

# Methods of Information in Medicine

## Cross-lingual Natural Language Processing on Limited Annotated Case/Radiology Reports in English and Japanese: Insights from the Real-MedNLP Workshop

Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, Eiji Aramaki.

Affiliations below.

DOI: 10.1055/a-2405-2489

**Please cite this article as:** Yada S, Nakamura Y, Wakamiya S et al. Cross-lingual Natural Language Processing on Limited Annotated Case/Radiology Reports in English and Japanese: Insights from the Real-MedNLP Workshop. *Methods of Information in Medicine* 2024. doi: 10.1055/a-2405-2489

**Conflict of Interest:** The authors declare that they have no conflict of interest.

**This study was supported by** Ministry of Health, Labour and Welfare (<http://dx.doi.org/10.13039/501100003478>), JPMH21AC500111, Japan Science and Technology Agency (<http://dx.doi.org/10.13039/501100002241>), JPMJCR18Y1, JPMJCR20G9

### Abstract:

**Background:** Textual datasets (corpora) are crucial for the application of natural language processing (NLP) models. However, corpus creation in the medical field is challenging, primarily because of privacy issues with raw clinical data such as health records. Thus, the existing clinical corpora are generally small and scarce. Medical NLP (MedNLP) methodologies perform well with limited data availability.

**Objectives:** We present the outcomes of the Real-MedNLP workshop, which was conducted using limited and parallel medical corpora. Real-MedNLP exhibits three distinct characteristics: (1) Limited Annotated Documents: The training data comprises only a small set (approximately 100) of case reports (CRs) and radiology reports (RRs) that have been annotated. (2) Bilingually Parallel: The constructed corpora are parallel in Japanese and English. (3) Practical Tasks: The workshop addresses fundamental tasks, such as named entity recognition and applied practical tasks.

**Methods:** We propose three tasks: named entity recognition (NER) of approximately 100 available documents (Task 1), NER based only on annotation guidelines for humans (Task 2), and clinical applications (Task 3) consisting of adverse drug effects (ADE) detection for CRs and identical case identification (CI) for RRs.

**Results:** Nine teams participated in this study. The best systems achieved 0.65 and 0.89 F1-scores for CRs and RRs in Task 1, whereas the top scores in Task 2 decreased by 50–70%. In Task 3, ADE reports were detected by up to 0.64 F1-score, and CI scored up to 0.96 binary accuracy.

**Conclusions:** Most systems adopt medical-domain-specific pre-trained language models using data augmentation methods. Despite the challenge of limited corpus size in Tasks 1 and 2, recent approaches are promising because the partial match scores reached approximately 0.8–0.9 F1-scores. Task 3 applications revealed that the different availabilities of external language resources affected the performance per language.

### Corresponding Author:

Dr. Shuntaro Yada, Nara Institute of Science and Technology, Division of Information Science, 8916-5 Takayama-cho, 6300192 Ikoma, Japan, s-yada@is.naist.jp

### Affiliations:

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Shuntaro Yada, Nara Institute of Science and Technology, Division of Information Science, Ikoma, Japan  
Yuta Nakamura, The University of Tokyo Hospital, 22nd Century Medical and Research Center, Bunkyo-ku, Japan  
Shoko Wakamiya, Nara Institute of Science and Technology, Ikoma, Japan  
Eiji Aramaki, Nara Institute of Science and Technology, Ikoma, Japan



This article is protected by copyright. All rights reserved.

Accepted Manuscript

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

# Cross-lingual Natural Language Processing on Limited Annotated Case/Radiology Reports in English and Japanese: Insights from the Real-MedNLP Workshop

Shuntaro Yada, Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan

Yuta Nakamura, 22nd Century Medical and Research Center, The University of Tokyo Hospital, Tokyo, Japan

Shoko Wakamiya, Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan

**Eiji Aramaki**, PhD, Professor, Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan, [aramaki@is.naist.jp](mailto:aramaki@is.naist.jp)

## ABSTRACT

**Background** Textual datasets (corpora) are crucial for the application of natural language processing (NLP) models. However, corpus creation in the medical field is challenging, primarily because of privacy issues with raw clinical data such as health records. Thus, the existing clinical corpora are generally small and scarce. Medical NLP (MedNLP) methodologies perform well with limited data availability.

**Objectives** We present the outcomes of the Real-MedNLP workshop, which was conducted using limited and parallel medical corpora. Real-MedNLP exhibits three distinct characteristics: (1) Limited Annotated Documents: The training data comprises only a small set (approximately 100) of case reports (CRs) and radiology reports (RRs) that have

been annotated. (2) Bilingually Parallel: The constructed corpora are parallel in Japanese and English. (3) Practical Tasks: The workshop addresses fundamental tasks, such as named entity recognition and applied practical tasks.

**Methods** We propose three tasks: named entity recognition (NER) of approximately 100 available documents (Task 1), NER based only on annotation guidelines for humans (Task 2), and clinical applications (Task 3) consisting of adverse drug effects (ADE) detection for CRs and identical case identification (CI) for RRs.

**Results** Nine teams participated in this study. The best systems achieved 0.65 and 0.89 F1-scores for CRs and RRs in Task 1, whereas the top scores in Task 2 decreased by 50–70%. In Task 3, ADE reports were detected by up to 0.64 F1-score, and CI scored up to 0.96 binary accuracy.

**Conclusions** Most systems adopt medical-domain-specific pre-trained language models using data augmentation methods. Despite the challenge of limited corpus size in Tasks 1 and 2, recent approaches are promising because the partial match scores reached approximately 0.8–0.9 F1-scores. Task 3 applications revealed that the different availabilities of external language resources affected the performance per language.

## Keywords

Natural Language Processing

Deep Learning

Drug-Related Side Effects and Adverse Reactions

## INTRODUCTION

The rise of electronic medical records (EMRs) has heightened the importance of natural language processing (NLP) techniques in healthcare due to the vast amount of textual data they generate<sup>1</sup>. Given the widespread interest in NLP within computer science, the volume of research on medical NLP has experienced a remarkable surge annually. This trend has also been supported by numerous medical NLP workshops, such as CLEF eHealth<sup>2,3</sup>, n2c2<sup>4</sup> (formerly known as i2b2), MADE<sup>5</sup>, and MEDIQA<sup>6</sup>. However, despite the substantial body of research, the availability of privacy-compliant medical text data remains limited, particularly in non-English languages<sup>7-9</sup>.

To address this limitation, we organized a series of medical NLP workshops with open datasets (the MedNLP Series) at an international conference, NII Testbeds, and Community for Information Access Research (NTCIR): MedNLP-1<sup>10</sup>, MedNLP-2<sup>11</sup>, MedNLPDoc<sup>12</sup>, and MedWeb<sup>13</sup>. In MedNLP-1, we introduced a foundational NLP task, named entity recognition (NER), utilizing dummy medical records crafted by medical professionals. MedNLP-2 focused on a term normalization task, again employing dummy medical records developed by medical experts. The MedNLPDoc workshop was designed to encompass a comprehensive task. Beginning with a medical record sourced from a medical textbook, participants were tasked with identifying an appropriate disease name represented by ICD codes. In MedWeb, a disease tweet classification task was designed to simulate the use of social media data in the medical and healthcare domains; dummy Twitter data were created in Japanese and translated into English and Chinese.

Past workshops in the MedNLP Series, summarized in Table 1, successfully produced valuable datasets. However, two major problems have been identified. (1) The data were *not real* clinical texts but were *dummy* records or sample texts from medical textbooks. (2) The datasets were limited to Japanese, which made it difficult to compare the results with those of other English-based workshop results.

To address these aspects, during 2021–2022, we proposed and organized **Real-MedNLP**, the first workshop in the MedNLP Series that handles *real* and *parallel* medical text. Our data comprised two document types: (1) case reports (**MedTxt-CR**) and (2) radiology reports (**MedTxt-RR**). Both corpora are *realistic* medical/clinical texts based on the materials available on the Internet, where *realistic* means that *real* case-report articles constitute MedTxt-CR, and MedTxt-RR contains newly written (dummy) radiology reports that interpret commonly available *real* radiology images. Furthermore, we manually translated the original Japanese text into English, enabling us to develop the first benchmark for cross-lingual medical NLP. Considering the data, we redesigned the workshop scheme to achieve our goal of promoting systems applicable at the bedside. This reintroduces the aforementioned challenging restrictions in medical NLP: limited dataset sizes. The proposed task format is as follows:

- **Low-resource NER (Tasks 1 & 2):** Participants are supposed to extract medical expressions from text, although only a limited number of annotated documents are available for training the machine learning models. This reflects the real-world MedNLP, which often suffers from a scarcity of available annotated text in hospitals or their departments owing to annotation costs. We further defined two Tasks: **Just**

**100 Training** (SubtaskTask 1) and **Guideline Learning** (Task 2). This task set is called *low-resource* in the NLP research<sup>14</sup>; Task 2 corresponds explicitly -to *zero* or *few-shot learning* in the machine learning field.

- **Applications (Task 3):** Corresponding to the two document types, we propose two practical and useful MedNLP applications in actual clinical work. For case reports, we designed an information extraction task for **adverse drug events (ADE) reporting** (i.e., pharmacovigilance) characterized by a different approach from relation extraction, which is usually adopted in existing workshops such as i2b2 2009<sup>15</sup>. We propose a novel **case identification (CI)** task for radiology reports to detect reports originating from identical patients.

These demanding tasks offer exciting prospects for advancing practical systems that can enhance various medical services, including phenotyping<sup>16</sup>, drug repositioning<sup>17</sup>, drug target discovery<sup>18</sup>, precision medicine<sup>19</sup>, clinical text-input methods<sup>20,21</sup>, and Electronic Health Record (EHR) summarization/aggregation<sup>22</sup>.

