



Introduction



Methods

Outline



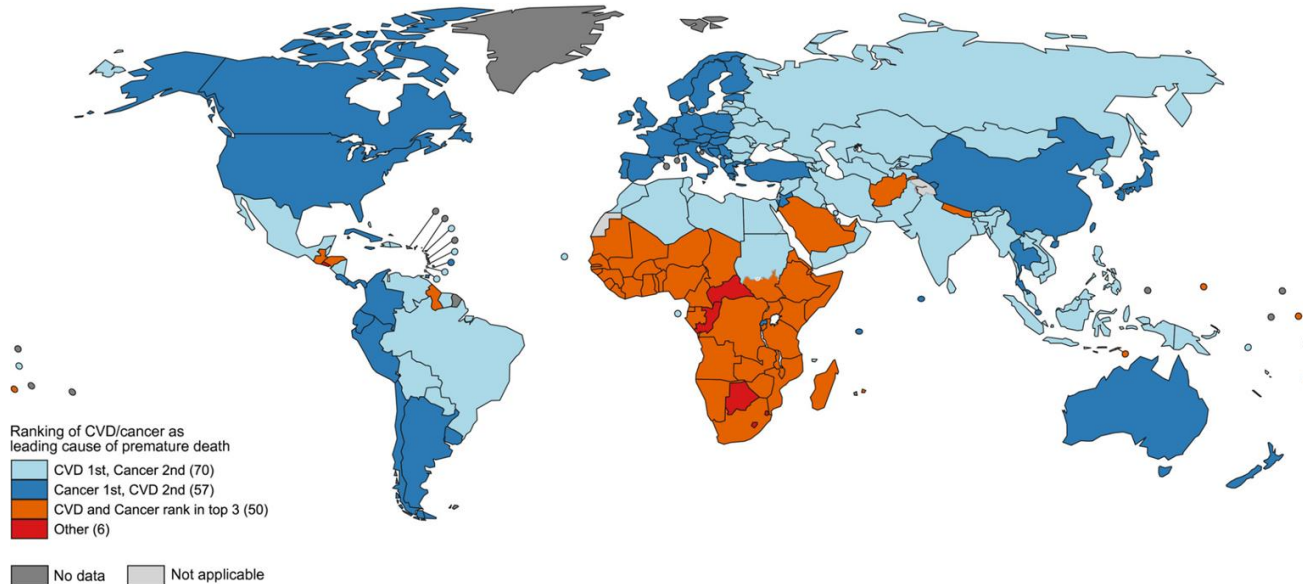
Results



Conclusion

Please, consider this is work in
progress !!!

Cancer Burden (leading cause of death)



The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data source: GHE 2020
Map production: CSU
World Health Organization

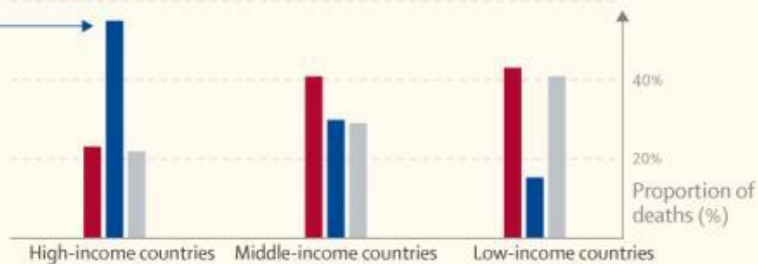
Cancer Burden (leading cause of death)

Cardiovascular disease is the leading cause of death worldwide



100% of deaths globally

But in high-income countries, **cancer** causes twice as many deaths as **cardiovascular disease**

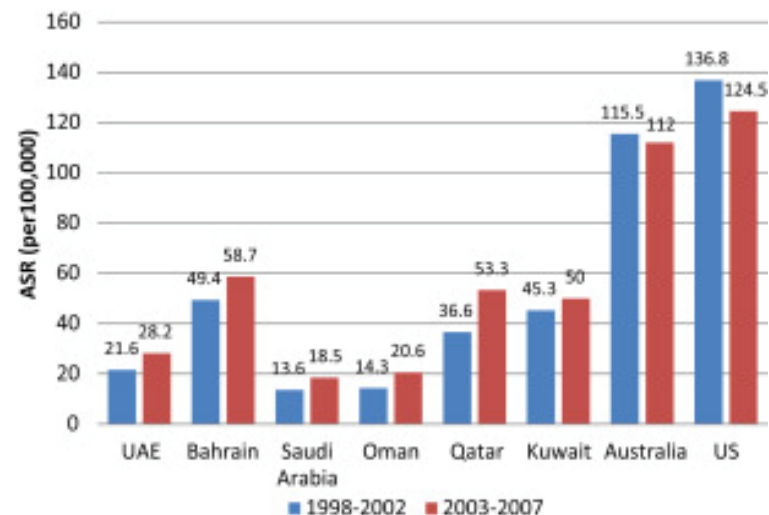
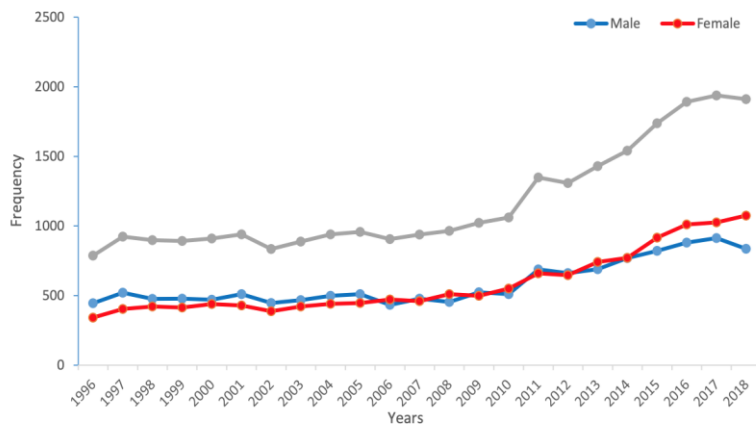


