NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001328

NTCIR-17 MedNLP-SC Radiology Report Subtask Overview: Dataset and Solutions for Automated Lung Cancer Staging

Yuta Nakamura
The University of Tokyo
Japan
yutanakamura-tky@umin.ac.jp

Shouhei Hanaoka
The University of Tokyo
Japan
hanaokalog@gmail.com

Shuntaro Yada
Nara Institute of Science and Technology, Japan
s-yada@is.naist.jp

Shoko Wakamiya
Nara Institute of Science and Technology, Japan
wakamiya@is.naist.jp

Eiji Aramaki
Nara Institute of Science and Technology, Japan
aramaki@is.naist.jp

ABSTRACT

This paper describes the Radiology Report TNM staging (RR-TNM) subtask as a part of NTCIR-17 Medical Natural Language Processing for Social Media and Clinical Texts (MedNLP-SC) shared task in 2023. This subtask focused on automated lung cancer staging based on radiology reports. We created a dataset of 243 Japanese radiology reports containing no personal health information. A total of three teams with 16 members participated and submitted seven solutions. The best accuracy scores for the T, N, and M categories reached 67%, 80%, and 93%, respectively. Through the RR-TNM subtask, we have provided a valuable open Japanese clinical corpus and useful insights to apply natural language processing for secondary usage of staging information.

KEYWORDS

Medical Natural Language Processing, Radiology Reports, Lung Cancer, Cancer Staging

SUBTASK

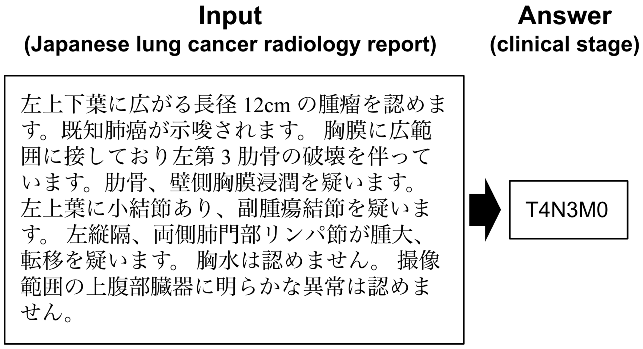
RR-TNM

1 INTRODUCTION

Radiologists serve physicians by planning, performing, and reporting results of imaging studies such as X-ray or computed tomography (CT). Radiology reports include image findings, diagnoses, or occasionally recommendations of next actions, for which radiologists are often called *doctor's doctor* [4]. However, rapidly increasing imaging studies hinder radiologists from creating thorough radiology reports and inhibit physicians from carefully reviewing them [1].

Natural language processing (NLP) has the potential to aid radiological workflow [3, 15, 22]. Since the potential has been proven mainly on English datasets [7, 13], we have provided non-English public datasets mainly in Japanese [2, 17–19, 27, 28, 30].

This Radiology Report TNM staging (RR-TNM) subtask is a part of NTCIR-17 Medical Natural Language Processing for Social Media and Clinical Texts (MedNLP-SC) shared task[1]. Similarly to our last NTCIR-16 Real-MedNLP shared task [30], the RR-TNM subtask is targeted at lung cancer, the largest cause of cancer-related mortalities worldwide [24].



左上下葉に広がる長径 12cm の腫瘤を認めます。既知肺癌が示唆されます。 胸膜に広範囲に接しており左第 3 肋骨の破壊を伴っています。肋骨、壁側胸膜浸潤を疑います。左上葉に小結節あり、副腫瘍結節を疑います。 左縦隔、両側肺門部リンパ節が腫大、転移を疑います。 胸水は認めません。 撮像範囲の上腹部臓器に明らかな異常は認めません。

Figure 1: Scheme for the RR-TNM subtask. The RR-TNM subtask is a multi-label document classification to assign T, N, and M categories to lung cancer radiology reports.

The RR-TNM subtask evaluates NLP systems to automate cancer staging from Japanese radiology reports. Staging is an assessment of stage, or cancer progression, using the TNM classification system in three categories: primary tumor (T), lymph nodes (N), and distant metastasis (M) [9, 12]. Early-stage cancer will be labeled with smaller numbers, such as T1N0M0, whereas advanced-stage cancers will be labeled with larger numbers, such as T3N2M1. Stage information is used primarily for treatment and secondarily for research or public monitoring [26]. However, the stage is only sometimes directly mentioned in radiology reports [23], and readers must often carefully scrutinize them. Thus, automated staging is a highly valuable NLP application. Moreover, staging is a technically challenging task that requires intensive clinical knowledge and reasoning ability.