This study provides an account of the materials used, detailed task definitions, evaluation metrics employed, an overview of participants' approaches, and the overall results achieved during the Real-MedNLP workshop.

## **MATERIALS**

### **Corpora: MedTxt**

#### **Overview**

The textual datasets (corpora) released by workshop participants were named **MedTxt**.

Two types of medical and clinical documents were used as corpora: case reports (CR) and

radiology reports (RR). These two corpora are parallel in Japanese (JA, original) and English (EN, translated). For example, we identified the Japanese case report corpus as MedTxt-CR-JA (MedTxt-CR-JA).

### **Case Reports: MedTxt-CR**

This case report is a medical research paper in which doctors describe specific clinical cases. Case reports aimed to share clinically notable issues with other doctors, particularly those in medical societies. The format of case reports is similar to that of discharge summaries, which are clinical documents written by doctors to record the treatment history of discharged patients. While popular English medical NLP corpora are often composed of discharge summaries (e.g., MIMIC-III), techniques for case report analysis are smoothly extended to analyze discharge summaries.

MedTxt-CR-JA comprises open-access case reports obtained from the Japanese scholarly publication platform J-Stage<sup>23</sup>. Figure 1 shows its sample. As the number of medical societies that produce open-access publications is limited, the types of patients and diseases reported in open-access case reports are highly biased. To reduce the bias caused by the publication policy (whether to prefer open access or not) of each medical society, we selected 224 case reports based on the “frequencies” in a Japanese disease-name dictionary MANBYO-DIC (J-MeDic)<sup>24</sup>, which contains the frequency of each term in Japanese medical corpora. These case reports were manually translated from Japanese (MedTxt-CR-JA) to English (MedTxt-CR-EN) while retaining named entity annotations (described later). They were divided into 148 training and 76 test datasets.



## **Radiology Reports: MedTxt-RR**

A radiology report is a clinical document written by a radiologist to share a patient's status with physicians. Each radiology report discusses a single radiological examination such as radiography, CT, or MRI. A radiology report contains (i) descriptions of all normal and abnormal findings, and (ii) interpretations of the findings, including disease diagnosis and recommendations for the next clinical test or treatment. Although most radiology AI research focuses only on images because image-based AI has drawn much attention, NLP on radiology-report text also has the potential for a wide variety of clinical applications<sup>25</sup>.

MedTxt-RR<sup>26</sup> consists of 135 radiology reports. Figure 2 shows its sample. The MedTxt-RR aims to provide information on the diversity of expressions used by different radiologists to describe the same diagnosis. One of the difficulties in analyzing radiology reports is the variety of writing styles. However, relying solely on radiology reports from medical institutions presents limitations. In typical clinical settings, only one report is generated per radiological exam, restricting the available data for in-depth analysis. MedTxt-RR-JA was created to overcome this problem by crowdsourcing, in which 9 radiologists independently wrote radiology reports for the same series of 15 lung cancer cases.

MedTxt-RR-EN is an English translation of MedTxt-RR-JA by nine translators corresponding to radiologists. We divided them into 72 training datasets (8 cases) and 63 test datasets (7 cases).

## Tasks and Annotations

### **Tasks 1 & 2: Low-resource NER challenge**

Because NER is the most fundamental information extraction for medical NLP, we designed a challenge regarding NER for a limited number of clinical reports (approximately 100). This corpus size fits into so-called *low-resource* NLP, in which training models become challenging<sup>14</sup>. This setting is the de facto standard in medical NLP mainly because of the innate difficulty of medical concept annotation and privacy concerns regarding patient health records. To address this challenge, we defined two *Tasks* based on the size of the available training data:

- **Task 1) Just 100 Training:** We provided approximately 100 reports for training, corresponding to the standard few-resource (or few-shot) supervised learning.
- **Task 2) Guideline Learning:** We provided annotation guidelines containing only a handful of annotated sentences, simulating human annotator training, in which human annotators usually learn from annotation guidelines defined by NLP researchers.

Although we provided only a few training data for both tasks, workshop participants could use any other resources outside this project (e.g., medical dictionaries and medically pretrained language models) if they found them useful for their methods.

We adopted the following entity types from an existing medical NER scheme<sup>27,28</sup>:

- **Diseases and symptoms** <d> with the modality 'certainty':
  - positive – the existence is recognized

- negative – the absence is confirmed
- suspicious – the existence is speculated
- general – hypothetical or common knowledge mentions

- **Anatomical parts** <a>

- **Time** <timeX3> with the modality 'type':

- date – a calendar date
- time – a clock time
- duration – a continuous period
- set – frequency consisting of multiple dates, times, and periods
- age – a person's age
- med – medicine-specific time expressions such as 'post-operative.'
- misc – exceptional time expressions other than the above

- **Test** <t-test/key/val> with the modality 'state':

- scheduled – treatment is planned
- executed – treatment was executed
- negated – treatment was canceled
- other – an exceptional state other than the above

- **Medicine** <m-key/val> with the modality 'state':

- scheduled – treatment is planned
- executed – treatment was executed
- negated – treatment was canceled
- other – an exceptional state other than the above

Detailed definitions and information on modality are provided in<sup>27</sup> and Chapter 2 of our annotation guidelines<sup>28</sup>. Several batches of the Japanese corpus were annotated separately by multiple native Japanese speakers without medical knowledge and then translated into English while retaining the annotated entities. This annotation scheme enables a reasonable quality of coherent annotation even if annotators lack medical knowledge<sup>27,29</sup>.

The detailed statistics of the entity annotations in the datasets are presented in Table 3. We evaluated the following tag sets.

- **MedTxt-CR:** <d>, <a>, <t-test>, <timex3>, <m-key>, <m-val>, <t-key>, and <t-val> (all types above)
- **MedTxt-RR:** <d>, <a>, <t-test>, and <timex3> (a subset of the types above)

The teams may choose whether or not to predict the modalities of the entities.

### **Task 3: Applications**

#### **Task3-CR: Adverse Drug Effect Extraction (ADE)**

In this application, tailored for MedTxt-CR, the systems were supposed to analyze input case reports and extract any Adverse Drug Event (ADE) information. Unlike the typical relation-extraction formulation in past ADE extraction tasks<sup>4</sup>, we set the objective to *table slot filling*, that is, to independently predict the degree of involvement in ADEs for each mentioned disease and medicine. Thus, we attempted to address an issue with standard ADE extraction in which even medical professionals find it difficult to identify ADEs only from the text, leading to annotation difficulties. As portrayed in Figure 3, the ADE information consists of two tables: <d>-table for disease/symptom names and <m-key>-

table for drug names. For each entity in these tables, the four levels of ADE certainty (**ADEval**) based on Kelly et al.<sup>30</sup> were as follows:

**3** – Definitely | **2** – Probably | **1** – Unlikely | **0** – Unrelated

For disease names, ADEvals were interpreted as the likelihood of *being an ADE*, and for medicine names, the interpretation was the likelihood of *causing an ADE*. To annotate these labels, we let two annotators follow the author’s perspective on whether drugs and symptoms were related to ADE (i.e., the writer’s perception). However, if the annotators noticed other possibilities of ADEs that were not explicitly pointed out in the report, we allowed them to label ADEval as well (i.e., the reader’s perception). Note that one annotator has experience as a pharmacist and the other possesses a master’s degree in biology. Table 4 presents the distribution of ADEvals in the training set.

### ***Task3-RR: Case Identification (CI)***

This application was specifically designed for MedTxt-RR. Given the unsorted radiology reports, the participants were required to identify sets of radiology reports that diagnosed identical images, as depicted in Figure 4. MedTxt-RR was created by collecting radiology reports from multiple radiologists who independently diagnosed the same CT images; this original correspondence between radiology reports and CT images was used as the gold standard label (document level).

In addition to an educational purpose in which trainee radiologists practice writing reports on shared images, this task would contribute to better NLP models that accurately understand the clinical content of radiology reports without being confused by synonyms

or paraphrases, as MedTxt-RR contains radiology reports with almost the same clinical content but with various expressions.

## **Data Availability**

The training portions of MedTxt corpora are publicly available at NTCIR Test Collection<sup>31</sup>.

## **METHODS**

### **Baseline Systems**

#### ***Overview***

We propose baseline systems for each Task in *Japanese* corpora<sup>32</sup>. All the systems adopted straightforward approaches to solving tasks. The models proposed below are based on a BERT<sup>33</sup> model pre-trained on Japanese corpora<sup>34</sup>, which tokenizes Japanese text using the morphological analyzer MeCab<sup>35</sup>.

#### ***Task1&2: NER models***

We fine-tuned the individual models on each training set using the same NER-training scheme as Devlin et al.<sup>33</sup>. The model predicts all entity types defined in Yada et al.<sup>27</sup>, instead of only targeting the subsets for the tasks, because more entity types would provide more context to the model, even though task complexity may increase.

We released the baseline model trained on MedTxt-CR-JA at the Hugging Face Hub<sup>36</sup>.

### ***Task3-CR: ADE classifier***

We solved this application task using an entity-wise classification scheme. For each disease or drug entity in the table row, we designed a model input consisting of four parts separated by [SEP] special tokens (Figure 5): (i) the document ID, (ii) contextual tokens around the targeted entity, (iii) the targeted entity itself, and (iv) its entity type name (i.e., “disease” or “drug”). Specifically, Part (ii) contains 50 + 50 characters before and after the target entity, including the entity itself. For simplicity, the context around the first mention is extracted when the targeted token appears multiple times in the document.

### ***Task3-RR: CI clusterer***

We exhaustively classified all document pairs to identify radiology reports describing identical patients. For each pair of given documents, the model judges whether the inputs describe an identical patient (i.e., binary classification; Figure 6). Each document pair is separated by a [SEP] token. Considering permutation, we regard a document pair (Article text 1, Article text 2) as identical patient reports if and only if both predictions of (Article text 1, Article text 2) and (Article text 2, Article text 1) result in “identical-patient.”