For more, visit www.thelancet.com

- Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE)
- Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study

Oman and Gulf Cooperation Council cancer incidence

Figure 4: Trends of cancer cases 1996-2018



Cancer care continuum & surveillance



CANCER REGISTRY



TRACK AND MONITOR CANCER TRENDS OVER TIME
AND PROVIDE VITAL INFORMATION

FOR ALLOCATING RESOURCES, IMPLEMENTING PREVENTION, SCREENING AND TREATMENT PROGRAMS,
AND EVALUATING THE IMPACT AND EFFECTIVENESS OF CANCER PROGRAMS AND POLICIES

Cancer Registry

- Track trends over time (Incidence, mortality and survival)
- Allocate resources, prevention, screening and treatment
- Evaluate effectiveness of cancer programs and policies

GLOBAL CANCER BURDEN

In 2012 there were 14.1 million new cancer cases, which is estimated to rise 54% by 2030



SCARCITY OF CANCER DATA

Percentage of population covered by high quality cancer registries



FACTORS THAT AFFECT COSTS¹

Cancer registries can improve operations and efficiency



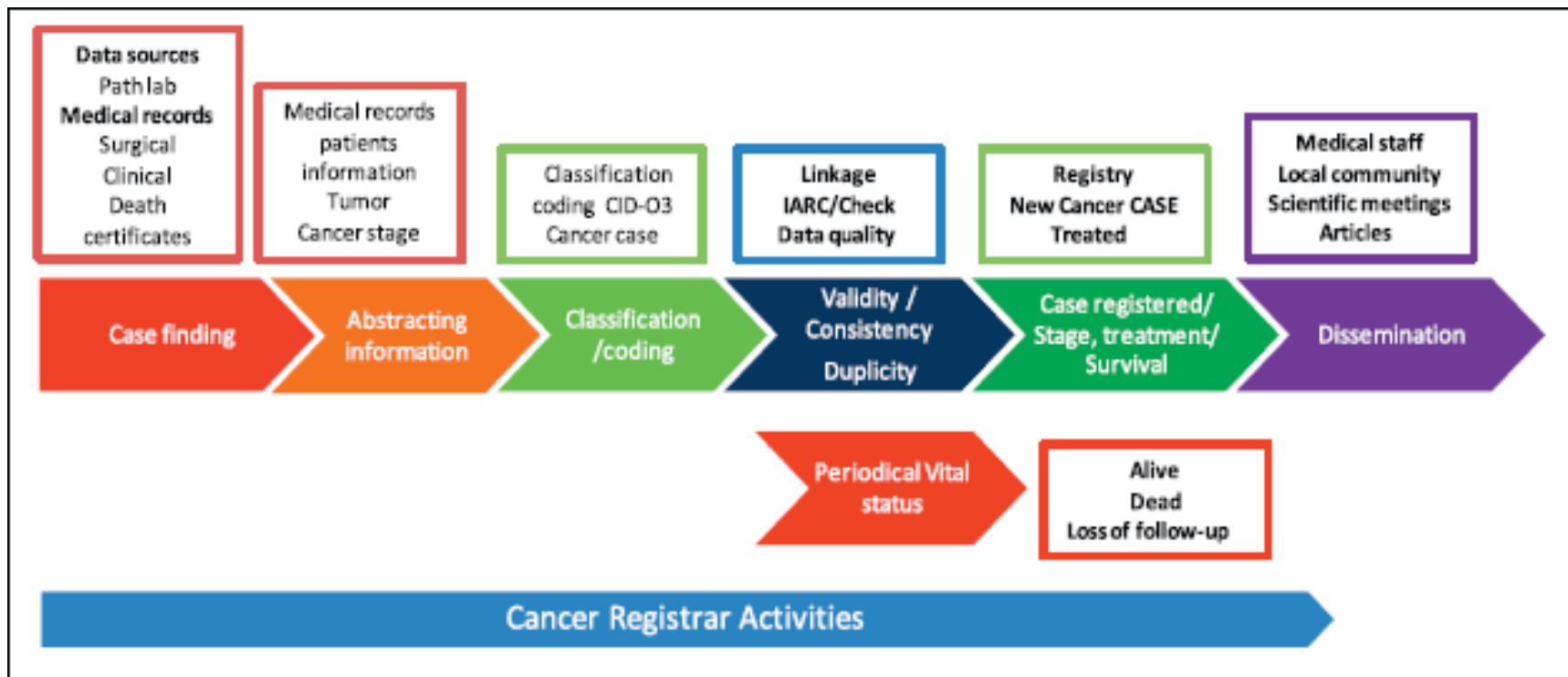
COST AT THE POPULATION LEVEL²

Spread over the population covered by the registries, registry costs per person are low