This overview paper describes the dataset and the solutions by the participating teams as well as their performance. We aim to provide valuable insights to apply NLP systems to utilize cancer stage information.

2 TASK SCHEME

As in Figure 1, the RR-TNM subtask is a three-label document classification to predict T, N, and M categories for each radiology report.

[1] Another subtask is Social Media Adverse Drug Event Detection (SM-ADE) [29].

The RR-TNM dataset was labeled under the 8th edition by the Japan Lung Cancer Society (JLCS) [12] described in Table 1. The JLCS criteria are widely accepted in Japan and closely align with the worldwide standard [9].

All stage labels were aggregated to a coarse level. The choices were T0, T1, T2, T3, or T4 for the T category; N0, N1, N2, or N3 for the N category; and M0 or M1 for the M category.

Registration was open from January 6 to June 26, 2023. First, participating teams received the training and validation sets of labeled radiology reports. Then, the test set of unlabeled radiology reports was released on July 10, 2023. The submission deadline was July 17, 2023.

3 MATERIAL

The RR-TNM dataset comprises 243 Japanese radiology reports and contains no personal health information. Nine board-certified radiologists diagnosed the same open 27 lung cancer cases independently. We split the 27 cases into the training set (108 documents for 12 cases), validation set (54 documents for six cases), and test set (81 documents for nine cases).

3.1 Corpus Creation

We created a raw corpus following the method to build our previous MedTxt-RR dataset [20, 30].

3.1.1 Case Selection. We used Radiopaedia [2] , an open-access radiology reference, as the source of lung cancer cases.

Figure 2 summarizes the case selection protocol. First, we conducted keyword searches on Radiopaedia and obtained 256 lung cancer cases (see Appendix A for the details). Second, we excluded 28 cases that overlapped with MedTxt-RR [20], or lacked patient demographics or case presentation. Third, we performed a brief manual review to check the case titles and CT images. We excluded 161 cases that were not primary lung cancer or had missing or truncated images. Fourth, we performed a detailed manual review to observe CT images. We excluded 16 cases that had highly equivocal findings. Finally, we manually selected 27 cases for RR-TNM. We chose the cases to cover various image findings related to staging as widely as possible. The complete list is available in Appendix A.

All the case selection procedures were conducted by one board-certified radiologist with six years of experience in diagnostic radiology.

3.1.2 Radiology Reporting. In cooperation with Y's Reading, Inc.[3], a teleradiology company in Japan, we recruited nine board-certified radiologists.

We asked the radiologists to create free-text radiology reports in Japanese. We prepared a reporting form for each case in the .docx format, indicating the hyperlink of the Radiopaedia URL, age, sex, and case presentation. The forms were sent and received via e-mail.