### **Participating Teams and Systems**

Nine teams participated in the-MedNLP workshops. Table 5 lists the teams, their demographics, and number of systems submitted by each team for each task. Notably, our workshop attracted global industries (i.e., five of the nine teams) from China, Japan, and the US, demonstrating a high demand for practical medical NLP solutions worldwide. Two were multidisciplinary, and medical and computer science researchers collaborated. Most

teams have adopted pre-trained language models as their methodological basis, often combined with either or both external medical knowledge and data augmentation. Each team's approach is summarized: Refer to the corresponding system papers for NTCIR-16 Real-MedNLP<sup>32,37-45</sup> for more information.

- **AMI** (*Task1-CR-JA, Task1-RR-JA, Task2-CR-JA, Task2-RR-JA*): This team<sup>37</sup> adopted the medically pretrained Japanese BERT (UTH-BERT<sup>46</sup>) as its base model. For Task 1, two systems were proposed: an ensemble of four UTH-BERT models and a fine-tuned UTH-BERT with a CRF layer. For Task 2, a knowledge-guided pipeline was proposed in which UTH-BERT's NER predictions were corrected using medical dictionaries such as J-MeDic<sup>24</sup>, Hyakuyaku-Dictionary<sup>47</sup>, and comeJisyo<sup>48</sup> along with an additional vocabulary extended by bootstrapping.
- **FRDC** (*Task1-CR-EN, Task3-CR-EN (ADE), Task3-RR-EN (CI)*): This team<sup>39</sup> submitted two systems utilizing BioBERT<sup>49</sup> for Task1-CR-EN. One system involved fine-tuning exclusively, while the other integrated data augmentation techniques, including label-wise token replacement, synonym replacement, mention replacement, and shuffling within segments<sup>50</sup>. For Task 3, this team proposed a vocabulary-adapted BERT (VART) model that was continuously trained from a fine-tuned BERT, but with out-of-vocabulary words from the initial fine-tuning. In Task3-CR-EN (ADE), VART was trained with multiple NLP tasks to classify the ADEval for each entity based on its contextual text and predict the entity type (disease or drug). Task3-RR-EN (CI) was solved using a combination of two main methods: (1) key feature clustering to find case-specific information, such as tumor size, and (2) K-means clustering based



on document embedding using sentence BERT<sup>51</sup> to cluster the remaining cases unidentified in Step (1).

- **GunNLP** (*Task3-RR-JA (CI)*): This team<sup>39</sup> applied collaborative filtering to an entity-frequency matrix created from the bag-of-words representation of radiology reports.
- **NAISTSOC (Baseline)** (*Task1-CR-JA, Task1-RR-JA, Task2-CR-JA, Task2-RR-JA, Task3-CR-JA (ADE), Task3-RR-JA (CI)*): This multidisciplinary team<sup>32</sup> provides the aforementioned task baselines for Japanese corpus tracks.
- **NICTmed** (*Task1-CR-EN, Task1-CR-JA, Task3-CR-EN (ADE), Task3-CR-JA (ADE), Task3-RR-EN (CI), Task3-RR-JA (CI)*): This team<sup>40</sup> investigated the effectiveness of two close multilingual language models, multilingual BERT (mBERT)<sup>33</sup> and XLM-RoBERTa<sup>52</sup>. While simply fine-tuning them for Task1-CR (NER), the team addressed Task3-CR (ADE) by considering ADEval as an additional attribute of the named entities. Task3-RR (CI) was solved by the k-means clustering of documents vectorized by mBERT.
- **NTTD** (*Task1-CR-JA, Task1-RR-JA*): This team fine-tuned the Japanese BERT using data augmentation (i.e., synonym replacement and shuffling within segments)<sup>50</sup>.
- **SRCB** (*Task1-CR-EN, Task1-RR-EN, Task3-CR-EN (ADE)*): This team<sup>42</sup> adopted BERT<sup>53</sup>, BioBERT, clinical BERT<sup>54</sup>, PubMed BERT<sup>55</sup>, and entity BERT<sup>56</sup> as the base models for Task 1. These are fine-tuned by span-based prompting (e.g., token prediction with the prompt “[span] is a [MASK] entity,” where [span] is one of the possible spans and [MASK] is the span’s entity type), along with data augmentation (i.e., language model-based token replacement). For Task 3 (ADE), the team used

PubMed BERT, Clinical BERT, and BioBERT, and after multitask learning (medicine/disease classification and cloze-test tasks) and two-stage training<sup>57</sup>, they were fine-tuned with prompt learning (i.e., ADE-causing drug and disease pair prediction) assisted by data augmentation (back translation via Chinese and random feature dropout).

- **Syapse** (*Task2-RR-EN, Task3-CR-EN (ADE), Task3-RR-EN (CI)*): This team<sup>43</sup> did not perform any fine-tuning of the given training datasets. Instead, it applies standard NLP modules to pipelines such as MetaMap<sup>58</sup> and ScispaCy<sup>59</sup>. First, the pipeline obtains entities for task 2. For ADE applications, an additional SciBERT<sup>60</sup> module measures the similarity between medicine and disease-embedding pairs to regard high-similarity pairs as high-ADEval entities. A bag-of-entity representation of documents was used for the CI application to measure document similarity.
- **Zukyo** (*Task1-CR-JA, Task1-RR-JA, Task1-CR-EN, Task1-RR-EN, Task3-RR-JA (CI)*): This multidisciplinary team addresses tasks according to language. The Japanese sub-team<sup>44</sup> fine-tuned an ensemble of Japanese BERT models using data augmentation (random entity replacement constrained by entity types). For the Japanese CI application, the subteam manually re-annotated each sentence of the given RR corpus using the TNM classification<sup>61</sup>, the international standard of cancer staging. The same ensemble NER architecture was fine-tuned separately in the sentence-wise sequential labeling of TNM.  
The English sub-team<sup>45</sup> adopted domain-specific BERT models to tackle Task 1: BioBERT<sup>53</sup>, ClinicalBERT<sup>54</sup>, and RoBERTa (general domain)<sup>62</sup>. Furthermore, entity attributes were predicted by SVM using bags of contextual words around the

entities. The training dataset was augmented by random entity replacement and constrained by entity types.

## Evaluation Metrics

### **Tasks 1 and 2**

We employed the common F1 measure as the NER metric; specifically, we adopted its micro-average over entity classes (i.e., micro F1). Furthermore, we considered the following two factors to enable an evaluation specific to few-resource NER:

- **Boundary factor (exact/partial):** The standard NER metric treats a correct NE match if and only if the predicted span is identical to the gold standard (i.e., *exact match*). We also introduce *partial match* to Tasks 1 & 2: if the predicted span overlaps the gold-standard span, the prediction is regarded as “partially correct,” obtaining a diminished score. This is intended for downstream tasks in which partially identified NEs are still useful, such as large-scale medical record analysis and highlighting important note sections. We considered the proportion of common sub characters between the gold standard and predicted entities to calculate the partial match score. Given that a gold standard entity and predicted entity overlap in their spans, we first calculate *entity-level* partial-match precision and recall, where  $l_1$  stands for the character length of  $e_1$  and  $l_2$  denotes the common character length between  $e_1$  and  $e_2$ . We then obtain *the system-level* partial-match precision and recall as follows:

Finally, we calculate the partial F1 score, i.e., their harmonic mean.

- **Frequency factor:** In our few-resource NER setting, substantial portions of NEs in the corpora appear only a few times, making the tasks challenging for machine-learning models. To measure model performance in identifying such rare NEs, we designed a novel weighted metric that penalizes the correct guesses of the system more heavily as the predicted entity appears more frequently in the training dataset. For each gold-standard entity in the test set, we multiplied the entity-level precision and recall scores by the weight based on the term frequency of the same entity in the training set. This weighting portrays the extent to which the system relied on high-frequency entities in the training phase, as well as how well the system captured low-frequency entities.

### **Task 3**

For the ADE application, we employed two levels of evaluation for the ADEval classification: **entity level** and **report level**.

- **Entity level:** The precision (P), recall (R), and F1-score (F) of each ADEval (= 0, 1, 2, 3) were micro-averaged for the disease and medicine entities.
- **Report level:** We regard a report that contains at least one entity with ADEval as a POSITIVE-REPORT, otherwise NEGATIVE-REPORT. This binary classification scheme evaluates the report-wise P, R, and F values of the POSITIVE REPORT.

For the CI application, we adopted standard metrics for supervised clustering: adjusted normalized mutual information (AdjMI)<sup>63</sup>, Fowlkes-Mallows (FM) scores<sup>64</sup>, and binary accuracy. We aim to penalize random predictions or split clinically similar documents into

numerous clusters. Both the AdjMI and FM were robust in addressing these errors. AdjMI is an adjustment of mutual information (MI), which is a popular clustering metric. The FM also provides a useful means of estimating performance distinctly, as it spans from zero to one.

### **Ethical considerations**

This study did not require the participants to be involved in any physical or mental intervention. Furthermore, as it did not utilize personally identifiable information, the study was exempt from institutional review board approval in accordance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects outlined by the Japanese national government.

## **RESULTS**

### **System Notations**

Distinct systems proposed by a team X, for example, are denoted in combination with numbers, such as “X-1” and “X-2.” For readability, we multiplied the scores by 100, except for the Task 3 metrics.

### **Tasks 1 and 2**

Table 6 lists the F1 measures per evaluation factor obtained for Tasks 1 and 2. Since predicting the entity modalities is optional, we separately report the scores of the modality-aware match (“+mod”) from the modality-agnostic match (“label”).