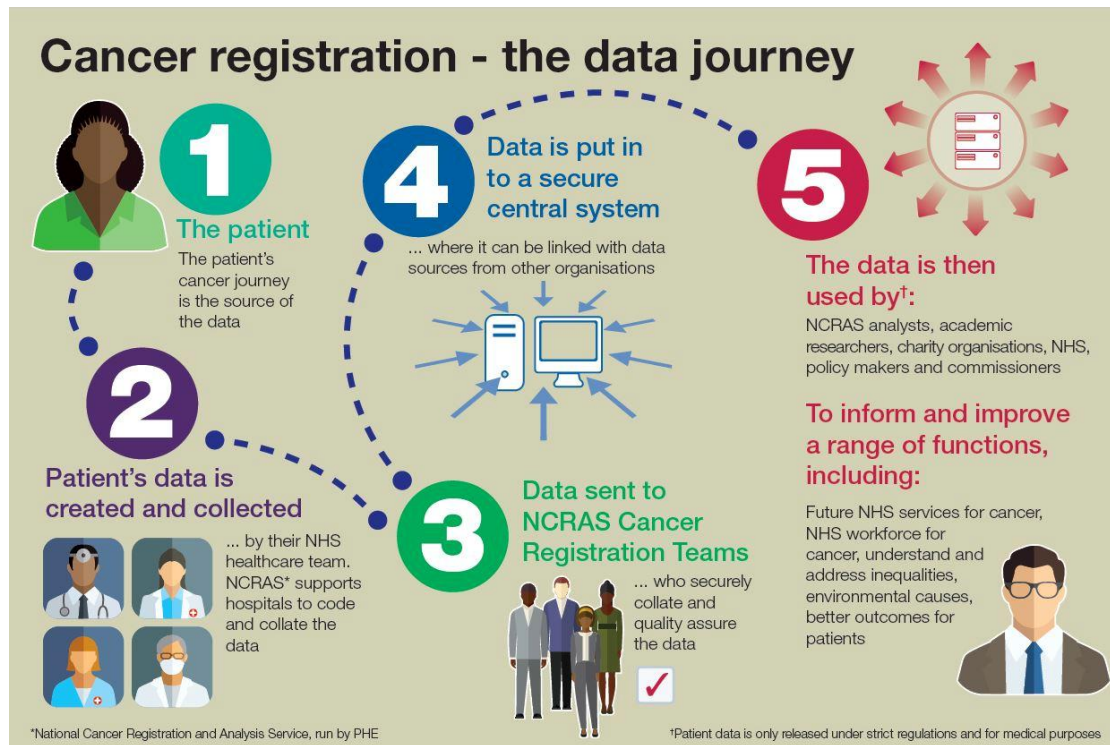
COST PER PERSON
(lowest to highest in study)



Cancer Registration



Manual Cancer Registration



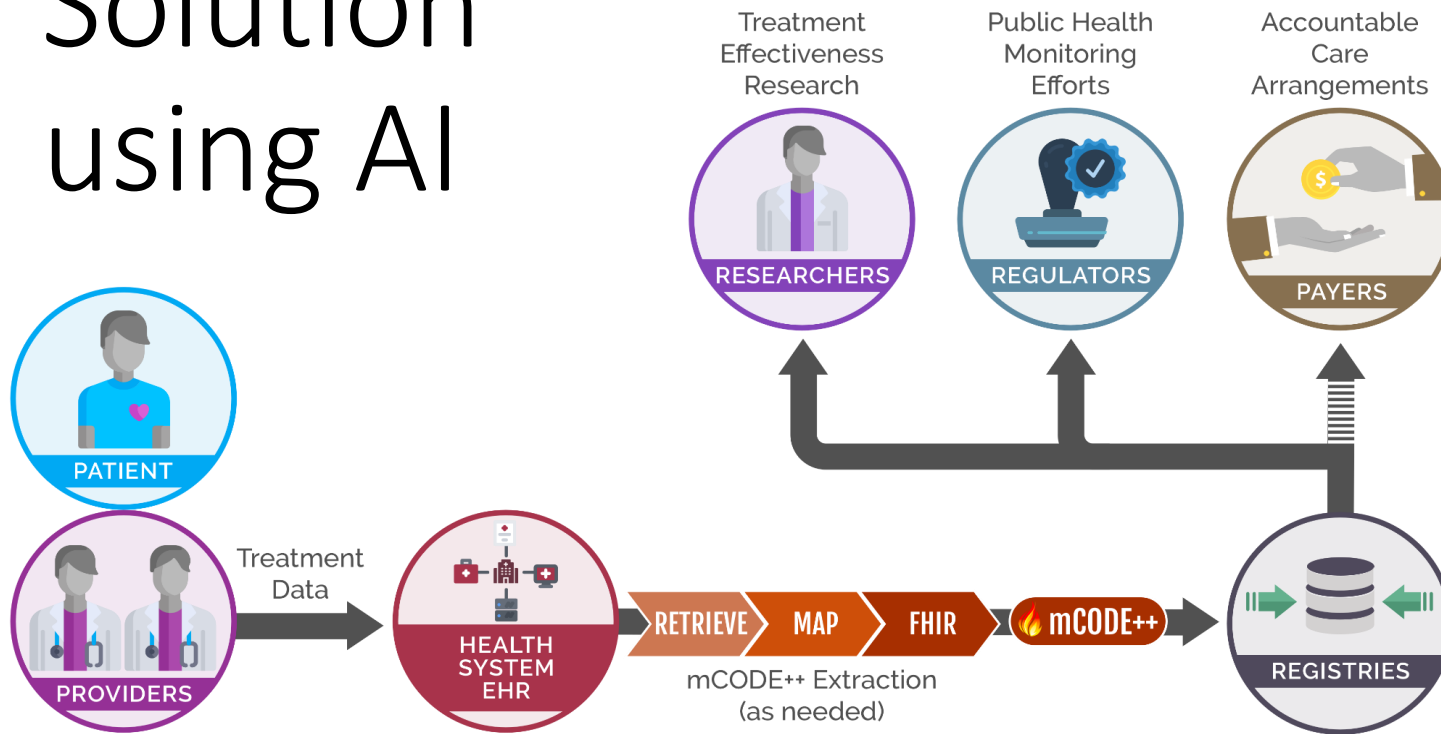
Challenges - Manual abstraction

- 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 **24 months** or greater after a diagnosis

