MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Classification of Diagnostic Certainty in Radiology Reports with Deep Learning

Kento SUGIMOTO

Department of Medical Informatics,
Osaka University Graduate School of Medicine

AIDH IMIA

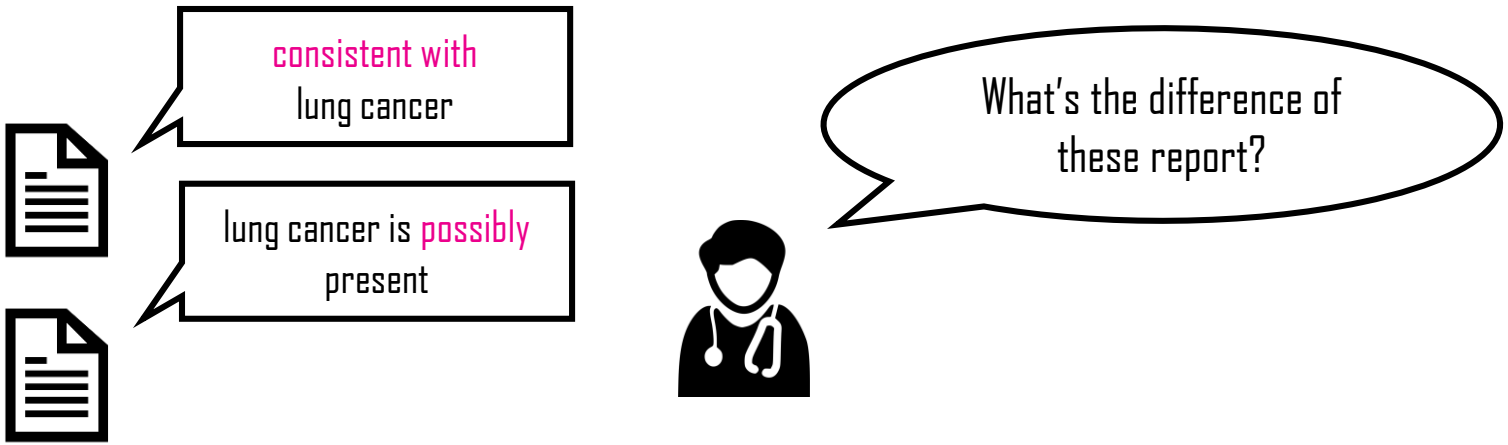
What is diagnostic certainty?

mean the level of confidence in making a diagnosis

- radiologists use various ambiguous expressions in their reports

Need for effective communication

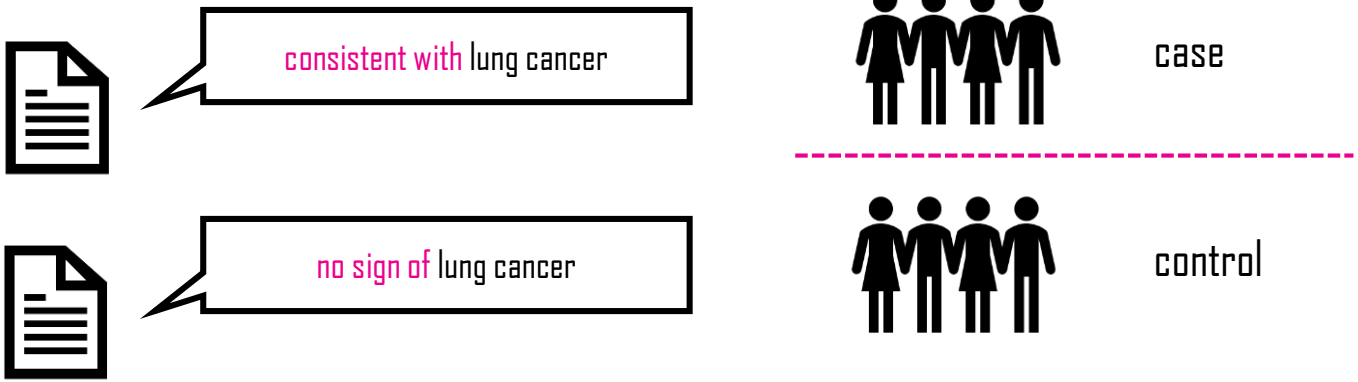
Referring physicians find radiology reports occasionally confusing