The best scores range from 49.93 (exact, +mod, weighted) to 78.46 (partial, label) among the evaluation factors in Task1-CR-JA, most of which were achieved by our baseline, whereas SRCB-2 consistently achieved the best scores ranging from 51.81 to 78.80 in Task1-CR-EN. The best scores in Task1-RR-JA range from 46.74 to 96.39 (by the AMI systems and our baseline), whereas the scores of 52.62–92.93 (by SRCB-1 and 2) were the best in Task1-RR-EN. In Task2-CR-JA and Task2-RR-JA, AMI-1 outperformed our baseline, achieving the scores 36.44–61.63 and 49.58–88.43, respectively. The only participating system, Syapse-1, scored 49.67–82.89 in Task2-RR-EN. No system was submitted to Task2-RR-JA.

### **Task 3**

#### ***ADE application***

Tables 7 and 8 list the results of the ADE application on MedTxt-CR. Note that the test dataset does not contain any ADEval entities. Three systems, including our baseline, participated the Japanese corpus track (JA). At the entity level, the NICTmed systems performed better than our baseline in prioritizing precision; P and F tended to be higher. At the reporting level, our baseline achieved the best recall (77.78), whereas NICTmed-1 performed the best in terms of P (37.50) and F (48.00).

Further, 19 systems were submitted to the English track (EN). Syapse-1 achieved the best scores most frequently (i.e., in the five metrics: P in ADEval=0, F in ADEval=1, P in ADEval=3, and R and F at the reporting level). SRCB-2, 5, and 6 performed the best in several metrics. These scores were higher than those for JA, with an average difference of approximately 20–30 points.

## **CI application**

Table 9 shows the evaluation scores for Task3-RR, the CI application. In the Japanese corpus (JA), Zukyo-1 achieved the highest scores for all metrics, 34.09 in AdjMI and 36.22 in FM. For the English corpus (EN), FRDC-1 achieved 84.37 AdjMI and 84.36 FM. Overall, the EN scores tended to be higher than those of JA.

## **DISCUSSIONS**

### **Task 1**

A distinction in the nature of the corpus is evident in the higher scores achieved for the RR corpus compared to the CR corpus. This observation aligns with the tendency for radiology reports to exhibit linguistic simplicity when compared to case reports<sup>27</sup>. The CR corpus has a large vocabulary (7369 unique tokens) that covers most medical fields, whereas the RR corpus has a smaller number of unique tokens (i.e., 1182).

The performances of the top-tier systems in the two languages (*JA* vs. *EN*) were similar (an average difference of approximately 5 points), indicating that task difficulty was independent of the language in this size of training data (approximately 100 documents).

This could benefit from pre-training the language models, which will be discussed subsequently.

For the boundary factor (*exact* vs. *partial* match), the partial scores were at least 10 points higher than the exact scores, regardless of the corpus or language. Remarkably, the best scores for the modality-agnostic unweighted partial match were close to 80 for CR and 95

for RR. This indicates that the best systems captured medically important phrases, at least partially, despite the relatively small training data.

For the frequency factor (*weighted* or not), we did not observe a change in the rank of the top systems even after weighting (except for the partial unweighted modality-agnostic match in RR-JA), which suggests that the best systems did not rely too much on high-frequency entities.

Finally, we discuss the approaches adopted by the participating teams. Their common methods are (i) language models and (ii) data augmentation.

1. **Language models:** Almost all systems employ Transformer-based language models, and many teams adopt domain-specific pre-trained models, such as BioBERT and Clinical BERT in EN, and UTH-BERT in JA. Now that these pre-trained models drive contemporary NLP, even models without additional techniques, such as our baseline, perform well enough.
2. **Data augmentation:** Given the few-resource issues, many systems use data augmentation techniques. The results showed that machine translation-based methods (e.g., SRCB) contributed to the performance more than simple rule-based methods (e.g., FRDC). Owing to improvements in machine translation, round-trip translation would generate semantically correct samples; conversely, rule-based augmentation, such as random word swapping, might break the medical appropriateness of sentences.



## **Task 2**

While Task 1 provided a small corpus of approximately 100 *documents*, our new challenge, Task 2, only included approximately 100 *sentences* in the annotation guidelines for model training. This challenge can be observed in the exact match performance: even the best systems resulted in only 50–70% of the highest scores of their Task 1 counterparts.

However, the partial match scores of the best systems in Task 2 were rather close to those in Task 1, that is, within only 10-point difference in most cases. For instance, AMI-1 scored 61.63 (partial, +mod, unweighted) in CR-JA, which was an 8.14-point difference from our baseline of 69.77 in Task1-CR-JA. AMI-1 also achieved 88.43 (partial, label, and unweighted) in RR-JA, the performance of which seems sufficiently high for certain practical applications, such as medical concept-based document retrieval. Thus, this challenge revealed the potential feasibility of NER based on only a few samples for human annotators.

## **Task 3**

### ***ADE Application***

At the entity level, the average F-scores of the submitted systems were proportional to the number of corresponding ADEval entities in the training set ( Table 4). Report-level ADE performance tended to be inconsistent with entity-level performance; a better entity-level system was not necessarily a better report-level system. Although the corpora were parallel, most EN systems performed much better than the JA systems. For this task, the domain-specific language models effectively contributed to the results. Most EN systems

are based on medically pre-trained language models, such as BioBERT, clinical BERT, and PubMed BERT, whereas JA systems only adopt general-domain BERT models.

We then focused on effective approaches, particularly for EN. Regarding the F-scores for ADEval = 3 and at the report level, which mostly corresponded to ADE signal detection, the SRCB systems generally performed well (average of 61.2, ADEval = 3, and 52.3, respectively). They trained models on automatically generated snippets that explicitly explained which entity in a report was related to an ADE, which seemed to enhance the local and global ADE contexts. In addition, Syapse-1 performed best at the report level (64.00 F), whose method compares medicine and disease entities embedded by SciBERT per document; drug-disorder relations inside each document would contribute to report-level performance.

### ***CI application***

The system Zukyo-1 achieved the highest scores, suggesting the effectiveness of sentence classification in determining TNM staging, even with the limited availability of knowledge. FRDC-1, which uses heuristics for cancer size matching and Sentence-BERT encoding<sup>51</sup>, achieved the highest performance of all the systems. As shown in Table 10, the radiology reports of cases 4 and 5 were successfully grouped into a single cluster, suggesting that matching lesion size is helpful for case distinction.

Although both used a NER-based approach, a large discrepancy was observed between the scores of the GunNLP-1 and Syapse-1 systems. This may reflect differences in the availability of biomedical knowledge bases between Japanese and English. Whereas

Syapse-1 used UMLS to normalize biomedical entities, GunNLP-1 had to create bag-of-entity vectors only from the training set, which probably had difficulty dealing with unseen entities in the test set.

As listed in Table 11, most systems grouped the test cases into the same number of clusters as the gold standard, although the true cluster number was not clarified. In this task, the test sample size quickly determines the true cluster number, as exploited by FRDC-1.

In summary, the effective strategies differed between Japanese (RR-JA) and English (RR-EN). For the RR-EN, the embedding distance with the help of a knowledge base works well and can be applied in other clinical specialties beyond lung cancer. For the RR-JA, the lack of external public knowledge motivated participants to adopt a more dataset-specific approach, resulting in comparatively lower performance and a limited possibility of application beyond lung cancer.

### **Limitations**

Our workshop has two major limitations. First, relatively few teams participated in the new tasks that we designed: guideline learning, ADE, and CI. The numbers of participants in both languages were also unbalanced. Although few results prevented a finer analysis, we hope these tasks will attract more attention.

Second, we translated the original Japanese corpora into English to create bilingual parallel corpora for our tasks, which may have produced unnatural medical texts in English. It is generally known that the writing style of clinical documents varies in languages and nations. Our English corpora may deviate from the standard writing of typical English case

reports and radiology reports. However, we believe that medical parallel corpora will help international communities understand clinical writing styles in non-English languages, which is important for language-independent MedNLP applications in the future.

### **Clinical or Public Health Implications**

The designed tasks were oriented toward real-world clinical document processing. Although they do not directly affect patient health, the participating teams proposed MedNLP techniques to extract information useful for medical research and analysis from texts (e.g., phenotyping and ADE). In the future, application systems adapting these techniques will support the work and study of medical workers, benefiting patients.

## **Conclusions**

This study introduced the Real-MedNLP workshop, which encompassed three distinct medical NLP tasks conducted on bilingual parallel corpora (English and Japanese): named entity recognition (NER), adverse drug event extraction (ADE), and case identification (CI). The participating teams employed a dual approach, which involved (1) implementing data augmentation techniques and (2) utilizing domain-specific pre-trained language models like BioBERT and ClinicalBERT. These strategies partially addressed the challenges associated with limited resources in MedNLP. However, the performance in tasks involving extremely low-resource settings, such as Task 2 (guideline learning), remained insufficient. Specifically, for newly devised tasks like Task 3 ADE and CI applications, significant effort was required to establish evaluation methodologies that accurately captured their performance characteristics.

## **Future work**

Since our three tasks and other medical tasks are awaiting NLP solutions, organizing and sharing approaches and results worldwide is important. We believe that our datasets and results will boost future research. The results of this workshop provide the rigorous “baseline” for medical information extraction since it was held right before the rise of large language models (LLMs) such as ChatGPT<sup>65</sup> and Gemini<sup>66</sup>. By comparing their performance in our tasks with our results based on pre-LLM cutting-edge techniques, we can accurately gauge the capability of LLMs in low-resource medical NLP.

Furthermore, we worked on a successor of Real-MedNLP in NTCIR-17 (from 2022 to 2023), entitled “MedNLP-SC,” where “SC” stands for social media and clinical text<sup>67,68</sup>. This new workshop posed information extraction from patient-generated and doctor-generated texts, where the low-resource setting is still active, given our experience in Real-MedNLP. Evaluating and comparing the outcomes of this workshop with those of the current one will be another focus for future research.