Solution using AI



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

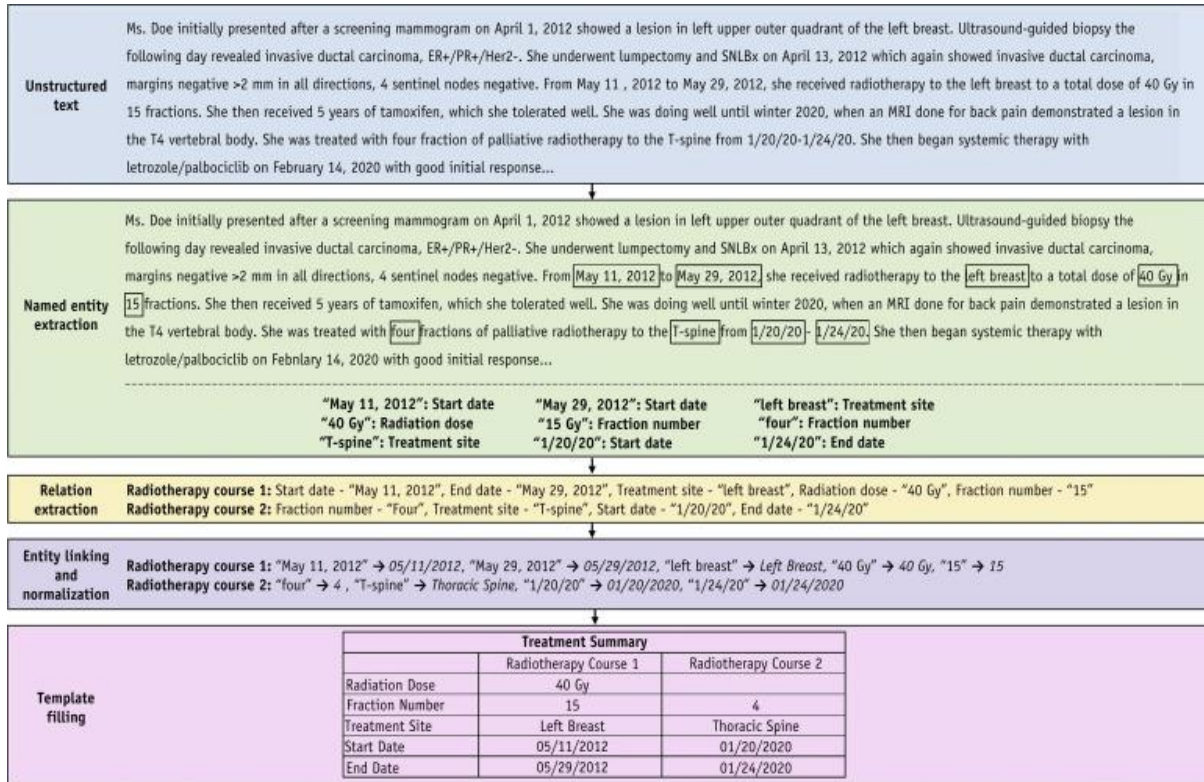
Applies natural language processing (NLP) and deep learning algorithms to population-based cancer data