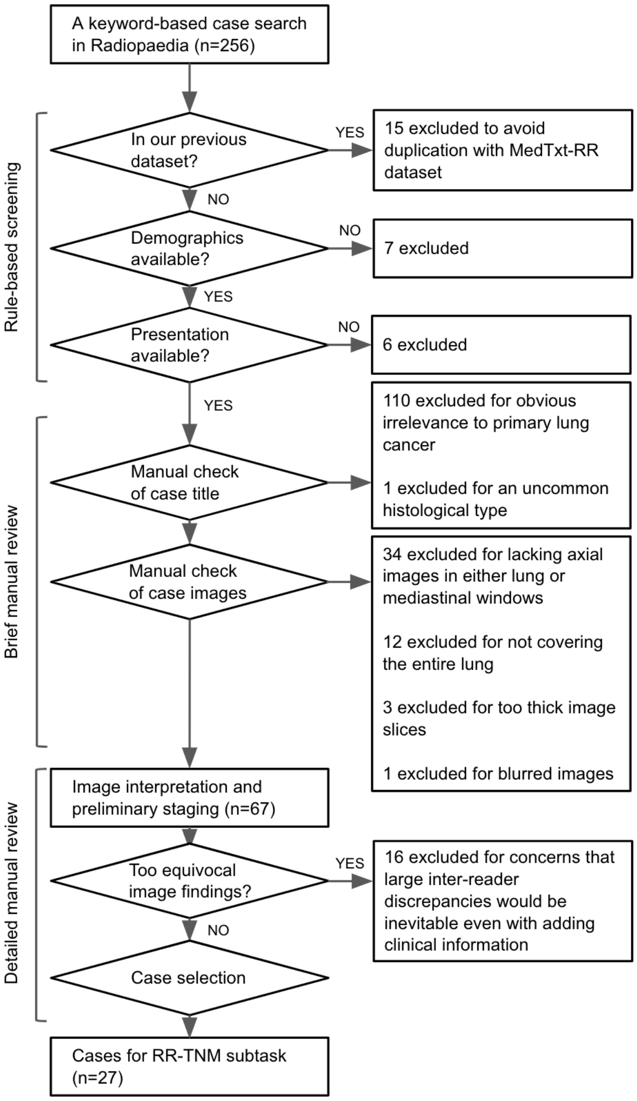

A keyword-based case search in Radiopaedia (n=256)
Rule-based screening
In our previous dataset?
YES
15 excluded to avoid duplication with MedTxt-RR dataset
NO
Demographics available?
NO
7 excluded
YES
Presentation available?
NO
6 excluded
YES
Brief manual review
Manual check of case title
110 excluded for obvious irrelevance to primary lung cancer
1 excluded for an uncommon histological type
Manual check of case images
34 excluded for lacking axial images in either lung or mediastinal windows
12 excluded for not covering the entire lung
3 excluded for too thick image slices
1 excluded for blurred images
Detailed manual review
Image interpretation and preliminary staging (n=67)
Too equivocal image findings?
YES
16 excluded for concerns that large inter-reader discrepancies would be inevitable even with adding clinical information
NO
Case selection
Cases for RR-TNM subtask (n=27)

Figure 2: Case selection protocol for the RR-TNM subtask.

We did not directly instruct the radiologists because we would like them to create radiology reports like the actual workflow. Instead, we indirectly guided the radiologists by adding fictitious patient information that would disambiguate equivocal image findings.

The radiologists diagnosed the cases independently without discussion for consensus. Moderate inter-observer agreement was sufficient because we aimed to build a text corpus, not a vision-and-language multimodal dataset.

We made no changes to the received radiology reports except for a few typographical errors affecting the cancer stage.

3.2 Stage Labeling

We asked the radiologists who created the radiology reports to assign TNM. Next, the labels were manually checked and corrected

[2] https://radiopaedia.org/
[3] https://www.ysreading.co.jp/

Table 1: Lung cancer staging criteria of the 8th edition by the Japan Lung Cancer Society (in Japanese) [12].

Stage	Definition
T: primary lesion	
TX	原発腫瘍の存在が判定できない，あるいは喀痰または気管支洗浄液細胞診でのみ陽性で画像診断や気管支鏡では観察できない
T0	原発腫瘍を認めない
Tis	上皮内癌 (carcinoma in situ): 肺野型の場合は，充実成分径 0 cm かつ病変全体径≦ 3 cm
T1	腫瘍の充実成分径≦ 3 cm, 肺または臓側胸膜に覆われている, 葉気管支より中枢への浸潤が気管支鏡上認められない (すなわち主気管支に及んでいない)
(T1mi	微少浸潤性腺癌: 部分充実型を示し, 充実成分径≦ 0.5 cm かつ病変全体径≦ 3 cm)
(T1a	充実成分径≦ 1 cm でかつ Tis・T1mi には相当しない)
(T1b	充実成分径>1 cm でかつ≦ 2 cm)
(T1c	充実成分径>2 cm でかつ≦ 3 cm)
T2	充実成分径>3 cm でかつ≦ 5 cm, または充実成分径≦ 3cm でも以下のいずれかであるもの • 主気管支に及ぶが気管分岐部には及ばない • 臓側胸膜に浸潤 • 肺門まで連続する部分的または一側全体の無気肺か閉塞性肺炎がある
(T2a	充実成分径>3 cm でかつ≦ 4 cm)
(T2b	充実成分径>4 cm でかつ≦ 5 cm)
T3	充実成分径>5 cm でかつ≦ 7 cm, または充実成分径≦ 5 cm でも以下のいずれかであるもの • 壁側胸膜, 胸壁 (superior sulcus tumor を含む), 横隔神経, 心膜のいずれかに直接浸潤 • 同一葉内の不連続な副腫瘍結節
T4	充実成分径>7 cm, または大きさを問わず横隔膜, 縦隔, 心臓, 大血管, 気管, 反回神経, 食道, 椎体, 気管分岐部への浸潤, あるいは同側の異なった肺葉内の副腫瘍結節
N: nodal involvement	
NX	所属リンパ節評価不能
N0	所属リンパ節転移なし
N1	同側の気管支周囲かつ/または同側肺門, 肺内リンパ節への転移で原発腫瘍の直接浸潤を含める
N2	同側縦隔かつ/または気管分岐下リンパ節への転移
N3	対側縦隔, 対側肺門, 同側あるいは対側の前斜角筋, 鎖骨上窩リンパ節への転移
M: distant metastasis	
M0	遠隔転移なし
M1	遠隔転移がある
(M1a	対側肺内の副腫瘍結節，胸膜または心膜の結節，悪性胸水 (同側・対側)，悪性心嚢水)
(M1b	肺以外の一臓器への単発遠隔転移がある)
(M1c	肺以外の一臓器または多臓器への多発遠隔転移がある)

