BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers

Yuta Nakamura[1,2*], Shouhei Hanaoka[1,2], Yukihiro Nomura[3], Takahiro Nakao[3], Soichiro Miki[3], Takeyuki Watadani[1,2], Takeharu Yoshikawa[3], Naoto Hayashi[3] and Osamu Abe[1,2]

Abstract

Background: It is essential for radiologists to communicate actionable findings to the referring clinicians reliably. Natural language processing (NLP) has been shown to help identify free-text radiology reports including actionable findings. However, the application of recent deep learning techniques to radiology reports, which can improve the detection performance, has not been thoroughly examined. Moreover, free-text that clinicians input in the ordering form (order information) has seldom been used to identify actionable reports. This study aims to evaluate the benefits of two new approaches: (1) bidirectional encoder representations from transformers (BERT), a recent deep learning architecture in NLP, and (2) using order information in addition to radiology reports.

Methods: We performed a binary classification to distinguish actionable reports (i.e., radiology reports tagged as *actionable* in actual radiological practice) from non-actionable ones (those without an *actionable* tag). 90,923 Japanese radiology reports in our hospital were used, of which 788 (0.87%) were actionable. We evaluated four methods, statistical machine learning with logistic regression (LR) and with gradient boosting decision tree (GBDT), and deep learning with a bidirectional long short-term memory (LSTM) model and a publicly available Japanese BERT model. Each method was used with two different inputs, radiology reports alone and pairs of order information and radiology reports. Thus, eight experiments were conducted to examine the performance.

Results: Without order information, BERT achieved the highest area under the precision-recall curve (AUPRC) of 0.5138, which showed a statistically significant improvement over LR, GBDT, and LSTM, and the highest area under the receiver operating characteristic curve (AUROC) of 0.9516. Simply coupling the order information with the radiology reports slightly increased the AUPRC of BERT but did not lead to a statistically significant improvement. This may be due to the complexity of clinical decisions made by radiologists.

Conclusions: BERT was assumed to be useful to detect actionable reports. More sophisticated methods are required to use order information effectively.

Keywords: Radiology reports, Actionable finding, Natural language processing (NLP), Bidirectional encoder representations from transformers (BERT), Deep learning

Background

A radiology report may include an actionable finding that is critical if left overlooked by the referring clinician [1]. However, clinicians can fail to see mentions of actionable findings in radiology reports for various reasons,

*Correspondence: yutanakamura-tky@umin.ac.jp
[1] Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan
Full list of author information is available at the end of the article

and such failure in communication can delay further procedures and impact the prognosis of the patient [2]. Therefore, fast and reliable communication on actionable findings is essential in clinical practice.

Information technologies are helpful in identifying and tracking actionable findings in radiology reports [3, 4]. Handling such information in radiology reports seems a difficult task because radiology reports usually remain unstructured free texts [5]. However, thanks to recently developed natural language processing (NLP) technologies, the detection of radiology reports with actionable findings has been achieved, as well as various other tasks using radiology reports [6]. The aim of this study is to automatically detect reports with actionable findings by NLP-technology-based methods.

Many researchers in previous studies have used NLP technologies to automatically detect specific findings or diseases in radiology reports. Some of them stated that their goal is to assist in tracking and surveillance of actionable findings, the details of which are summarized in Table 1 [7–26]. Some studies in Table 1 have the following features: (1) Multiple or all types of pathological entities are covered [7–15]. (2) The ground truth is based on clinical decisions, not just on the existence of specific expressions in radiology reports [16–18]. These two features can both lead to comprehensive detection of radiology reports with actionable findings. However, there have been no studies that use both features to the best of our knowledge.

In our hospital, for better communication and tracking of any actionable findings, an *actionable* tagging function was implemented in our radiological reporting system and this function has been in operation since September 9, 2019. Thus, adopting *actionable* tags for labeling can provide a dataset based on clinical decisions for all types of pathological entities.

In addition to the free texts in radiology reports, the free texts that are input in the ordering form by the referring clinician (hereafter, order information) may also be useful for detecting radiology reports with actionable findings. That is, if serious and incidental findings are present, some gaps can be found between the order information and the radiology report.