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

Unstructured data

Solutions - Automate Data collection using ML & NLP



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

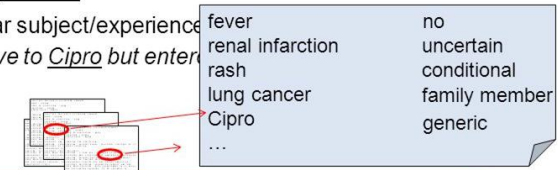
Corresponds to SHARPN conditions

Not Patient: problem associated with someone who is not the patient

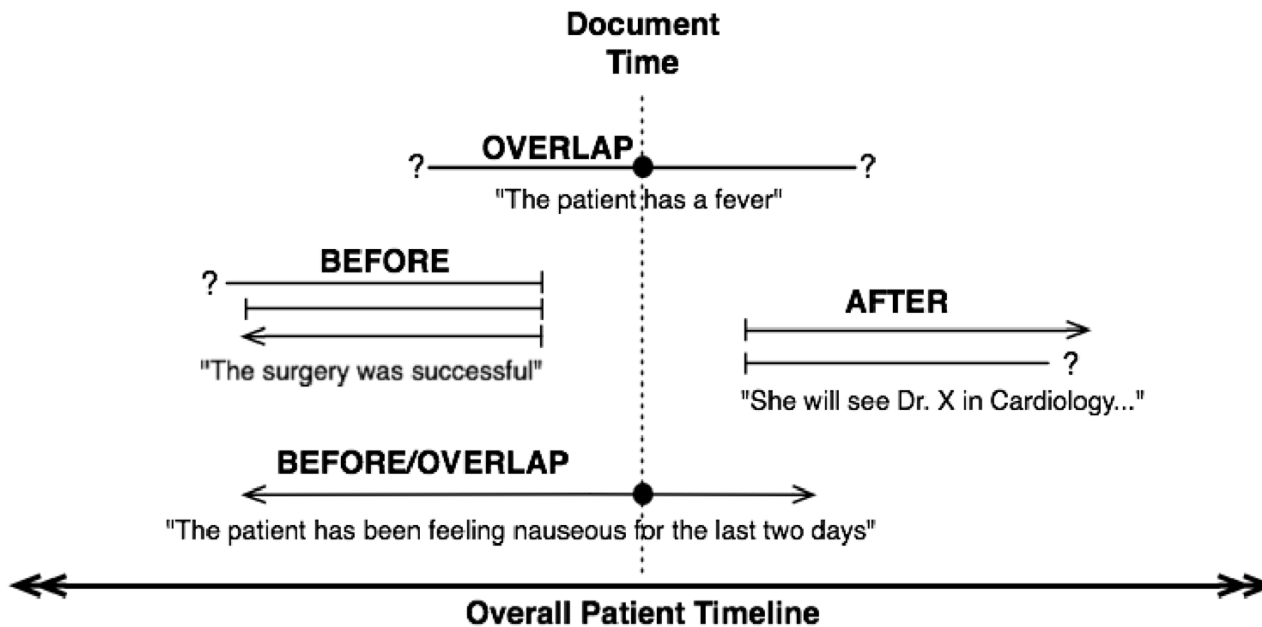
Family history of prostate cancer

The Challenge: Text Mentions versus Clinical Facts

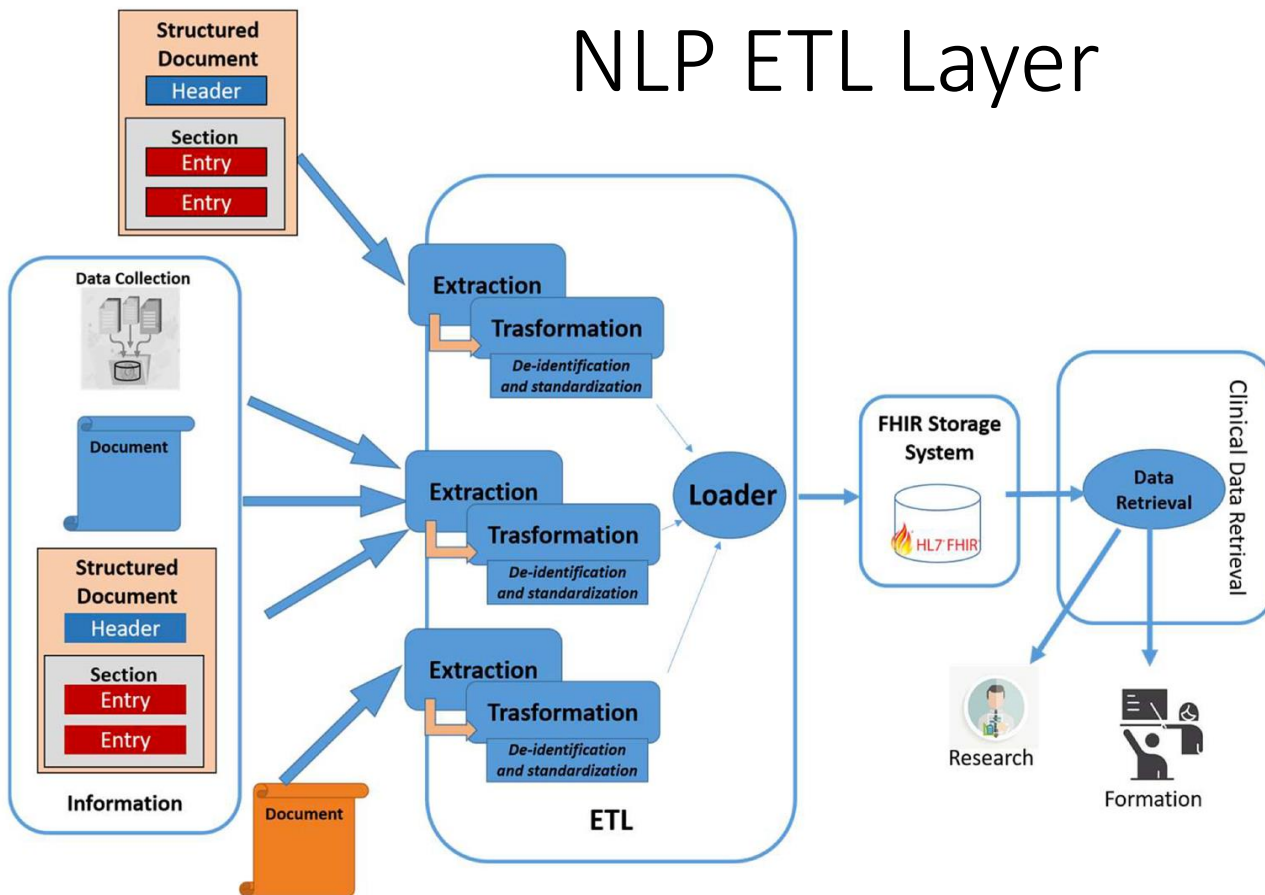
- **Negation:** event has not occurred or entity does not exist
She had no fever yesterday.
- **Uncertainty:** a measure of doubt
The symptoms are not inconsistent with renal failure.
- **Conditional:** could exist or occur under certain circumstances
The patient should come back to the ED if any rash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