### **Abbreviations**

ADE: adverse drug event

CI: case identification

CR: case reports

EHR: electronic health records

EN: English

HR: health records

JA: Japanese

MedNLP: medical natural language processing

NEN: named entity normalization

NER: named entity recognition

NLP: natural language processing

RR: radiology reports

TC: text classification

## **Funding/Acknowledgements**

This work was supported by JST AIP Trilateral AI Research Grant Number JPMJCR20G9 and MHLW Program Grant Number JPMH21AC500111 (formerly JST AIP-PRISM Grant Number JPMJCR18Y1), Japan.

The authors also greatly appreciate the NTCIR-16 chairs for their efforts in organizing the NTCIR-16 conference. Finally, the authors thank all participants for their contributions to our Real-MedNLP workshop.

### **Conflict of Interest**

We have nothing to disclose in terms of the conflict of interest.

## **REFERENCES**

1. Aramaki E, Wakamiya S, Yada S, Nakamura Y. Natural language processing: From bedside to everywhere. *Yearb Med Inform.* 2022;31(01): 243-253
2. Suominen H, Salanterä S, Velupillai S, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: *Information Access Evaluation. Multilinguality, Multimodality, and Visualization.* Springer Berlin Heidelberg; 2013:212-231
3. Névéal A, Cohen KB, Grouin C, et al. Clinical information extraction at the CLEF eHealth evaluation lab 2016. *CEUR workshop proceedings.* 2016;1609:28-42
4. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc.* 2020;27(1):3-12

5. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Saf.* 2019;42(1):99-111
6. Ben Abacha A, Mrabet Y, Zhang Y, Shivade C, Langlotz C, Demner-Fushman D. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In: *Proceedings of the 20th Workshop on Biomedical Language Processing. Association for Computational Linguistics; 2021:74-85*
7. He B, Dong B, Guan Y, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. *J Biomed Semantics.* 2017;69:203-217
8. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat AL, Névéol A. A French clinical corpus with comprehensive semantic annotations: Development of the medical entity and relation LIMSI annotated text corpus (MERLOT). *Lang Resour Eval.* 2018;52(2):571-601
9. Oliveira LESE, Peters AC, Silva AMP da, et al. SemClinBr - a multi-institutional and multi-speciality semantically annotated corpus for Portuguese clinical NLP tasks. *J Biomed Semantics.* 2022;13(1):13
10. Morita M, Kano Y, Ohkuma T, Miyabe M, Aramaki E. 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); 2013*
11. Morita M, Kano Y, Ohkuma T, Aramaki E. 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); 2014

12. Morita M, Kano Y, Ohkuma T, Aramaki E. 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); 2016
13. Wakamiya S, Morita M, Kano Y, Ohkuma T, Aramaki E. 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); 2017
14. Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc.* 2010;17(5):519-523
15. Hedderich MA, Lange L, Adel H, Strötgen J, and Klakow D. A survey on recent approaches for natural language processing in low-resource scenarios. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021:2545–2568
16. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc.* 2016;23(6):1046-1052
17. Xue H, Li J, Xie H, Wang Y. Review of drug repositioning approaches and resources. *Int J Biol Sci.* 2018;14(10):1232-1244

18. Öztürk H, Özgür A, Schwaller P, Laino T, Ozkirimli E. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discov Today*. 2020;25(4):689-705
19. Roberts K, Demner-Fushman D, Voorhees EM, et al. Overview of the TREC 2017 precision medicine track. *The text REtrieval conference: TREC Text REtrieval Conference*. 2017;26
20. Biswal S, Xiao C, Glass LM, Westover B, Sun J. CLARA: Clinical report auto-completion. In: *Proceedings of the Web Conference 2020. WWW '20*. Association for Computing Machinery; 2020:541-550
21. Yazdani A, Safdari R, Golkar A, R Niakan Kalhori S. Words prediction based on n-gram model for free-text entry in electronic health records. *Health Inf Sci Syst*. 2019;7(1):6
22. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *J Am Med Inform Assoc*. 2015;22(5):938-947
23. J-Stage. <https://www.jstage.jst.go.jp/> [accessed 2023-04-18]
24. Ito K, Nagai H, Okahisa T, Wakamiya S, Iwao T, Aramaki E. J-MeDic: A Japanese disease name dictionary based on real clinical usage. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA); 2018. <https://aclanthology.org/L18-1375>
25. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: A systematic review. *Radiology*. 2016;279(2):329-343

26. Nakamura Y, Hanaoka S, Nomura Y, et al. Clinical comparable corpus describing the same subjects with different expressions. *Stud Health Technol Inform.* 2022;290:253-257
27. Yada S, Joh A, Tanaka R, Cheng F, Aramaki E, Kurohashi S. Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: Starting from critical lung diseases. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association; 2020:4565-4572. <https://www.aclweb.org/anthology/2020.lrec-1.561>
28. Yada S, Aramaki E, Tanaka R, Cheng F, Kurohashi S. *Medical/Clinical Text Annotation Guidelines.*; 2021
29. Yada S, Tanaka R, Cheng F, Aramaki E, Kurohashi S. Versatile annotation guidelines for clinical-medical text with an application to critical lung diseases [in Japanese]. *Journal of Natural Language Processing.* 2022;29(4):1165-1197. doi:10.5715/jnlp.29.1165
30. Kelly CR, Kunde SS, Khoruts A. Guidance on preparing an investigational new drug application for fecal microbiota transplantation studies. *Clin Gastroenterol Hepatol.* 2014;12(2):283-288
31. NTCIR test collection. <https://research.nii.ac.jp/ntcir/data/data-en.html>
32. Nishiyama T, Nishidani M, Ando A, Yada S, Wakamiya S, Aramaki E. NAISTSOC at the NTCIR-16 Real-MedNLP task. In: *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*; 2022:330-333
33. Devlin J, Chang MW, Lee K, Toutanova K. 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; 2019:4171-4186. doi:10.18653/v1/N19-1423

34. BERT models for Japanese NLP. <https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking>. [accessed 2023-04-18]
35. MeCab. <https://taku910.github.io/mecab/> [accessed 2023-04-18]
36. MedNER-CR-JA. <https://huggingface.co/sociocom/MedNER-CR-JA> [accessed 2023-04-18]
37. Hiai S, Nagayama S, Kojima A. AMI team at the NTCIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:297-304
38. Zhong Z, Fang L, Cao Y. FRDC at NTCIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:311-315
39. Noguchi R. GunNLP at the NTCIR-16 Real-MedNLP task: Collaborative filtering-based similar case identification method via structured data “case matrix.” In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:349-352
40. Ideuchi M, Tsuchiya M, Wang Y, Utiyama M. NICTmed at the NCTIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:339-344

41. Shao S, Jin G, Satoh D, Nomura Y. NTTD at the NTCIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:345-348
42. Zhang Y, Cheng R, Luo L, Gao H, Jiang S, Dong B. SRCB at the NTCIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:305-310
43. Holmes B, Gagorik A, Loving J, Green F, Huang H. Syapse at the NCTIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:334-338
44. Fujimoto K, Nishio M, Sugiyama O, et al. Approach for named entity recognition and case identification implemented by ZuKyo-JA sub-team at the NTCIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:322-329
45. Nooralahzadeh F, Horvath AN, Krauthammer M. Leveraging token-based concept information and data augmentation in few-resource NER: ZuKyo-EN at the NTCIR-16 Real-MedNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.; 2022:316-321
46. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. PLoS One. 2021;15(11):e0259763
47. Hyakuyaku-Dictionary. <https://sociocom.naist.jp/hyakuyaku-dic/> [accessed 2023-04-18]
48. ComeJisyo. <https://ja.osdn.net/projects/comedic/> [accessed 2023-04-18]

49. Lee J, Yoon W, Kim S, et al. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240
50. Dai X, Adel H. An analysis of simple data augmentation for named entity recognition. In: *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics; 2020:3861-3867. doi:10.18653/v1/2020.coling-main.343
51. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019:3982-3992. doi:10.18653/v1/D19-1410
52. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:8440-8451. doi:10.18653/v1/2020.acl-main.747
53. Devlin J, Chang MW, Lee K, Toutanova K. 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; 2019:4171-4186. doi:10.18653/v1/N19-1423

54. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics; 2019:72-78. doi:10.18653/v1/W19-1909
55. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM trans comput healthc.* 2021;3(1):1-23
56. Lin C, Miller T, Dligach D, Bethard S, Savova GK. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In: BIONLP.; 2021
57. Zhou B, Cui Q, Wei XS, Chen ZM. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. Published online 2019. doi:10.48550/ARXIV.1912.02413
58. Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001:17-21.
59. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and robust models for biomedical natural language processing. Published online February 2019. <https://arxiv.org/abs/1902.07669>
60. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. Published online March 2019. <https://arxiv.org/abs/1903.10676>
61. Brierley JD, Gospodarowicz MK, Wittekind C. TNM Classification of Malignant Tumours. John Wiley & Sons; 2017
62. Zhuang L, Wayne L, Ya S, Jun Z. A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on

Computational Linguistics. Chinese Information Processing Society of China;  
2021:1218-1227. <https://aclanthology.org/2021.ccl-1.108>

63. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09. Association for Computing Machinery; 2009:1073-1080. doi:10.1145/1553374.1553511
64. Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. *J Am Stat Assoc.* 1983;78(383):553-569. doi:10.1080/01621459.1983.10478008
65. ChatGPT. <https://chat.openai.com/>
66. Gemini Team Google. Gemini: A family of highly capable multimodal models. Published online December 2023. <https://arxiv.org/abs/2312.11805>
67. Nakamura Y, Hanaoka S, Yada S, Wakamiya S, Aramaki E. NTCIR-17 MedNLP-SC Radiology Report Subtask overview: Dataset and solutions for automated lung cancer staging. In: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies; 2022:145-151. doi:10.20736/0002001328
68. Wakamiya S, Pereira LK, Raithel L, Yeh HS, Han P, Shimizu S, et al. 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; 2022:131-141. doi:10.20736/0002001327



**Figure 1** Annotated sample of the case report corpora in English (MedTxt-CR-EN). The entity notations stand for D = diseases and symptoms with the modality “certainty” such as positive (+) and negative (-); A = anatomical parts; Time = time expressions with the modality “type” such as date (DATE), age (AGE), and medically specific (MED); Tt/k/v = Test set/item/values with the modality “state” such as executed (+); and Mk = Medicine name with the modality “state” such as executed (+).