when inconsistent with the radiology reports. Then, all the labels were aggregated into a coarse level, namely either of T0, T1, T2, T3, or T4 for the T category; N0, N1, N2, or N3 for the N category; and M0 or M1 for the M category.

3.3 Dataset Splitting

We randomly split the RR-TNM dataset into training, validation, and test sets in the ratio of 12:6:9 ($\approx 0.44 : 0.22 : 0.33$). The split was on a case basis so that radiology reports of the same case did not overlap in multiple sets.

3.4 Dataset Statistics

Figure 3 presents the label distribution of the training, validation, and test sets. According to the chi-square test, there was a significant discrepancy in the distribution of the T category among the three splits ($p < 0.01$), but not of the N category ($p \approx 0.11$) or M category ($p \approx 0.07$).

4 METHODS

Three teams with 16 members in total formally participated and submitted their results. The team KRad had eight members and submitted one result. The team kuhp had seven members and submitted three results. The team NAIST-SOCRR had five members and submitted three results. The teams KRad and kuhp had six members in common.

We denote each system in combination with the team name and submission number, such as 'KRad-1.'

4.1 Team KRad

They employed zero-shot in-context learning to `gpt-3.5-turbo` [6, 21], a ChatGPT model, to make it follow the staging criteria without updating the parameters.

System KRad-1 They used prompts that included (i) the model persona: "you are a skilled thoracic surgeon," (ii) the staging criteria, (iii) instructions to follow the staging criteria

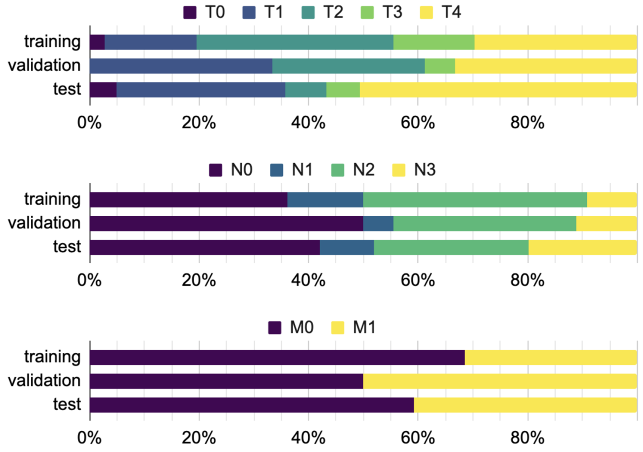


Figure 3: Label distribution of the RR-TNM dataset.

and return results in a JSON format, and (iv) the radiology report. Then, the staging results were extracted from the model's output using regular expressions. Lastly, output values in incorrect formats were replaced with T0, N0, or M0.

4.2 Team kuhp

They fine-tuned `open-calm-7b`[4], an open Japanese large language model built on GPT-NeoX [5]. They employed an instruction tuning framework [10]. They applied data augmentation 100 times to randomly drop characters and shuffle sentences of radiology reports. They further extended the fine-tuning dataset with handcraft question-answer pairs to enhance clinical knowledge.

System kuhp-1 Inference by the model fine-tuned for around 0.35 epoch.

System kuhp-2 Inference by the model fine-tuned for around 0.69 epoch.

System kuhp-3 Inference by the model fine-tuned for around 0.69 epoch.

4.3 Team NAISTSOCRR

They solved the RR-TNM subtask as multi-class document classification [11], using bidirectional Transformer models such as BERT [8] and RoBERTa [16].