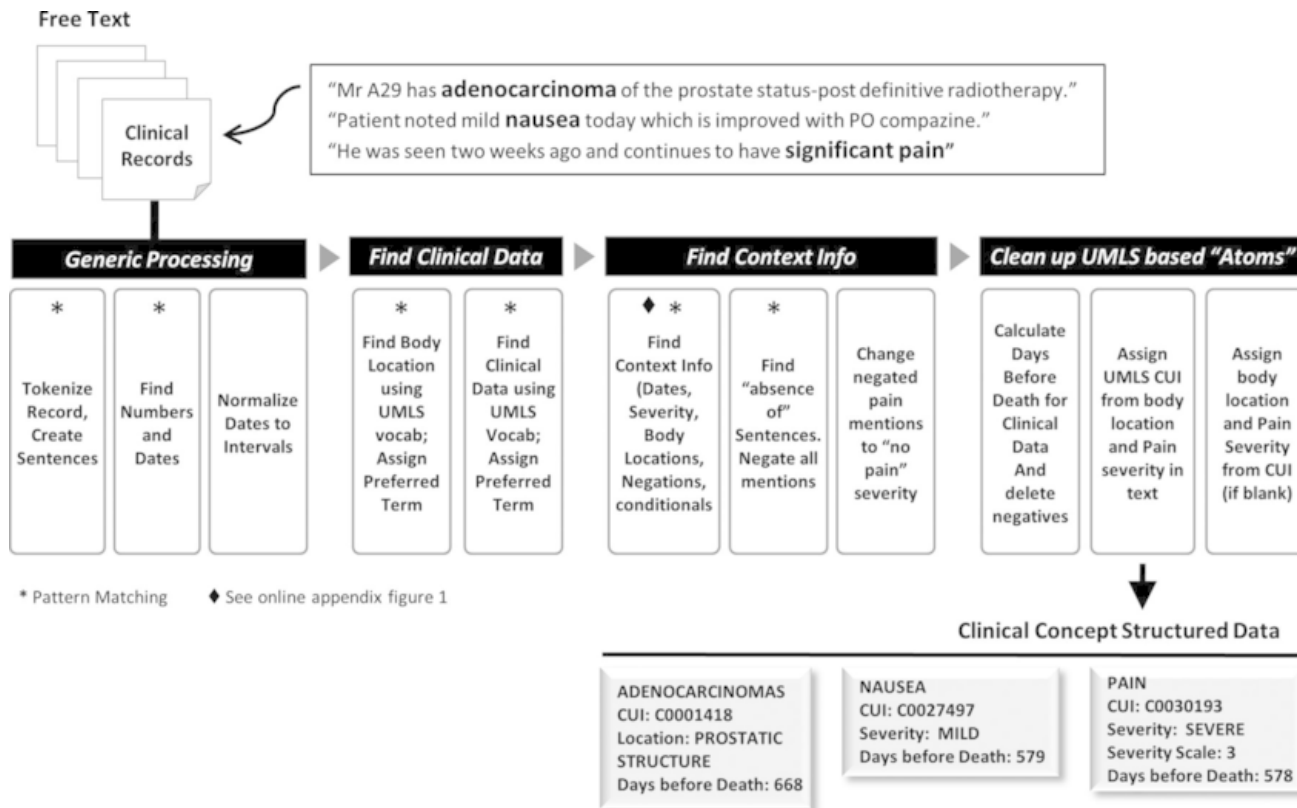
Clinical text is temporal



NLP ETL Layer



Methods - NLP Pipeline



Methods - Annotation or labeling (Gold standard)

Absent^[Z] Past^[X] Hypothetical^[C] Family^[V] SomeoneElse^[M] Possible^[N] Planned^[B] Allergy^[m]

PAST SURGICAL HISTORY: Section_Header Colon resection Procedure Past in 1990 and sinus surgeries Procedure Past in 1987, 1990 and 2005.

ALLERGIES: Section_Header PENICILLIN Drug Allergy.

SOCIAL HISTORY: Section_Header The patient is married.

She uses no ethanol Substance Absent, no tobacco Substance Absent and no il

FAMILY HISTORY: Section_Header Positive for diabetes mellitus type 2 D

REVIEW OF SYSTEMS: Section_Header The patient currently denies any vision Symptom Absent.

Denies chest pain Symptom Absent or shortness of breath Symptom Absent.

She denies any nausea Symptom Absent or vomiting Symptom Absent.

Otherwise, systems are negative.

PLAN: Section_Header Left breast excisional biopsy Procedure Planned with p radiography Test Planned.

Clinical indications: Prostate CA with radical prostatectomy. PSA recurrence. Pre adjuvant radiotherapy.

CT - ABDOMEN & PELVIS contrast

Technique: Multiplanar CT images through the abdomen and pelvis were performed after administration of oral and IV contrast.

Findings: No previous images are available for comparison at the time of reporting.

There is elevation of the left hemidiaphragm. Cardiac size is within normal limits. No focal pulmonary nodules at the lung bases.

A small 5 mm low density lesion was seen adjacent to the IVC within the caudate lobe of the liver. This lesion is too small to characterise its density in CT. An ultrasound is suggested to determine whether or not it is a cystic lesion. Multiple calcified gallstones are seen within the gallbladder. There is no significant dilatation of the biliary system.

A 9 mm low density lesion seen in a subcapsular position within the spleen, adjacent to the splenic hilum (Image 14 on the axial images). A second subcapsular low density lesion is seen in the inferior portion of the spleen. This measures approximately 7 mm (Image 28). Again these lesions are too small to characterise the density using CT. Ultrasound assessment of these two lesions is also suggested.

Annotation Types

- Ungrouped
 - CCV_Justification
- Contextual Polarity
- Cytomorphology
 - De:Cell Growth Pattern
 - De:Cell Type
 - De:Tissue Type
- Descriptor
 - De:Modality Type
- Entity
 - En:Generic Disorder
 - En:Generic Lesion
 - En:Metastases
 - En:Node
 - En:Primary
 - En:Recurrence
- Extent
 - Ex:Clear
 - Ex:Extent
 - Ex:In-Situ
 - Ex:Invasive

Annotation Instances

lymphnodes

Objective 1

Consolidate 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

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

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

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

[Abdulrahman KAAbdulsalam](#)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 information](#)

PMCID: PMC5961766 PMID: [29888032](#)

Abstract

Cancer stage is one of the most important prognostic parameters. The American Joint Committee on Cancer (AJCC) defines cancer stage 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

^aBiomedical Informatics, University of Utah, Salt Lake City, UT

^bEpidemiology, University of Utah, Salt Lake City, UT

^cUtah Cancer Registry, University of Utah, Salt Lake City, UT

^dMedical 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.

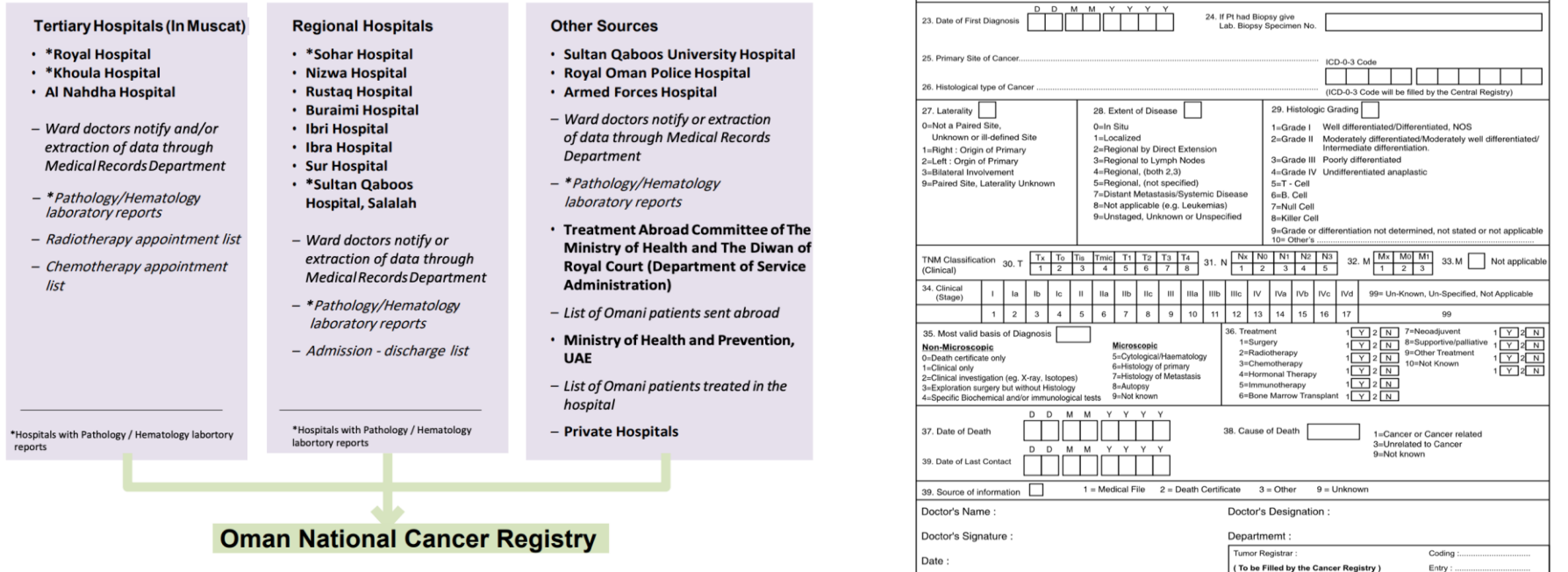
Materials and Methods Records collected at the Utah Cancer Registry (UCR) for patients with colon, lung, or prostate cancers were used for this study. UCR receives multiple