**Figure 2** Annotated sample of the radiology report corpora in English (MedTxt-RR-EN). The entity notations correspond stand for D = diseases and symptoms with the modality “certainty” such as positive (+) and suspicious (?); A = anatomical parts; Time = time expressions with the modality “type” such as date (DATE); and Tt = Test sets with the modality “state” such as executed (+).

**Figure 3** Task 3 - ADE application, wherein each disease or medication entity mentioned in case reports is labeled with the degree of involvement in adverse drug events (ADEval), ranging from 0 to 3.

**Figure 4** Task 3 - CI application, where radiology reports written by different radiologists are grouped by the described cases.

**Figure 5** Input and output formats of the baseline model for Task3-CR (ADE). Although the real inputs were in Japanese, an English sample is used in this figure for readability.

**Figure 6** Input and output formats of the baseline model for Task3-RR (CI). Although the real inputs were in Japanese, an English sample is used in this figure for readability.

**Table 1** Past MedNLP Series workshops proposed by the authors

Workshop	Year	Corpus	Task
MedNLP-1 <sup>10</sup>	2012–13	dummy HR written by clinicians	NER
MedNLP-2 <sup>11</sup>	2013–14	dummy HR written by clinicians	NEN
MedNLPDoc <sup>12</sup>	2015–16	dummy HR extracted from clinical textbooks	NEN
MedWeb <sup>13</sup>	2016–17	dummy Tweets obtained by crowdsourcing	TC

**Table 2** Real-MedNLP Tasks

Task	Corpus	Format
1) Just 100 Training	CR/RR	NER
2) Guideline Learning	CR/RR	NER
3) Applications	CR	ADE
	RR	CI

**Table 3** Named entities of the training sets of MedTxt-CR and MedTxt-RR

	Dataset	CR-JA	CR-EN	RR-JA	RR-EN
	# of texts	148	148	72	72
	# of characters	84471	40383	16861	8488
	(mean)	(570)	(272)	(234)	(117)
<a>	total	823	819	464	465
<d>	total	2348	2346	884	883
	"positive"	1695	1693	465	462
	"suspicious"	80	80	191	191
	"negative"	251	251	149	148
	"general"	302	302	1	1
<t-test>	total	387	388	26	27

	"scheduled"	0	0	0	0
	"executed"	362	363	19	19
	"negated"	7	7	2	2
	"other"	18	18	5	6
<b>&lt;timex3</b>	total	1353	1353	29	29
<b>&gt;</b>					
	"date"	539	539	26	26
	"time"	53	53	0	0
	"duration"	82	82	2	2
	"set"	34	34	0	0
	"age"	189	189	0	0
	"med"	428	428	1	1
	"misc"	28	28	0	0
<b>&lt;m-key&gt;</b>	total	344	344	0	0
	"scheduled"	0	0	0	0
	"executed"	266	266	0	0
	"negated"	27	27	0	0
	"other"	51	51	0	0
<b>&lt;m-val&gt;</b>	total	64	64	0	0
	"scheduled"	0	0	0	0
	"executed"	0	0	0	0
	"negated"	2	2	0	0
	"other"	0	0	0	0
<b>&lt;t-key&gt;</b>	total	524	524	1	1
<b>&lt;t-val&gt;</b>	total	427	427	0	0

**Table 4** ADEval distributions for each entity type in the training set of Task 3 ADE application

ADEva	Disease	Medicine	(total)
<b>1</b>			
0	1217	103	1320
1	33	28	61
2	57	22	79
3	123	47	170
total	1430	200	1630

**Table 5** Team demographic and the number of systems developed by each team. \* stands for multidisciplinary (medicine + computer science) teams

Team	Team demographic			# of submitted systems											
				Task 1				Task 2				Task 3			
	#members	Country	Affiliation	CR		RR		CR		RR		CR		RR	
			JA	EN	JA	EN	J	EN	J	EN	JA	EN	J	EN	
AMI	3	Japan	Industry	2		2		1		1					
FRDC	4	China	Industry		2							10		10	
GunNLP	1	Japan	University											1	
Baseline*	6	Japan	University	1		1		1		1		1		1	
NICTmed	4	Japan	Institute	4	4							2	2	1 1	
NTTD	4	Japan	Industry	1		1									
SRCB	6	China	Industry		5		3					6			
Syapse	5	US	Industry							1		1		1	
Zukyo*	11	Japan & Switzerland	University & Institute	4	4	4	4							1	
			<b>Total</b>	12	15	8	7	2	0	2	1	3	19	4 12	

**Table 6** Results of Tasks 1 and 2. Bold font indicates the best score for each evaluation metric

Task	Corpus	Language	System ID	Exact match				Partial match			
				label	weighted	+mod	weighted	label	weighted	+mod	weighted
1	CR	JA	AMI-1	61.3	51.95	-	-	78.4	68.12	-	-
			3				1				
			AMI-2	61.2	51.88	-	-	78.4	68.19	-	-
			4				6				
			Baseline	65.2	55.50	59.2	49.93	77.2	66.89	69.7	59.93
			5		1		7		7		
			NICTmed-1	56.9	47.37	52.4	43.33	72.6	62.30	65.5	55.74
			6		9		7		2		
			NICTmed-2	60.7	50.48	56.0	46.21	72.5	61.64	65.9	55.62
			6		2		7		6		
NICTmed-3	55.5	46.50	51.7	43.15	75.2	64.89	68.2	58.50			
0		1		2		8					
NICTmed-	58.1	48.63	54.2	45.15	74.6	63.81	68.2	57.96			

			<b>4</b>	3		0		4	1		
			<b>NTTD-1</b>	61.8	51.98	-	-	73.6	62.93	-	
				9				1			
			<b>Zukyo-1</b>	30.8	23.83	25.9	19.63	55.1	47.12	44.8	37.77
				8		1		4		8	
			<b>Zukyo-2</b>	35.8	29.68	30.1	24.59	63.9	56.33	53.0	46.25
				5		3		5		7	
			<b>Zukyo-3</b>	26.5	21.95	22.4	18.36	58.6	52.04	48.2	42.32
				6		7		5		0	
			<b>Zukyo-4</b>	27.7	23.34	23.0	19.10	59.6	53.46	49.6	44.03
				3		8		7		3	
		<b>EN</b>	<b>FRDC-1</b>	43.2	38.50	-	-	56.4	51.24	-	-
				1				8			
			<b>FRDC-2</b>	43.7	38.90	-	-	56.5	51.22	-	-
				1				5			
			<b>NICTmed-1</b>	46.8	40.92	42.4	37.01	69.9	62.80	62.4	55.83
				3		5		9		2	
			<b>NICTmed-2</b>	48.6	42.47	44.0	38.43	69.9	62.52	62.9	56.16
				0		6		0		5	
			<b>NICTmed-3</b>	49.1	43.26	44.8	39.38	72.3	65.28	64.8	58.40
				8		0		9		6	
			<b>NICTmed-4</b>	51.4	45.25	46.9	41.27	71.4	64.04	64.8	58.08
				5		6		2		1	
			<b>SRCB-1</b>	59.8	52.55	54.8	48.09	73.7	65.35	67.6	59.94
				0		4		2		9	
			<b>SRCB-2</b>	63.3	56.16	58.5	51.81	78.8	70.42	72.6	64.88
				7		3		0		9	
			<b>SRCB-3</b>	62.3	55.15	57.4	50.80	77.9	69.47	71.8	63.94
				1		9		0		1	
			<b>SRCB-4</b>	59.3	52.65	54.5	48.31	77.8	70.05	71.5	64.35
				3		2		4		6	
			<b>SRCB-5</b>	60.3	53.64	55.4	49.17	78.2	70.34	71.8	64.44
				3		0		5		0	
			<b>Zukyo-1</b>	45.5	39.65	29.5	25.89	70.3	63.03	44.7	40.05
				6		7		2		9	
			<b>Zukyo-2</b>	51.9	45.89	33.3	29.50	73.7	66.38	47.1	42.28
				7		5		6		1	
			<b>Zukyo-3</b>	51.1	44.78	32.6	28.67	72.2	64.53	46.0	41.11
				6		3		0		9	
			<b>Zukyo-4</b>	49.1	43.17	30.7	27.05	71.9	64.55	45.2	40.46
				8		7		1		6	
	<b>RR</b>	<b>JA</b>	<b>AMI-1</b>	15.0	11.65	-	-	96.3	56.68	-	-
				5				9			
			<b>AMI-2</b>	89.2	51.81	-	-	96.1	57.69	-	-
				6				4			