Several research groups have investigated the automatic detection of actionable findings based on statistical machine learning [9–11, 16, 18, 22, 25, 26]. However, these methods are mainly based on the frequency of words in each document, and other rich features such as word order and context are hardly taken into account. Recently, bidirectional encoder representations from transformers (BERT), one of the Transformer networks [27, 28], has attracted much attention because it achieves state-of-the-art performance in various NLP tasks. For better detection of radiology reports with actionable findings, BERT is worth using for two reasons: (1) BERT can use linguistic knowledge not only from an in-house dataset but also from a corpus (a set of documents) for pre-training [29]. (2) BERT is able to capture the relationship between two documents [28], which may enable it to perform well for a pair comprising order information and a radiology report. BERT has been used in several very recent studies of classification tasks in radiology reports [30, 31]. To the best of our knowledge, however, there have been no attempt to use BERT for the automated detection of radiology reports with actionable findings.

In this study, we investigate the automated detection of radiology reports with actionable findings using BERT.

The contributions of this study are as follows.

- Examination of the performance of BERT for the automated detection of actionable reports
- Investigation of the difference in detection performance upon adding order information to the input data

Methods

Task description

This study was approved by the institutional review board in our hospital, and was conducted in accordance with the Declaration of Helsinki.

We define two collective terms: (1) "report body,[1]" referring to the findings and impression in radiology reports, and (2) "order information," referring to the free texts that are written in the ordering form by the referring clinician (e.g., the suspected diseases or indications), as explained in Introduction. Our task is thus defined as the detection of radiology reports with *actionable* tags using the report body alone, or both the order information and the report body.

Clinical data

We obtained 93,215 confirmed radiology reports for computed tomography (CT) examinations performed at our hospital between September 9, 2019, and April 30, 2021, all of which were written in Japanese. Next, we removed the following radiology reports that were not applicable for this study: (1) eight radiology reports whose findings and impressions were both registered as empty, (2) 254 reports for CT-guided biopsies, and (3) 2030 reports for CT scans for radiation therapy planning.

[1] For simplicity, we regarded impression as part of the report body, although this is different from the definition of the body of the report by the American College of Radiology [32].

Table 1 Summary of previous studies of automatic detection of radiology reports with actionable findings, along with this study

	Target language	Multiple diseases	Use of labels in clinical practice	Criteria for positive class	Target sections in radiology reports	Methods
Meng et al. [7]	English	Yes	No	Expressions suggesting the need to promptly communicate to the referring clinician	Impression	Existing tool
Helibrun et al. [8]	English	Yes	No	Expressions suggesting specific critical findings	Impression	Existing tool
Carrodeguas et al. [9]	English	Yes	No	Follow-up recommendations	Impression	SML, LSTM
Yetisgen-Yildiz et al. [10]	English	Yes	No	Follow-up recommendations	Order information, findings, impression	SML
Yetisgen-Yildiz et al. [11]	English	Yes	No	Follow-up recommendations	Order information, findings, impression	SML
Dutta et al. [12]	English	Yes	No	Follow-up recommendations	Findings, impression, recommendation	Existing tool
Lau et al. [13]	English	Yes	No	Follow-up recommendations	(Not specified)	GRU
Dang et al. [14]	English	Yes	No	Follow-up recommendations	(Not specified)	Decision tree
Imai et al. [15]	Japanese	Yes	No	Expressions suggesting malignancy	Findings	Syntactic analysis
Lou et al. [16]	English	No	Yes	Reports pointing at indeterminate or suspicious upper abdominal mass	(Not specified)	SML
Danforth et al. [17]	English	No	Yes	ICD-9 codes suggesting lung nodules	(Not specified)	Rule base
Garla et al. [18]	English	No	Yes	Expressions suggesting potentially malignant liver lesions	(Not specified)	SML
Farjah et al. [19]	English	No	No	Expressions suggesting lung nodules	(Not specified)	Existing tool
Gershanik et al. [20]	English	No	No	Expressions suggesting lung nodules	Findings, impression	Existing tool
Oliveira et al. [21]	English	No	No	Expressions suggesting incidental lung nodules	Order information, findings	Rule base
Pham et al. [22]	French	No	No	Expressions suggesting incidentalomas	Order information, findings, impression	SML
Mabotuwana et al. [23]	English	No	No	Follow-up recommendations	(Not specified)	Rule base
Morioka et al. [24]	English	No	No	Expressions suggesting abdominal aorta aneurysm	(Not specified)	Existing tool
Xu et al. [25]	English	(Not specified)	No	Follow-up recommendations	Order information, findings, impression	SML
Fu et al. [26]	English	No	No	Expressions suggesting silent brain infarction or white matter disease	(Not specified)	Rule base, SML, CNN
This study	Japanese	Yes	Yes	Reports with an *actionable* tag	Order information, findings, impression	SML, LSTM, BERT