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



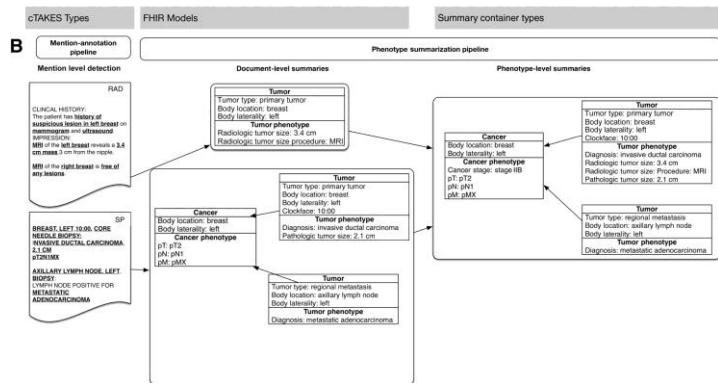
SAMPLE OF OMAN NATIONAL CANCER REGISTRY FORM

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)										
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 3= Caucasian 9= Not Known		2= Asian 4= Other		18. Occupation						
Patient's Address														
19. Telephone Land Line Mobile				21. Wilayat										
20. Other contact's Tel. No.				22. Village										
23. Date of First Diagnosis				24. If PI had Biopsy give Lab. Biopsy Specimen No.										
25. Primary Site of Cancer						ICD-0-3 Code								
26. Histological type of Cancer						(ICD-0-3 Code will be filled by the Central Registry)								
27. Laterality <input type="checkbox"/> Not a Paired Site Unknown or ill-defined Site 1=Right : Origin of Primary 2=Left : Origin of Primary 3=Bilateral Involvement 9=Paired Site, Laterality Unknown			28. Extent of Disease <input type="checkbox"/> 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=Distal Metastatic/Systemic Disease 8=Not applicable (e.g. Leukemias) 9=Unstaged, Unknown or Unspecified			29. Histologic Grade <input type="checkbox"/> 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 30. T _x T ₁ T ₂ T ₃ T ₄ T ₅ T ₆ T ₇ T ₈ T ₉ N ₁ N ₂ N ₃ N ₄ N ₅ M ₁ M ₂ M ₃ M ₄ M ₅ 31. 32. 33. 34. Not applicable														
34. Clinical (Stage) I Ia Ib Ic II Ila Ilb Ilc III IliIa IliIb IliIc IVd 99= Un-Known, Un-Specified, Not Applicable														
35. Most valid basis of Diagnosis Non-Microscopic 0=Death certificate only 1=Clinical only 2=Clinical investigation (eg. X-ray, isotopes) 3=Exploratory 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 2=Radiotherapy 3=Chemotherapy 4=Hormonal therapy 5=Immunotherapy 6=Bone Marrow Transplant 7=Neoadjuvant 8=Supportive/palliative 9=Other Treatment 10=Not Known														
37. Date of Death						38. Cause of Death 1=Cancer or Cancer related 3=Unrelated to Cancer 9=Not known								
39. Date of Last Contact														
39. Source of information <input type="checkbox"/> 1 = Medical File 2 = Death Certificate 3 = Other 9 = Unknown														
Doctor's Name :						Doctor's Designation :								
Doctor's Signature :						Department :								
Date :						Tumor Registrar : Coding : (To be Filled by the Cancer Registry) 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

928701 7/13/2004 10:00:00 AM

Admission Date : 07/03/2004

Discharge Date : 07/12/2004

DISCHARGE DIAGNOSIS : RIGHT
BICONDYLAR TIBIAL PLATEAU
FRACTURE .

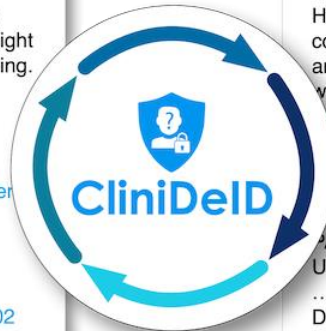
HISTORY OF PRESENT ILLNESS :

Mr. Jones is an otherwise healthy 32 year old male attorney who was vacationing at Richesson Valley when he fell off his moped at a speed of approximately 25 miles per hour. He remembers the accident with no loss of consciousness. He landed on his right knee and noted immediate pain and swelling. He was taken by ambulance to Justice Healthcare where he had plain films that revealed a comminuted bicondylar tibial plateau fracture on the right. He was transferred to the Midvalley Medical Center for further evaluation and treatment.