			<b>Baseline</b>	84.8	48.71	80.7	46.74	92.6	55.36	87.7	52.81
				8		9		9		8	
			<b>NTTD-1</b>	87.0	49.92	-	-	93.8	55.80	-	-
				3				5			
			<b>Zukyo-1</b>	58.1	31.91	42.5	25.50	82.0	49.71	57.2	36.93
				1		9		1		7	
			<b>Zukyo-2</b>	60.2	32.78	43.6	25.72	83.7	50.78	58.9	37.81
				2		3		0		4	
			<b>Zukyo-3</b>	57.7	31.27	42.2	24.80	82.1	50.03	58.5	37.64
				9		4		3		7	
			<b>Zukyo-4</b>	56.7	30.96	42.1	24.77	82.0	50.24	58.8	38.03
				4		6		1		4	
		<b>EN</b>	<b>SRCB-1</b>	82.6	54.96	79.1	52.62	92.8	64.02	88.6	60.95
				0		9		6		2	
			<b>SRCB-2</b>	82.6	55.00	78.7	52.31	92.9	64.06	88.0	60.59
				6		4		3		5	
			<b>SRCB-3</b>	80.6	53.58	77.1	51.05	92.2	63.88	87.8	60.50
				1		9		4		7	
			<b>Zukyo-1</b>	75.9	49.57	63.5	41.07	90.8	63.16	74.1	50.62
				2		0		5		0	
			<b>Zukyo-2</b>	79.9	52.99	67.0	44.00	91.3	63.25	75.5	51.63
				7		7		2		1	
			<b>Zukyo-3</b>	78.7	51.92	65.3	42.64	91.5	63.46	74.6	51.04
				7		2		6		9	
			<b>Zukyo-4</b>	78.9	52.45	65.4	43.09	91.7	63.81	75.1	51.65
				5		5		0		3	
<b>2</b>	<b>CR</b>	<b>JA</b>	<b>AMI-1</b>	37.1	36.44	37.1	36.44	61.6	60.91	61.6	60.91
				0		0		3		3	
			<b>Baseline</b>	25.1	24.74	19.4	19.12	45.8	45.47	34.6	34.24
				2		9		9		4	
	<b>RR</b>	<b>JA</b>	<b>AMI-1</b>	64.8	62.17	51.3	49.58	88.4	85.71	68.6	66.85
				5		3		3		4	
			<b>Baseline</b>	62.5	60.13	46.6	44.62	82.8	80.39	60.9	58.80
				5		8		9		4	
		<b>EN</b>	<b>Syapse-1</b>	54.9	54.22	50.3	49.68	82.8	81.79	75.9	74.95
				6		7		9		9	

**Table 7** Results of Task 3 for MedTxt-CR-JA. *Italic font* indicates the best score for each evaluation metric

System ID	ADEval=0			ADEval=1			ADEval=3			Report-level		
	P	R	F	P	R	F	P	R	F	P	R	F
<b>Baseline</b>	95.2	76.0	84.5	0.00	0.0	0.0	6.98	52.9	12.3	12.7	77.7	21.88

	1	4	5		0	0		4	3	3	8	
<b>NICTmed-1</b>	95.7	97.6	96.7	0.00	0.0	0.0	12.5	11.7	12.1	37.5	66.6	48.00
	6	7	1		0	0	0	6	2	0	7	
<b>NICTmed-2</b>	96.0	97.0	96.5	0.00	0.0	0.0	27.5	47.0	34.7	25.0	44.4	32.00
	5	0	2		0	0	9	6	8	0	4	

**Table 8.** Results of Task 3 for MedTxt-CR-EN. *Italic font* indicates the best score for each evaluation metric.

System ID	ADEval=0			ADEval=1			ADEval=3			Report-level		
	P	R	F	P	R	F	P	R	F	P	R	F
<b>FRDC-1</b>	95.7	94.94	95.3	20.00	5.26	8.33	62.5	26.3	37.04	22.2	66.67	33.33
	0		2				0	2		2		
<b>FRDC-2</b>	95.7	97.00	96.3	14.29	5.26	7.69	43.7	36.8	40.00	29.4	55.56	38.46
	9		9				5	4		1		
<b>FRDC-3</b>	95.9	93.52	94.7	6.25	5.26	5.71	28.5	21.0	24.24	19.3	66.67	30.00
	5		2				7	5		5		
<b>FRDC-4</b>	96.0	92.10	94.0	25.00	5.26	8.70	22.2	42.1	29.09	18.9	77.78	30.43
	5		3				2	1		2		
<b>FRDC-5</b>	95.8	95.26	95.5	0.00	0.00	0.00	56.2	47.3	51.43	25.9	77.78	38.89
	7		6				5	7		3		
<b>FRDC-6</b>	96.1	94.47	95.3	25.00	10.5	14.81	50.0	21.0	29.63	21.2	77.78	33.33
	4		0		3		0	5		1		
<b>FRDC-7</b>	95.6	94.31	94.9	0.00	0.00	0.00	33.3	26.3	29.41	19.3	66.67	30.00
	7		9				3	2		5		
<b>FRDC-8</b>	96.4	97.79	97.1	20.00	5.26	8.33	47.6	52.6	50.00	50.0	77.78	60.87
	2		0				2	3		0		
<b>FRDC-9</b>	96.3	91.79	94.0	0.00	0.00	0.00	23.8	52.6	32.79	18.9	77.78	30.43
	5		1				1	3		2		
<b>FRDC-10</b>	95.8	95.26	95.5	7.14	5.26	6.06	26.9	36.8	31.11	23.0	66.67	34.29
	7		6				2	4		8		
<b>NICTmed-1</b>	96.5	96.68	96.6	0.00	96.6	0.00	31.2	52.6	39.22	25.0	55.56	34.48
	3		1		8		5	3		0		
<b>NICTmed-2</b>	95.3	98.10	96.7	0.00	0.00	0.00	40.0	42.1	41.03	40.0	44.44	42.11
	9		3				0	1		0		
<b>SRCB-1</b>	96.5	97.95	97.2	14.29	5.26	7.69	60.0	63.1	61.54	50.0	66.67	57.14

	7		5				0	6		0		
<b>SRCB-2</b>	96.5	97.95	97.2	0.00	0.00	0.00	59.0	68.4	63.41	50.0	66.67	57.14
	7		5				9	2		0		
<b>SRCB-3</b>	96.2	98.10	97.1	0.00	0.00	0.00	60.0	63.1	61.54	50.0	55.56	52.63
	8		8				0	6		0		
<b>SRCB-4</b>	96.4	97.63	97.0	0.00	0.00	0.00	57.1	63.1	60.00	50.0	66.67	57.14
	1		2				4	6		0		
<b>SRCB-5</b>	95.8	99.37	97.6	0.00	0.00	0.00	78.5	57.8	66.67	60.0	33.33	42.86
	8		0				7	9		0		
<b>SRCB-6</b>	95.9	98.26	97.1	33.33	5.26	9.09	55.5	52.6	54.05	50.0	44.44	47.06
	9		1				6	3		0		
<b>Syapse-1</b>	97.0	97.63	97.3	30.00	31.5	30.77	100.	26.3	41.67	50.0	88.89	64.00
	2		2		8		0	2		0		

**Table 9** Performance of each system of Task 3 for MedTxt-RR (CI) in multiple evaluation metrics. *Italic font* indicates the best score for each evaluation metric

Language	System ID	AdjMI	FM	Binary Acc
JA	GunNLP-1	0.1988	0.267	0.7675
			4	
	Baseline	0.1489	0.181	0.8131
			4	
	NICTmed-1		0.117	0.7680
			0	
	Zukyo-1	0.3409	0.362	0.8285
			2	
EN	FRDC-1	0.8437	0.843	0.9595
			6	
	GunNLP-2	0.8116	0.811	0.9508
			0	
	GunNLP-3	0.8122	0.812	0.9514
			6	
	GunNLP-4	0.8122	0.812	0.9514



			6	
	GunNLP-5	0.8122	0.812	0.9514
			6	
	GunNLP-6	0.8122	0.812	0.9514
			6	
	GunNLP-7	0.8261	0.816	0.9524
			6	
	GunNLP-8	0.8123	0.811	0.9514
			9	
	GunNLP-9	0.8123	0.811	0.9514
			9	
	GunNLP-10	0.8255	0.815	0.9519
			0	
	NICTmed-1		0.108	0.7809
			5	
	Syapse-1	0.7309	0.699	0.9206
			2	
<b>Extreme prediction (Isolate all samples)</b>			0.000	0.0000
			0	

**Table 10** Number of clusters into which each case was split by each system in Task 3 for MedTxt-RR (CI)

Case ID	4	5	7	8	10	14	15
TNM	T2aN0M0	T2bN0M0	T3N1M0	T3N3M	T4N0M0	T4N3M1a	T2N2M1c
				0			
<b>GunNLP-1</b>	5	6	3	3	3	5	2
<b>Baseline</b>	6	7	5	8	6	5	5
<b>NICTmed-1</b>	6	5	5	6	5	5	5
<b>Zukyo-1</b>	4	5	3	4	4	3	2
<b>FRDC-1</b>	1	1	2	2	1	2	3



Case Report: **Time(AGE)** 60 years old , male.

Chief Complaint: **D(+)** Positive fecal occult blood .

Current Medical History: Patient had **D(+)** psoriasis vulgaris for **Time(AGE)** 5 years and had been treated with topical **Mk(+)** steroids and **Mk(+)** adalimumab at the dermatology department of our hospital.

In **Time(AGE)** September 2012 , he was found to be **D(+)** positive for occult blood in the stool .

**Tt(+)** Colonoscopy revealed a **D(+)** 0-IIa lesion in the **A** cecum and a **D(+)** type 1 lesion in the **A** sigmoid colon , and the patient was referred to surgery .

Past Medical History: **D(+)** Hypertension, hyperuricemia .

Present Condition: **Time(MED)** at Admission : **Tk** Height **Tv** 159.5 cm , **Tk** weight **Tv** 66.1 kg , **Tk** body temperature **Tv** 35.6°C , **Tk** blood pressure **Tv** 149/99 mmHg , **Tk** pulse **Tv** 68 beats/min .