System NAISTSOCRR-1 A single deep learning model was employed to infer T, N, and M categories jointly. They fine-tuned JMedRoBERTa [25], a biomedical Japanese pre-trained RoBERTa model [16].

System NAISTSOCRR-2 Predictions for T, N, and M categories were treated as three separate tasks. Each was solved by a fine-tuned Japanese bidirectional model, namely JMedRoBERTa [25], Tohoku-BERT-v3[5], and UTH-BERT [14].

System NAISTSOCRR-3 Majority baseline. The most frequent training label, T1N0M0, was used for all predictions.

5 EVALUATION

5.1 Evaluation Metrics

5.1.1 *Main Evaluation.*

We employ two types of evaluations and four metrics:

- Separate evaluation:
  - **T accuracy**: fraction of cases whose T category was correctly predicted. Formally, given gold standards and predictions $\{(y_i^T, \hat{y}_i^T)\}_{i=1}^K$ for K test samples, *T accuracy* is calculated as:

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{1}\left[y_i^T = \hat{y}_i^T\right], \quad y_i^T \in \{\mathrm{T0}, \mathrm{T1}, \mathrm{T2}, \mathrm{T3}, \mathrm{T4}\}.$$

  - **N accuracy**: fraction of cases whose N category was correctly predicted. Formally, given gold standards and predictions $\{(y_i^N, \hat{y}_i^N)\}_{i=1}^K$ for K test samples, *N accuracy* is calculated as:

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{1}\left[y_i^N = \hat{y}_i^N\right], \quad y_i^N \in \{\mathrm{N0}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}\}.$$

  - **M accuracy**: fraction of cases whose M category was correctly predicted. Formally, given gold standards and predictions $\{(y_i^M, \hat{y}_i^M)\}_{i=1}^K$ for K test samples, *M accuracy* is calculated as:

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{1}\left[y_i^M = \hat{y}_i^M\right], \quad y_i^M \in \{\mathrm{M0}, \mathrm{M1}\}.$$

- Joint evaluation:
  - **Joint accuracy**: fraction of cases whose T, N, and M categories were all correctly predicted. Formally, given gold standards and predictions $\{((y_i^T, y_i^N, y_i^M), (\hat{y}_i^T, \hat{y}_i^N, \hat{y}_i^M))\}_{i=1}^K$ for K test samples, *joint accuracy* is calculated as:

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{1}\left[(y_i^T, y_i^N, y_i^M) = (\hat{y}_i^T, \hat{y}_i^N, \hat{y}_i^M)\right],$$
$$y_i^T \in \{\mathrm{T0}, \mathrm{T1}, \mathrm{T2}, \mathrm{T3}, \mathrm{T4}\},$$
$$y_i^N \in \{\mathrm{N0}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}\},$$
$$y_i^M \in \{\mathrm{M0}, \mathrm{M1}\}.$$

5.1.2 *Additional Analysis.*

Accuracy metrics cannot differentiate between minor errors and significant mistakes. In a clinical context, larger discrepancies, such as between T1 and T4, have a more substantial impact on decision-making than smaller discrepancies like those between T2 and T3. To address this perspective, we also computed the weighted Kappa coefficient, which penalizes larger discrepancies more than minor ones. We applied a quadratic weight.

5.2 Results

Table 2 shows the results. The highest N and M accuracy of 0.8025 and 0.9259 were achieved by KRad-1 using gpt-3.5-turbo, which was a promising performance. In contrast, T accuracy dropped in almost all systems. NAISTSOCRR-2 achieved the highest T accuracy of 0.6667 and the highest joint accuracy of 0.3704.

[4] https://huggingface.co/cyberagent/open-calm-7b

[5] https://huggingface.co/cl-tohoku/bert-base-japanese-v3

Table 2: Results of the RR-TNM subtask. The best scores are shown in bold text.

