Comparing rule-based and machine learning methods to automate Urology Cancer Registry data collection from unstructured data

Eileen Bei[2], Hong Hong Huang[1], Andrew Fang[4], Fiona Lim[5], Matthew Han[2], Elian Chia[3], Priyanka Grover[3],Aixin Sun[5], Jay Lim[1]

1 Singapore General Hospital (SGH), Department of Urology, 2 Singapore General Hospital (SGH), Department of Future Health System, 3 SingHealth, Office for Insights & Analytics (OIA)
4 SingHealth Polyclinics, Department of Research, 5 Nanyang Technological University (NTU), School of Computer Science and Engineering (SCSE)

Introduction:

In 2017, the Singapore General Hospital Urology Cancer Registry (UroCaRe) transited to a semi-automated system to retrieve data

Objective:

To further automate extraction of unstructured data from histopathology reports by applying rule-based and machine learning (ML)/ deep learning (DL) methods

Method:

IRB review is exempted.
The dataset were distributed based on past trend of urology cancer cases (Table 1)

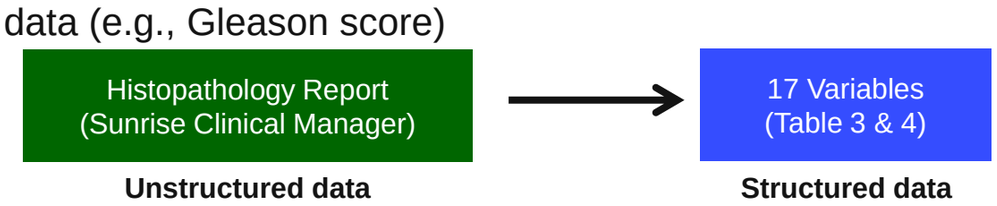
Table 1: Breakdown of number of reports by cancer types (2008–2018)

Cancer types	Number of reports
Renal Cell Carcinoma	49
Prostate Cancer	109
Urothelial Cancer	66
Penile Cancer	20
Testicular cancer	17

Report type: Histology
Source: Sunrise Clinical Manager
Number of records: 250
Time period: 2008–2018

Natural Language Processing (NLP) were applied to the histopathology report for:

- Extraction of unstructured data (e.g. sentences) into structured data (e.g., Gleason score)



- Rule-based VS ML/DL approaches (Table 2)
- Minimum accuracy 95% (based on UroCaRe manual data collection standards)

Table 2: Rule-based approach VS ML/DL approach

	Rule-based	ML/DL
Extraction method	By using patterns/ keywords	By linear classifiers (SVM)/ Neural Networks (BioBert)
Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	K-Fold Cross-validation

FN: False Negative, FP: False Positive, TN: True Negative, TP: True Positive

Results:

This is an interim result

Table 3: 7 variables with >95% accuracy

Variables	Rule-based Accuracy (%)	ML/DL Accuracy (%)
Primary Site	98.5	NA
Tumour Size (1st, 2nd, 3rd Dimensions)	88.0	99.0
Margin Status	97.0	80.0
Gleason Score (Prostate only)	99.0	74.5
Grade (WHO/ISUP)	97.0	NA
Gross Laterality	87.0	98.1
Number of LN Positive	98.8	NA

NA: Not applicable *Variables that are not suitable to use ML/DL approaches

Table 4: 10 variables with < 95% accuracy or WIP

Variables	Rule-based Accuracy (%)	ML/DL Accuracy (%)
Histological Type	82.8	NA
Pathological T*	NA	73.3
Pathological N	NA	79.0
Pathological M*	NA	79.0
Clinical T#	WIP	WIP
Clinical N#	WIP	WIP
Clinical M#	WIP	WIP
Grade (Fuhrman) (Kidney only)	94.0	52.2
Number of LN Resected	78.0	NA
Others Histological Type#	WIP	WIP

DL: Deep Learning, ML: Machine Learning, NA: Not applicable,
WIP: Work in Progress
*Variables were not suitable for the respective approach
#Variables are still in development process

- Accuracy rate of 52.2–99.0%
- 7 variables achieved >95% accuracy (Table 3)
- In this group, rule-based outperforms ML in 5 variables
- 10 variables did not reach minimum accuracy of 95% (Table 4)

Discussion:

- Applying both methods results in manpower savings of estimated 6 min/report

Table 5: Mean accuracy of Rule-based VS ML/DL approach

	Rule-based	ML/DL
Number of variables, n	10	8
Mean accuracy, %	92.0	79.4

DL: Deep Learning, ML: Machine Learning

- Rule-based consistently outperforms ML in most variables
- ML performs better in variables with higher variabilities
- Algorithm modification could improve variable accuracy:
  - Selecting reports from same AJCC edition time period would improve overall accuracy of extraction (i.e. AJCC 8th Ed.)
- Larger annotated dataset could improve accuracy
- Variability would impact model performance
- Our algorithm shows potential in other application
  - Radiology reports
  - Operating theatre reports

Conclusion:

- Both approaches show preliminary success
- Rule-based being the promising approach

Acknowledgement:

This project funded by $500,000 grant from AISG 100E for R.