BERT = Bidirectional Encoder Representations from Transformers, CNN = Convolutional Neural Network, GRU = Gated Recurrent Units, LSTM = Long Short-Term Memory, and SML = Statistical Machine Learning

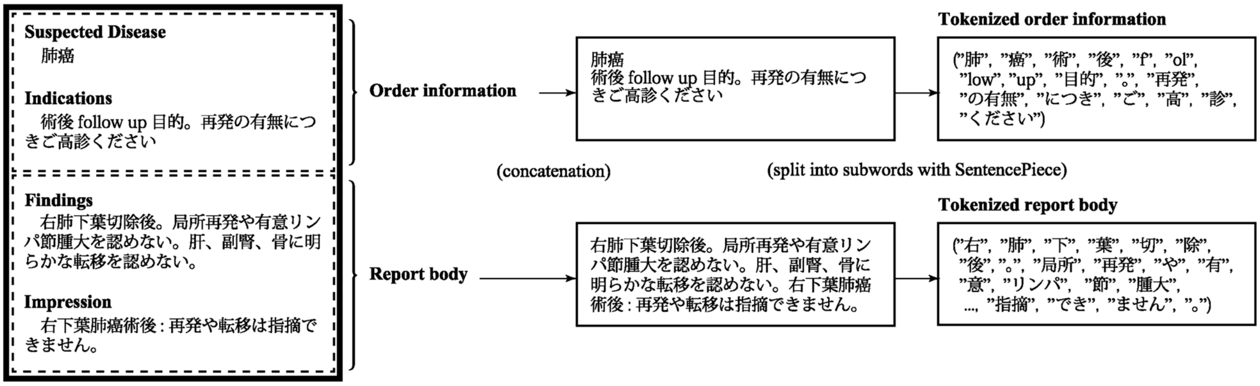

Suspected Disease
肺癌
Indications
術後 follow up 目的。再発の有無につきご高診ください
Findings
右肺下葉切除後。局所再発や有意リンパ節腫大を認めない。肝、副腎、骨に明らかな転移を認めない。
Impression
右下葉肺癌術後 : 再発や転移は指摘できません。
Order information
Report body
(concatenation)
肺癌
術後 follow up 目的。再発の有無につきご高診ください
右肺下葉切除後。局所再発や有意リンパ節腫大を認めない。肝、副腎、骨に明らかな転移を認めない。右下葉肺癌術後 : 再発や転移は指摘できません。
(split into subwords with SentencePiece)
Tokenized order information
(”肺”, ”癌”, ”術”, ”後”, ”f”, ”ol”, ”low”, ”up”, ”目的”, ”。”, ”再発”, ”の有無”, ”につき”, ”ご”, ”高”, ”診”, ”ください”)
Tokenized report body
(”右”, ”肺”, ”下”, ”葉”, ”切”, ”除”, ”後”, ”。”, ”局所”, ”再発”, ”や”, ”有”, ”意”, ”リンパ”, ”節”, ”腫大”, ..., ”指摘”, ”でき”, ”ません”, ”。”)

Fig. 1 Preprocessing for each radiology report. Note that subwords recognized by SentencePiece are not precisely substrings of words in the grammatical sense, because SentencePiece automatically constructs its vocabulary without a dictionary. For this reason, SentencePiece sometimes treats long phrases as one subword or conversely one character as one subword

The remaining 90,923 radiology reports corresponded to 18,388 brain, head, and neck; 64,522 body; 522 cardiac; and 5673 musculoskeletal reports; and 3209 reports of other CT examinations whose body parts could not be determined from the information stored in the Radiology Information System (RIS) server. The total was greater than the number of reports because some reports mentioned more than one part.