	Model		Accuracy				Weighted kappa		
System	Name(s)	Type	T	N	M	Joint	T	N	M
KRad-1	`gpt-3.5-turbo`	Causal	0.3951	**0.8025**	**0.9259**	0.2716	0.3660	**0.8422**	**0.8436**
kuhp-1	`open-calm-7b`	Causal	0.4815	0.6049	0.7407	0.2346	0.6536	0.5699	0.6386
kuhp-2	`open-calm-7b`	Causal	0.4568	0.5185	0.7778	0.2099	0.0000	0.0000	0.0000
kuhp-3	`open-calm-7b`	Causal	0.4444	0.5062	0.7901	0.1975	0.5016	0.6453	0.4220
NAISTSOCRR-1	JMedRoBERTa	Bidirectional	0.6049	0.6049	0.8765	0.3086	0.6018	0.6577	0.7419
NAISTSOCRR-2	JMedRoBERTa, Tohoku-BERT-v3, UTH-BERT	Bidirectional	**0.6667**	0.5679	0.8395	**0.3704**	**0.6556**	0.6933	0.6692
NAISTSOCRR-3	(Majority baseline)	-	0.3086	0.4198	0.5926	0.1975	0.0000	0.0000	0.0000

6 DISCUSSION

We held the RR-TNM subtask to evaluate the state-of-the-art NLP systems in automating lung cancer staging. To this end, we created an open Japanese radiology report dataset. We received seven solutions from three teams.

We have created a new open Japanese clinical corpus to alleviate scarcity in open Japanese clinical data. In particular, our previous MedTxt-RR corpus and RR-TNM subtask dataset here have been the only ones covering radiology reports. MedTxt-RR had 135 documents, and this time, we almost tripled the corpus size by adding 243 documents.

We have provided an open Japanese radiology corpus with TNM labels for the first time. In our last shared task, we assigned NER labels to the MedTxt-RR corpus, but NER labels were far from direct real-world applications. Conversely, TNM labels in this RR-TNM subtask are closer to real-world applications.

Participants' solutions indicate that ChatGPT is promising because it determined N and M categories with over 80% accuracy. ChatGPT required no training to surpass the other solutions, including fine-tuned Japanese large language models or BERT-like models.

Most solutions had more room for enhancement for the T category than the N and M categories. We attribute the discrepancy to the task complexity. Although the N and M categories are only based on anatomy, the T category's criteria include anatomy and tumor size. The tumor size must be numerically compared to the criteria, where language models have difficulty.

Our future shared task could continue to target automated staging by providing more fine-grained annotations or an additional corpus.

7 CONCLUSION

We organized the RR-TNM subtask focusing on automated lung cancer staging based on Japanese radiology reports. The RR-TNM subtask brought a valuable open Japanese clinical corpus and helpful insights to apply NLP for secondary usage of staging information.

ACKNOWLEDGEMENTS

This work was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G9, JST AIP-PRISM Grant Number JPMJCR18Y1, Japan, as well as ANR grant ANR-20-IADJ-0005-01, France, under the trilateral ANR-DFG-JST AI call.

REFERENCES

[1] R. Alexander, S. Waite, M. A. Bruno, E. A. Krupinski, L. Berlin, S. Macknik, and S. Martinez-Conde. 2022. Mandating limits on workload, duty, and speed in radiology. *Radiology* 304, 2 (aug 2022), 274–282.

[2] Eiji Aramaki, Yoshinobu Kano, Tomoko Ohkuma, and Mizuki Morita. 2016. MedNLPDoc: Japanese Shared Task for Clinical NLP. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP).* The COLING 2016 Organizing Committee, Osaka, Japan, 13–16. https://aclanthology.org/W16-4203

[3] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. 2022. Natural Language Processing: from Bedside to Everywhere. *Yearbook of medical informatics* (June 2022).

[4] L. Berlin. 1977. The radiologist: doctor's doctor or patient's doctor. *AJR Am J Roentgenol* 128, 4 (Apr 1977), 702.

[5] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models.* Association for Computational Linguistics, virtual+Dublin, 95–136. https://doi.org/10.18653/v1/2022.bigscience-1.9

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

[7] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc* 23, 2 (3 2016), 304–310.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] The Union for International Cancer Control. 2016. *The TNM Classification of Malignant Tumours* (8th ed.).

[10] Koji Fujimoto, Mizuho Nishio, Chikako Tanaka, Morteza Rohanian, Farhad Nooralahzadeh, Michael Krauthammer, and Fabio Rinaldi. 2023. CClassification of cancer TNM stage from Japanese radiology report using on-premise LLM at NTCIR-17 MedNLP-SC RR-TNM subtask. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17.* https://doi.org/10.20736/0002001299

[11] Takuya Fukushima, Yuka Otsuki, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2023. NAISTSOCRR at the NTCIR-17 MedNLP-SC Radiology Report. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17.* https://doi.org/10.20736/0002001285

[12] The Japan Lung Cancer Society. 2021. *General RuIe for Clinical and Pathological Record of Lung Cancer* (8th ed.). Revised Version.