For reusing the radiology reports

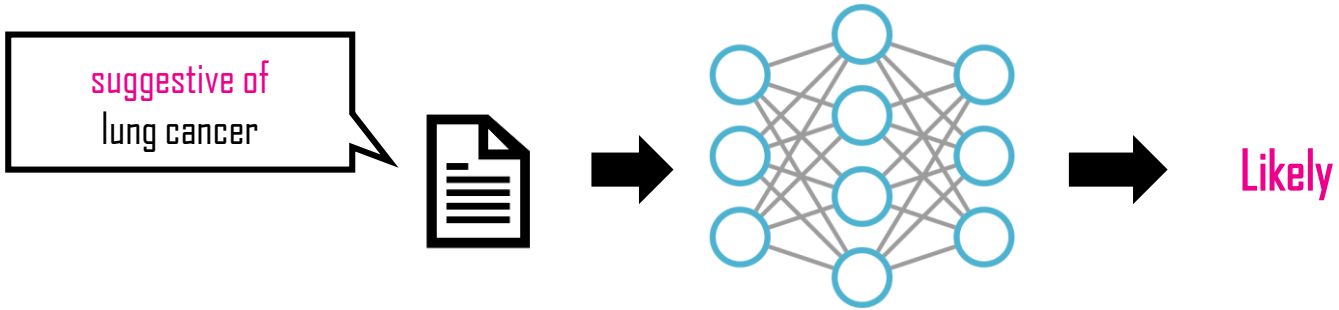
Automatic classification is expected for secondary use

(e.g., cohort selection)



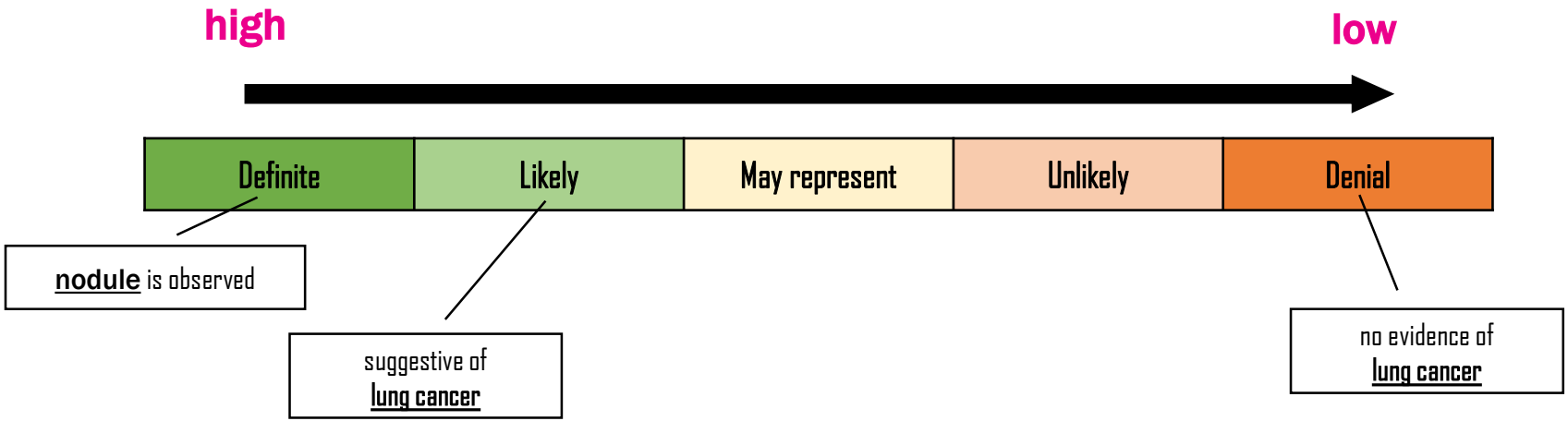
Determining certainty scale

- We introduce a Natural Language Processing (NLP) method
- NLP determine the certainty scale of lesions or findings in the reports