PAST MEDICAL/SURGICAL HISTORY :
Unremarkable .

...

Dictated By : ALBERTS JOHN , M.D. RY02
Attending : JOHN R. STETSON , M.D.



327468 6/17/1994 12:00:00 AM

Admission Date : 06/07/1994

Discharge Date : 06/16/1994

DISCHARGE DIAGNOSIS : RIGHT
BICONDYLAR TIBIAL PLATEAU
FRACTURE .

HISTORY OF PRESENT ILLNESS :

Mr. Fraser is an otherwise healthy 42 year old male physicist who was vacationing at Abertson Falls when he fell off his moped at a speed of approximately 25 miles per hour. He remembers the accident with no loss of consciousness. He landed on his right knee and noted immediate pain and swelling. He was taken by ambulance to Hasring healthcare where he had plain films that revealed a comminuted bicondylar tibial plateau fracture on the right. He was transferred to the Mercy Medical Center for further evaluation and treatment.

PAST MEDICAL/SURGICAL HISTORY :
Unremarkable .

...

Dictated By : SCHELIEFE BEN , M.D. DJ07
Attending : VITA T. JOHNSON , M.D.

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 positive for WT1
- Some errors with partial matching of “M1” in middle of words:
AIM140.6 AIM111.1

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, PhD¹ and Abdulrahman AAlAbdulsalam, PhD^{2*}

¹ College of Health Science and Medicine

Abstract—National cancer registries rely on manual abstraction of free-text clinical records to collect vital information about cancer diagnosis, stage, progression and treatment. 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, discovery, clinical trial matching, ..., etc.). In this study experimental results of applying NLP to extract information from the records of breast cancer patients for the cancer registry in Oman.

Keywords—Clinical information extraction, Natural Language Processing, Structured Data, Electronic Medical Records

INTRODUCTION

Cancer registries are important resource for monitoring the prevalence of cancer disease in the population and vital for research and decision-making. They are also important for disease control and prevention [1]. However,

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, PhD¹ and Abdulrahman AAlAbdulsalam, PhD²

¹ College of Health Science and Medicine, Sultan Qaboos University ² Department of Computer Science, Sultan Qaboos University

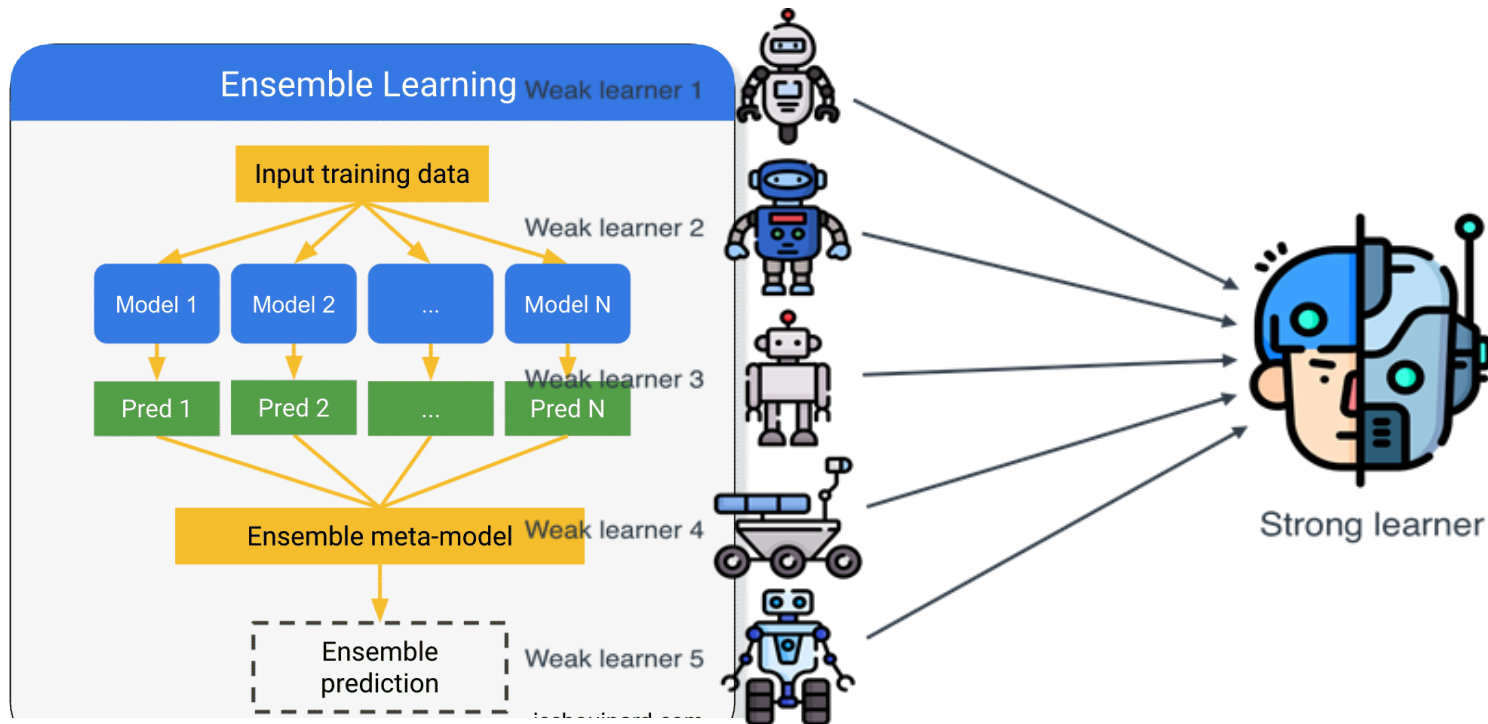
Background. National cancer registries rely on manual abstraction of free-text clinical records to collect vital information about cancer diagnosis, stage, progression and treatment [1]. Many prior studies have demonstrated the ability of natural language processing (NLP) based on machine learning to 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 registry, 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