**A** Abdomen : flat and soft , without **D(-)** tenderness .

**D(+)** Emphysematous changes are prominent in **A** both lungs , **A** mainly on the periphery .

An irregular **D(+)** intrapulmonary mass of  $\text{O} \times \text{O}$  cm in size is found in the **A** right lung apex , and **D(?)** lung cancer is suspected. The **D(+)** mass **D(+)** invades the **A** right 2nd and 3rd ribs and the Th2 and 3 vertebral bodies , and at the **A** Th2 level , it is suspected to **D(?)** extend **A** into the spinal canal ( **D** cT4 ).


Most of the **D(+)** nodules in the **A** right middle lobe and the outer layers of the lower lobes of both lungs have a transbronchial distribution, and **D(?)** inflammatory nodules are suspected. This is the patient's **Time(AGE)** first **Tt(+)** CT scan , so please follow the progress to see if this increases. There is nothing else to report in the **A** lungs or mediastinum .

## Case reports

Some may describe adverse drug events (ADEs)

Predict the certainty of involving ADEs (ADEval) for each disease and drug

### Given



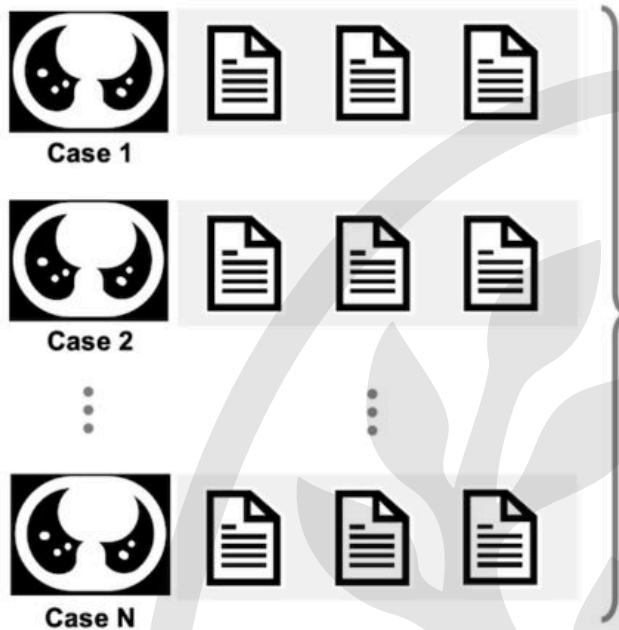
Case Report: **Time(AGE)** 14 year old , male.  
Chief Complaint: **D(+)** Fever, generalized erythema .  
Past Medical History: **D(+)** Mild Intellectual Disability .  
Current Medical History: In **Time(AGE)** April 2001 the patient was **D(+)** having **D(+)** epileptic seizures and the administration of **D(+)** valproic acid (VPA) was started **Time(AGE)** on April 20 .  
**Time(AGE)** Subsequently , due to it becoming difficult to control the patient's **D(+)** convulsions , concomitant use of **D(+)** CBZ was started on **Time(AGE)** July 8 .  
On **Time(AGE)** July 23 , **D(+)** fever and erythema **D(+)** presented , **Time(AGE)** Subsequently , **D(+)** liver dysfunction and thrombocytopenia were also observed.  
The patient was **D(+)** admitted and seen at our department on **Time(AGE)** August 3 .  
He presented with **D(+)** facial edema, lymphadenopathy coupled with **D(+)** downward trending lab results.  
Lab Results **Time(AGE)** Upon Admission : **Tk** WBC **Tv** 21,700/uL (eosinophil% **A** 12.5%) , **Tk** Hb **Tv** 169,000/uL , **Tk** TP **Tv** 5.2 g/dL , **Tk** AST **Tv** 1371 U/L , **Tk** ALT **Tv** 2021 U/L , **Tk** LDH **Tv** 7141 U/L , **Tk** CRP **Tv** 1.8 mg/dL , **Tk** IL-2R **Tv** 13,100 U/ml , **Tk** 2-SAS **Tv** 213 pMol/dL .  
Progress: The **D(+)** antiepileptic drug CBZ was **D(+)** discontinued with **D(+)** VPA alone being administered.  
**Time(AGE)** mPSL pulse therapy was given for **Time(AGE)** 3 days and the **D(+)** fever resolved .  
Symptoms and lab results also **D(+)** improved .  
Follow-up treatment started with **Time(AGE)** 30 mg/day of **D(+)** PSL , but because **D(+)** fever and skin rash **D(+)** once again presented the medication was changed to **Time(AGE)** 8 mg/day of **D(+)** betamethasone .

Disease	ADEval
Fever, generalized erythema	3
Mild Intellectual Disability	0
epileptic seizures	0
convulsions	0
fever and erythema	3
liver dysfunction and thrombocytopenia	3

Medicine	ADEval
valproic acid (VPA)	1
CBZ	3
VPA	1
PSL	2
betamethasone	0

Radiology reports by different radiologists

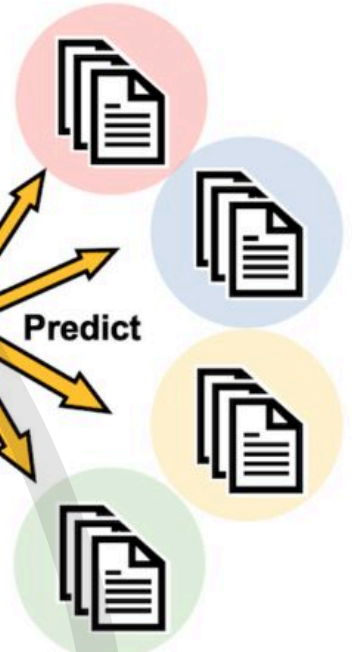


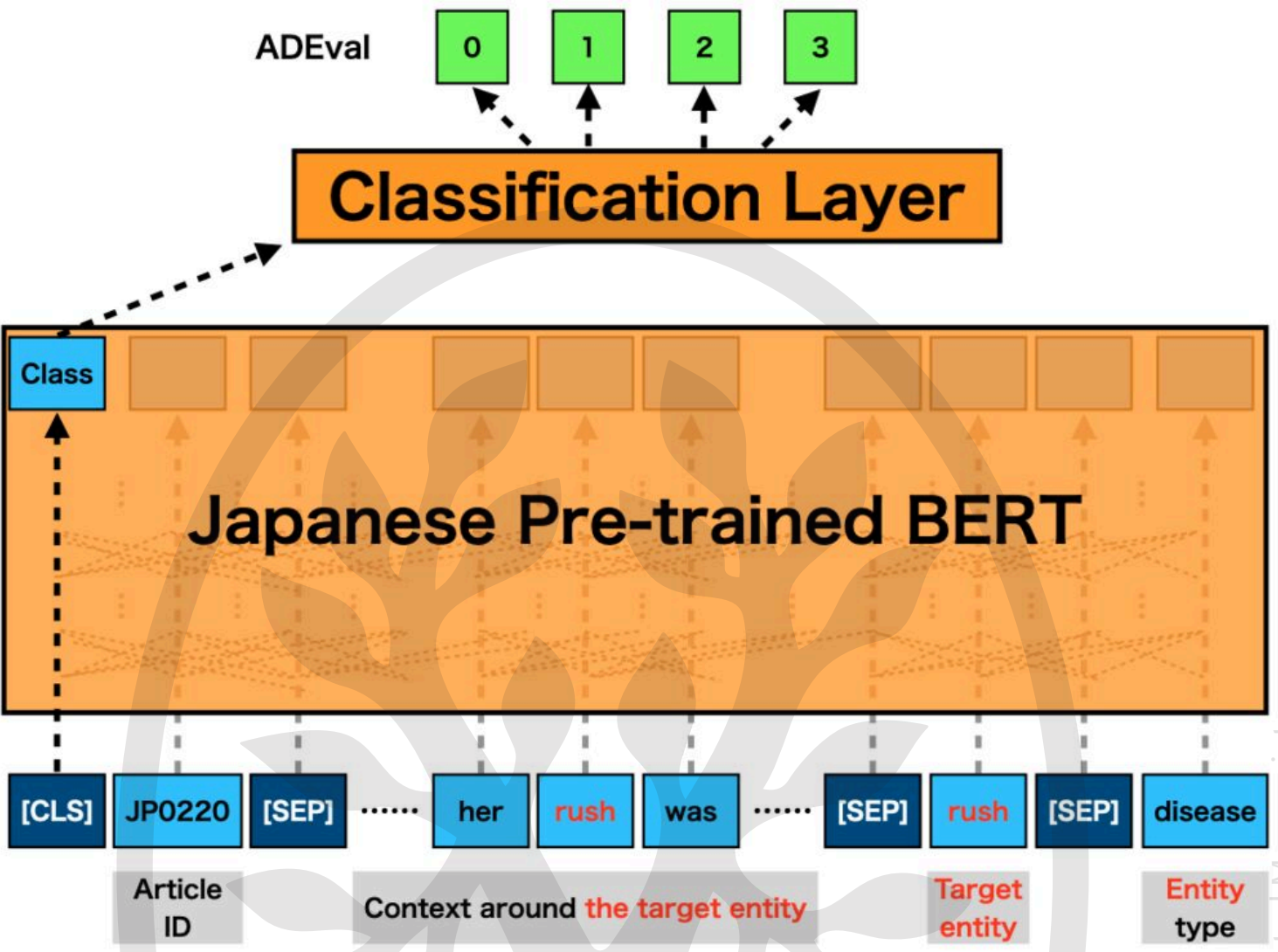
Group radiology reports of the same cases

Given



Predict





Do A and B describe  
the same patient?

True False

Classification Layer



[CLS] CT scan ..... found [SEP] The left ..... canal

Article A

Article B