[13] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L. H. Lehman, L. A. Celi, and R. G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 10, 1 (Jan 2023), 1.

[14] Y. Kawazoe, D. Shibata, E. Shinohara, E. Aramaki, and K. Ohe. 2021. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLoS One* 16, 11 (2021), e0259763.

[15] N. Linna and C. E. Kahn. 2022. Applications of natural language processing in radiology: A systematic review. *Int J Med Inform* 163 (Jul 2022), 104779.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:arXiv:1907.11692

[17] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2014. Overview of the NTCIR-11 MedNLP Task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[18] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2016. Overview of the NTCIR-12 MedNLPDoc Task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-12, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[19] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP Task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[20] Yuta Nakamura, Shohei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. Clinical Comparable Corpus Describing the Same Subjects with Different Expressions. *Stud Health Technol Inform* 290 (Jun 2022), 253–257.

[21] Mizuho Nishio, Hidetoshi Matsuo, Takaaki Matsunaga, Koji Fujimoto, Morteza Rohanian, Farhad Nooralahzadeh, Fabio Rinaldi, and Michael Krauthammer. 2023. Zero-shot classification of TNM staging for Japanese radiology report using ChatGPT at RR-TNM subtask of NTCIR-17 MedNLP-SC. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. https://doi.org/10.20736/0002001283

[22] E. Pons, L. M. Braun, M. G. Hunink, and J. A. Kors. 2016. Natural Language Processing in Radiology: A Systematic Review. *Radiology* 279, 2 (May 2016), 329–343.

[23] R. Sexauer, T. Weikert, K. Mader, A. Wicki, S. delin, B. Stieltjes, J. Bremerich, G. Sommer, and A. W. Sauter. 2018. Towards More Structure: Comparing TNM Staging Completeness and Processing Time of Text-Based Reports versus Fully Segmented and Annotated PET/CT Data of Non-Small-Cell Lung Cancer. *Contrast Media Mol Imaging* 2018 (2018), 5693058.

[24] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal. 2023. Cancer statistics, 2023. *CA Cancer J Clin* 73, 1 (Jan 2023), 17–48.

[25] Kaito Sugimoto, Taichi Iki, Yuki Chida, Teruhito Kanazawa, and Akiko Aizawa. 2023. JMedRoBERTa: a Japanese Pre-trained Language Model on Academic Articles. In *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*.

[26] T. C. Tucker, E. B. Durbin, J. K. McDowell, and B. Huang. 2019. Unlocking the potential of population-based cancer registries. *Cancer* 125, 21 (Nov 2019), 3729–3737.

[27] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the NTCIR-13 MedWeb Task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-13, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[28] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2019. Tweet Classification Toward Twitter-Based Disease Surveillance: New Data, Methods, and Evaluations. *J Med Internet Res* 21, 2 (20 Feb 2019), e12783. https://doi.org/10.2196/12783

[29] Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. https://doi.org/10.20736/0002001327

[30] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-16*. National Institute of Informatics (NII).

A DETAILS OF CASE SELECTION

We conducted keyword searches on Radiopaedia on January 18, 2023. The keywords are shown in Table 4. We used the following search options: the pathologically proven lung cancer (Diagnosis = "Certain"), the inclusion of CT images (Modality = "CT"), and the field-of-view covering the chest (Systems = "Chest").

Table 3: A list of keywords used to search lung cancer cases in Radiopaedia.

lung cancer
lung carcinoma
lung adenocarcinoma
lung scc
lung squamous cell cancer
lung squamous cell carcinoma
pulmonary cancer
pulmonary carcinoma
pulmonary adenocarcinoma
pulmonary scc
pulmonary squamous cell cancer
pulmonary squamous cell carcinoma
bronchial cancer
bronchial carcinoma
bronchial adenocarcinoma
bronchial scc
bronchial squamous cell cancer
bronchial squamous cell carcinoma
bronchogenic cancer
bronchogenic carcinoma
bronchogenic adenocarcinoma
bronchogenic scc
bronchogenic squamous cell cancer
bronchogenic squamous cell carcinoma
cancer of lung
cancer of the lung
carcinoma of lung
carcinoma of the lung
adenocarcinoma of lung
adenocarcinoma of the lung
scc of lung
scc of the lung
squamous cell cancer of lung
squamous cell cancer of the lung
squamous cell carcinoma of lung
squamous cell carcinoma of the lung
sclc
nsclc
small cell cancer
small cell carcinoma
small cell lung cancer
small cell lung carcinoma
GGO
GGN
subsolid
part-solid
part solid
ground glass opacity
ground glass nodule
ground glass
minimally invasive adenocarcinoma
atypical adenomatous hyperplasia

Table 4: A list of Radiopaedia cases used for RR-TNM dataset.