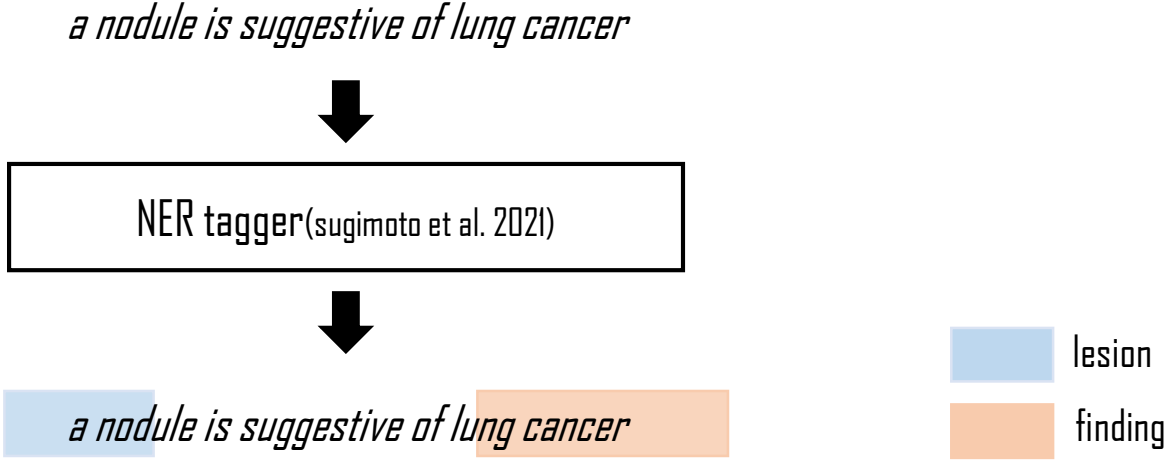
Our certainty scale

Five classes of certainty scale is defined



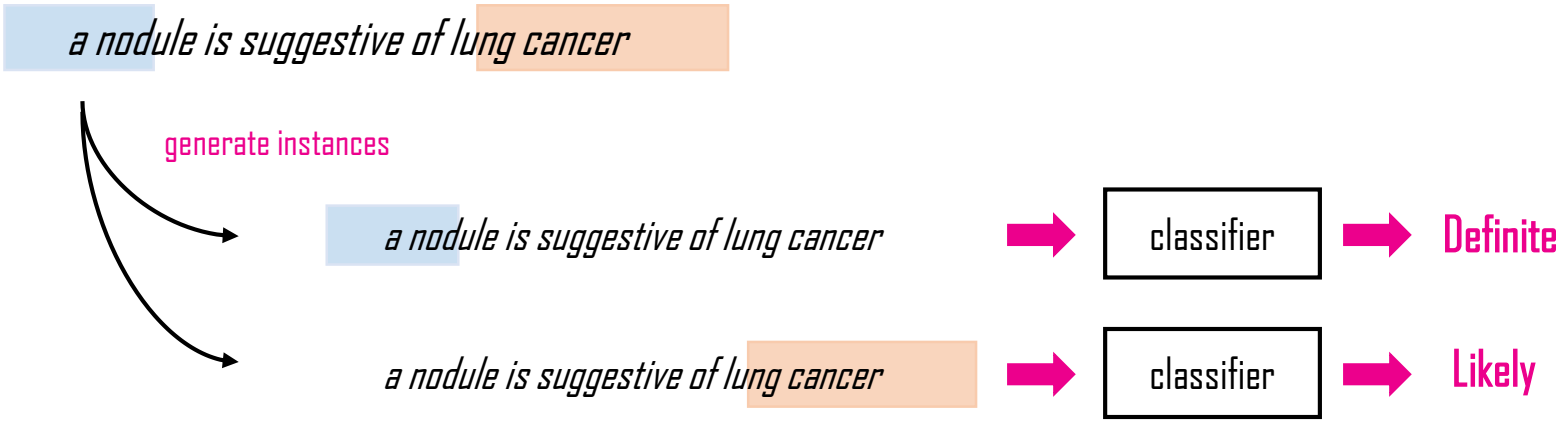
NLP architecture (preprocess)

The NER tagger (sugimoto et al. 2021) was used to detect lesions or findings



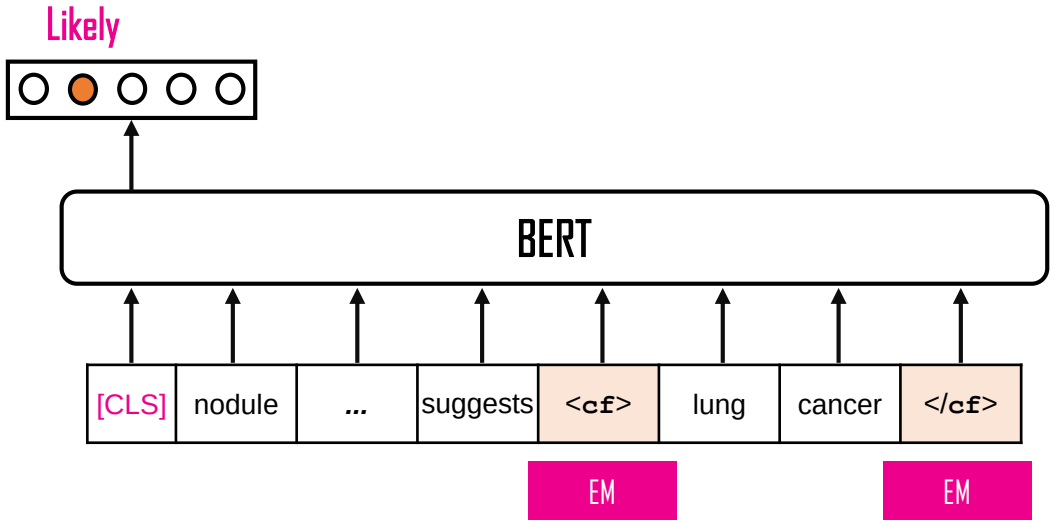
NLP architecture (preprocess)

Each detected entity is fed into classifier to determine its certainty



NLP architecture (main)