Class labeling and data split

Each of the 90,923 radiology reports was defined as actionable (positive class) if it had been provided with an *actionable* tag by the diagnosing radiologist, and it was otherwise defined as non-actionable (negative class). In other words, the gold standard had already been given to all of the reports in the clinical practice, which enabled a fully supervised document classification without additional annotations.

The radiologists in our hospital are requested to regard image findings as actionable when the findings were not supposed to be expected by the referring clinician and were potentially critical if left overlooked. Specific criteria for *actionable* tagging were not determined clearly in advance but left to clinical decisions of individual radiologists.

The numbers of actionable and non-actionable reports were 788 (0.87%) and 90,135 (99.13%), respectively. Then, these radiology reports were split randomly into a training set and a test set in the ratio of 7:3, maintaining the same proportions of actionable and non-actionable reports in each set, i.e., in the training set, there were 63,646 reports, where 552 were actionable and 63,094 were non-actionable, and in the test set, there were 27,277 reports, where 236 were actionable and 27,041 were non-actionable.

Preprocessing of radiology reports

To apply machine learning methods in the following sections, the same preprocessing was carried out on all radiology reports (Fig. 1). First, the contents in the order information and report body were respectively concatenated into passages. Then, the passages were individually tokenized with the SentencePiece model, whose vocabulary size is 32,000 [33, 34].

BERT

BERT is one of the Transformer networks [27, 28]. In general, “Transformer” refers to neural networks using multiple identical encoder or decoder layers with an attention mechanism [35]. Transformer networks have outperformed previous convolutional and recurrent neural networks in NLP tasks [27]. BERT has been proposed as a versatile Transformer network. BERT takes one or two documents as input, passes them into the inner stack of multiple Transformer encoder layers, and characteristically outputs both document-level and token-level representations. BERT can thus be applied to both document-level and token-level classification tasks [28]. Various BERT models pre-trained with large corpora are publicly available, which has established a new ecosystem for pre-training and fine-tuning of NLP models.

We used the Japanese BERT model developed by Kikuta [34]. This model is equivalent to “BERT-base” with 12 Transformer encoder layers and 768-dimensional hidden states. The model has been pre-trained using a Japanese Wikipedia corpus tokenized with the SentencePiece tokenizer [33].

We constructed a binary classifier (hereafter, a BERT classifier) by adding a single-layer perceptron with softmax activation after the pre-trained BERT model. The