		Data partition		
Case	Article title and URL	Training	Validation	Test
1	Invasive lung adenocarcinoma (https://radiopaedia.org/cases/invasive-lung-adenocarcinoma)		✓	
2	Lepidic predominant adenocarcinoma of the lung (https://radiopaedia.org/cases/lepidic-predominant-adenocarcinoma-of-the-lung-1)	✓		
3	Lung cancer with bronchoscopic biopsy (https://radiopaedia.org/cases/lung-cancer-with-bronchoscopic-biopsy)		✓	
4	Invasive mucinous adenocarcinoma (https://radiopaedia.org/cases/invasive-mucinous-adenocarcinoma-1)			✓
5	Left upper lobe collapse due to lung cancer (https://radiopaedia.org/cases/left-upper-lobe-collapse-due-to-lung-cancer)	✓		
6	Lung cancer causing complete atelectasis (https://radiopaedia.org/cases/lung-cancer-causing-complete-atelectasis)		✓	
7	Thickened right paratracheal stripe - small cell lung carcinoma (https://radiopaedia.org/cases/thickened-right-paratracheal-stripe-small-cell-lung-carcinoma-1)	✓		
8	Adenocarcinoma of the lung - EGFR-mutated (https://radiopaedia.org/cases/adenocarcinoma-of-the-lung-egfr-mutated-2)		✓	
9	Lung cancer with bone metastases (https://radiopaedia.org/cases/lung-cancer-with-bone-metastases-1)	✓		
10	Asbestosis complicated by lung cancer (https://radiopaedia.org/cases/asbestosis-complicated-by-lung-cancer)	✓		
11	T1b apical lung cancer (https://radiopaedia.org/cases/t1b-apical-lung-cancer)		✓	
12	Bronchial mucoepidermoid carcinoma (https://radiopaedia.org/cases/bronchial-mucoepidermoid-carcinoma-1)	✓		
13	Pancoast tumor with CT guided biopsy (https://radiopaedia.org/cases/pancoast-tumour-with-ct-guided-biopsy-1)	✓		
14	Lung cancer (massive) (https://radiopaedia.org/cases/lung-cancer-massive)	✓		
15	Invasive mucinous adenocarcinoma of the lung mimicking pneumonia (https://radiopaedia.org/cases/invasive-mucinous-adenocarcinoma-of-the-lung-mimicking-pneumonia)			✓
16	Metastatic adenocarcinoma of the lung (https://radiopaedia.org/cases/metastatic-adenocarcinoma-of-the-lung-2)			✓
17	Metastatic lung cancer (https://radiopaedia.org/cases/metastatic-lung-cancer-5)	✓		
18	Advanced metastatic lung cancer (https://radiopaedia.org/cases/advanced-metastatic-lung-cancer)		✓	
19	Lung adenocarcinoma (https://radiopaedia.org/cases/lung-adenocarcinoma-6)			✓
20	Pulmonary adenocarcinoma (https://radiopaedia.org/cases/pulmonary-adenocarcinoma-2)			✓
21	Pulmonary adenocarcinoma (https://radiopaedia.org/cases/pulmonary-adenocarcinoma-3)			✓
22	Lung cancer (https://radiopaedia.org/cases/lung-cancer-29)	✓		
23	Pancoast tumor (https://radiopaedia.org/cases/pancoast-tumour-20)			✓
24	Hypertrophic osteoarthropathy (https://radiopaedia.org/cases/hypertrophic-osteoarthropathy-8)	✓		
25	Common bile duct stone and lung cancer (https://radiopaedia.org/cases/common-bile-duct-stone-and-lung-cancer)	✓		
26	Primary sarcomatoid carcinoma of lung (https://radiopaedia.org/cases/primary-sarcomatoid-carcinoma-of-lung-3)			✓
27	Luftsichel sign in lung cancer (https://radiopaedia.org/cases/luftsichel-sign-in-lung-cancer)			✓