Entity Marker (EM) are inserted to recognize target lesions or findings



Corpus development

- Chest CT reports (2010~2018) at Osaka University Hospital
- 540 reports were randomly selected
- Three medical students did the annotation process

Annotation Process

- Given reports with highlighted terms for lesion or finding
- They annotate the certainty class independently

Report	Class
a **nodule** is suggestive of lung cancer	Definite
a nodule is suggestive of **lung cancer**	Likely
no evidence of **lung cancer**	Denial

Two evaluation metrics

	Definite	Likely	May represent	Unlikely	Denial
Definite	○	X	X	X	X
Likely	X	○	X	X	X
May represent	X	X	○	X	X
Unlikely	X	X	X	○	X
Denial	X	X	X	X	○

Strict

	Definite	Likely	May represent	Unlikely	Denial
Definite	○	○	X	X	X
Likely	○	○	○	X	X
May represent	X	○	○	○	X
Unlikely	X	X	○	○	X
Denial	X	X	X	X	○

Relaxed

Result (strict F1-scre)

Class	strict	relaxed
Definite	98.6%	99.0%
Likely	96.3%	100.0%
May represent	91.6%	97.3%
Unlikely	87.8%	97.3%
Denial	98.9%	98.9%
Total	97.6%	98.9%

- A total F1-score obtained 97.6%
- Definite and denial classes achieved satisfactory results
- May represent and unlikely classes had lower F1-scores

Result (relaxed F1-scre)

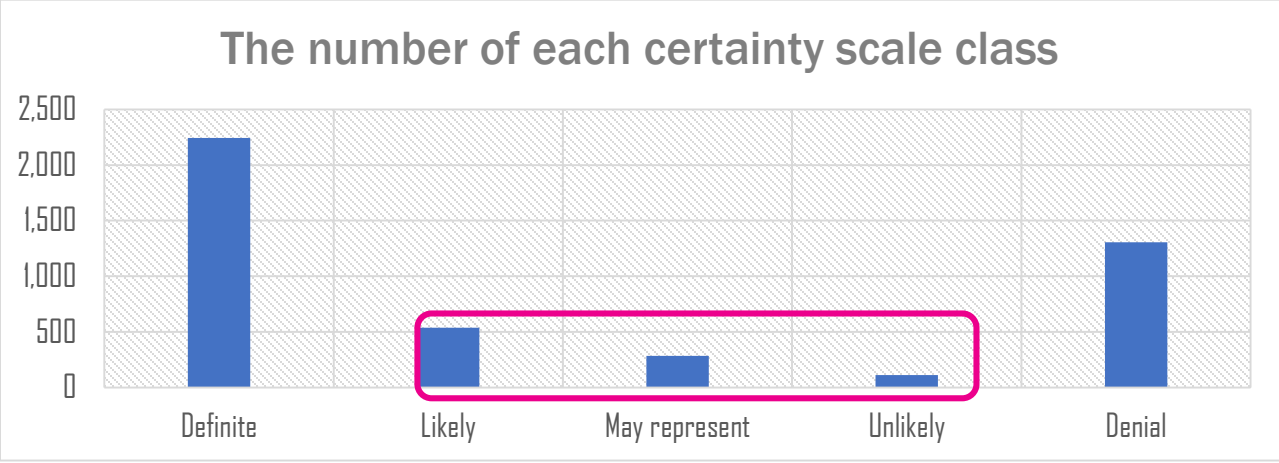
Class	strict	relaxed
Definite	98.6%	99.0%
Likely	96.3%	100.0%
May represent	91.6%	97.3%
Unlikely	87.8%	97.3%
Denial	98.9%	98.9%
Total	97.6%	98.9%

- A total F1-score obtained 98.9%
- Scores of uncertain (likely, may represent, and unlikely) classes improved

A lot of the discrepancy is between the nearest neighbor class

Error analysis

Insufficient sample size in uncertainty class causes lower performance



Error analysis

Model occasionally struggles to understand the context about the change of lesions

Report	Ground Truth	Predicted
Edema was improved	Denial	Definite

Annotation disagreement often occurred such cases ..

Limitation

- We only trained and evaluated the model using reports collected from a single institution
- To ensure generalizability, studies on datasets from outside our institution would be needed

Conclusions

- We presented an ordinal scale to measure the degree of diagnostic certainty
- The deep learning model achieved satisfactory results, which demonstrated that our certainty scale was sufficiently applicable to in-house radiology reports
- This automated classification system will be helpful to clinicians to reduce misinterpretation of radiology reports and contribute to building a curated dataset for secondary use

Confusion matrix

- While the definite and denial classes achieved satisfactory results, the may represent and unlikely classes had lower F1-scores in the strict metric.
- a lot of the discrepancy is between the nearest neighbor class