Large language models (LMs) speed up the process of extracting clinical data from un-structured clinical notes for cancer registry automation?

THANK YOU!

- List of Investigators in the study

  Dr AbdulRahman AlAbdulsalam (Computer Science)
  Dr. Rachid Hedjam (Computer Science)
  Dr. Dr. Najla Al Lawati (Non-communicable diseases, head of cancer regisrtry, MOH)

- The study is funded by generous His-Majesty SR grant

Adhari Abdullah AlZaabi
College of Medicine and Health Sciences
Sultan Qaboos University
adhari@squ.edu.om