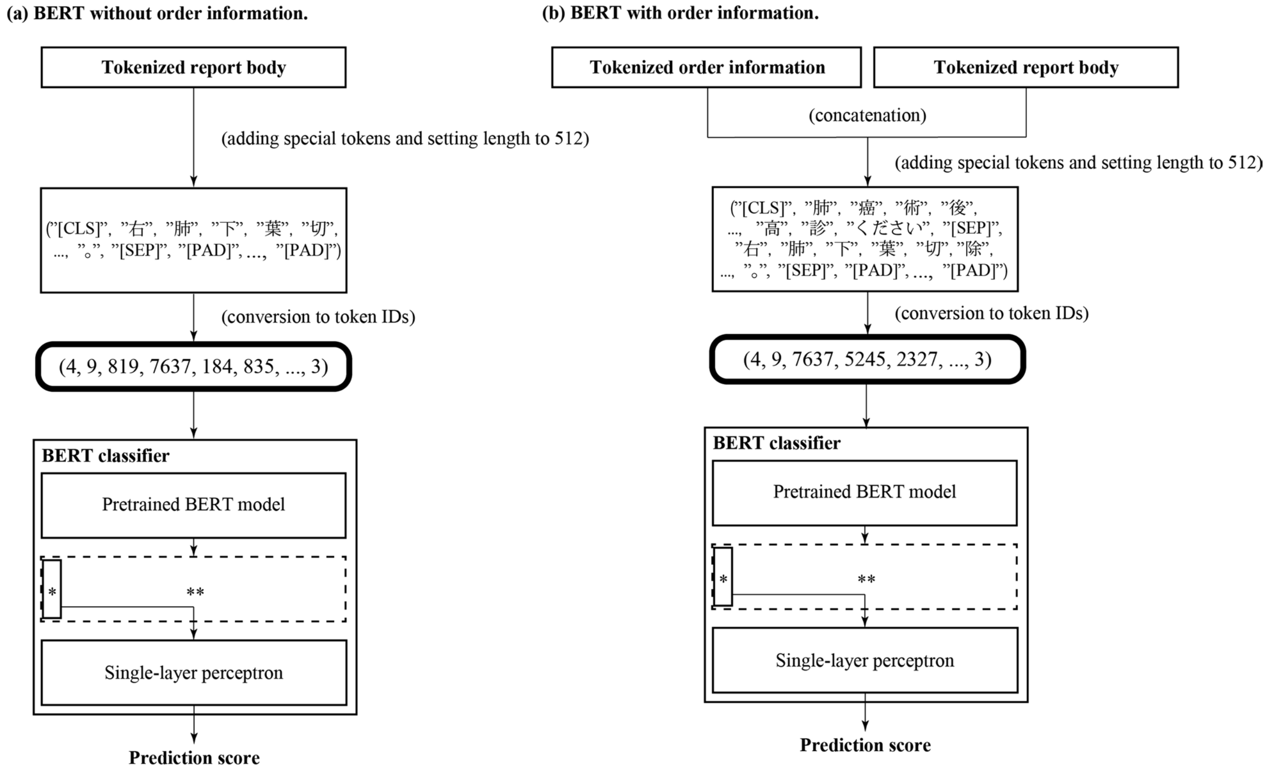

(a) BERT without order information.
Tokenized report body
(adding special tokens and setting length to 512)
("[CLS]", "右", "肺", "下", "葉", "切", ..., "。", "[SEP]", "[PAD]", ..., "[PAD]")
(conversion to token IDs)
(4, 9, 819, 7637, 184, 835, ..., 3)
BERT classifier
Pretrained BERT model
*
**
Single-layer perceptron
Prediction score
(b) BERT with order information.
Tokenized order information
Tokenized report body
(concatenation)
(adding special tokens and setting length to 512)
("[CLS]", "肺", "癌", "術", "後", ..., "高", "診", "ください", "[SEP]", "右", "肺", "下", "葉", "切", "除", ..., "。", "[SEP]", "[PAD]", ..., "[PAD]")
(conversion to token IDs)
(4, 9, 7637, 5245, 2327, ..., 3)
BERT classifier
Pretrained BERT model
*
**
Single-layer perceptron
Prediction score

Fig. 2 Detection of actionable reports with BERT (**a**) without and (**b**) with order information. Each sequence was fed into the BERT classifier after adding special tokens (e.g., [CLS] and [SEP]) and padding with [PAD] tokens that are required by the standard specification of BERT. Generally, BERT models output a 512 × 768 hidden state matrix (indicated by **), part of which is a 768-dimensional feature vector for classification tasks (indicated by *). We used only the feature vector for classification tasks and discarded the rest of the hidden state matrix, as in the standard procedure

perceptron converts a 768-dimensional document-level representation vector output by the pre-trained BERT model into a two-dimensional vector.

The procedure is shown in Fig. 2. For the detection experiment without order information, the sequences generated from the report body were fed to the BERT classifier. For the detection experiment with order information, each sequence pair generated from the order information and report body was fed to the BERT classifier.

Fine-tuning was performed on all embedding and Transformer encoder layers of the BERT model, and none of these layers were frozen. The maximum sequence length was set to 512 and the batch size[2] was set to 256. We used Adam optimizer [36] and binary cross-entropy loss function.

As in Table 2, the learning rate and the number of training epochs were set as follows. The learning rate was set to 5.0×10^{-5} for the experiment without order information and to 4.0×10^{-5} for the experiment with order information. The number of training epochs was set to 3 for both experiments. The learning rate and the number of training epochs were determined by the grid search and five-fold cross-validation using the training set. We tried all of the 25 direct groups of five learning rates, 1.0×10^{-5}, 2.0×10^{-5}, 3.0×10^{-5}, 4.0×10^{-5}, and 5.0×10^{-5}, and the five training epochs, 1 to 5. We calculated the averages of the area under the precision-recall curve (AUPRC) [37, 38] for the five folds, and chose the learning rate and the number of training epochs that gave the highest average AUPRC.

The learning environment was as follows: AMD EPYC 7742 64-Core Processor, 2.0 TB memory, Ubuntu 20.04.2 LTS, NVIDIA A100-SXM4 graphics processing unit (GPU) with 40 GB memory × 6, Python 3.8.10, PyTorch 1.8.1, Torchtext 0.6.0, AllenNLP 2.5.0, PyTorch-Lightning 0.7.6, scikit-learn 0.22.2.post1, Transformers 4.6.1, Tokenizers 0.10.3, SentencePiece 0.1.95, MLflow 1.17.0, and Hydra 0.11.3.

[2] The actual batch size was set to 16 owing to the limited computational resources. However, an effective batch size of 256 was realized by accumulating gradients of every 16 batches with the PyTorch-Lightning implementation.

Table 2 Details of hyperparameter tuning for each method. $xe+y$ means $x \times 10^y$ and $xe-y$ means $x \times 10^{-y}$

Method	Hyper-parameter	Candidates	Used hyperparameters			
			Order information (−)		Order information (+)	
			Oversampling (−)	Oversampling (+)	Oversampling (−)	Oversampling (+)
LR	C	1e−4, 1e−3, 1e−2, 1e−1, 1.0, 1e+1, 1e+2, 1e+3, 1e+4	1e+2	1e−2	1.0	1.0
	L1 ratio	0.0, 0.5, 1.0	1.0	0.5	1.0	1.0
GBDT	Iterations	500, 1,000, 1,500	500	1,000	1,500	1,000
LSTM	Learning rate	5e−6, 1e−5, 2e−5, 3e−5	5e−6	1e−5	5e−6	5e−6
	Epochs	5, 10, 15, 20, 25, 30	20	5	20	25
BERT	Learning rate	1e−5, 2e−5, 3e−5, 4e−5, 5e−5	5e−5	5e−5	4e−5	1e−5
	Epochs	1, 2, 3, 4, 5	3	1	3	1

Baselines: LSTM

As one of the baselines against BERT, we performed automated detections of actionable reports using a two-layer bidirectional long short-term memory (LSTM) model followed by a self-attention layer [27, 39]. As in BERT, the inputs to the LSTM model were report bodies in the experiments without order information and were concatenations of order information and report bodies in the experiments with order information. The lengths of the input documents in a batch were aligned to the longest one by adding special padding tokens at the end of the other documents in the same batch. Next, each document was tokenized and converted into sequences of vocabulary IDs using the SentencePiece tokenizer, and was then passed into a 768-dimensional embedding layer. In short, the preprocessing converted radiology reports in a batch into a batch size × length × 768 tensor.

The final layer of the LSTM model outputs two batch size × length × 768 tensors corresponding to the forward and backward hidden states. We obtained document-level representations by concatenating the two hidden states. The representations were further passed into a single-head self-attention layer with the same architecture as proposed by Vaswani et al. [27]. The self-attention layer converts the document-level representations to a batch size × 1536 matrix by taking the weighted sum of the document-level representations along the time dimension effectively by considering the importance of each token. Then, the matrix was converted into two-dimensional vectors using a single-layer perceptron with softmax activation. The resulting two-dimensional vectors were used as prediction scores. Hereafter, we collectively refer to the LSTM model, the self-attention layer, and the perceptron as the "LSTM classifier."

We trained the LSTM classifier from scratch. The same optimizer and loss function as those in BERT were used. The batch size was set to 256. As in BERT, the learning rate and the number of training epochs were determined by grid search and five-fold cross-validation. Table 2 shows the hyperparameter candidates on which the grid search was performed and the hyperparameters that were finally chosen for each experiment.

Baselines: statistical machine learning

Logistic regression (LR) [40] and the gradient boosting decision tree (GBDT) [41] were also examined for comparison.

Figure 3 shows the procedures. The tokenized report body and order information were individually converted into term frequency-inverse document frequency (TF-IDF)-transformed count vectors of uni-, bi-, and trigrams (one, two, and three consecutive subwords). The two vectors were concatenated for the detection experiment with order information, and only the vector from the report body was used for the detection experiment without order information.

Here, we describe the details of hyperparameters of the LR and GBDT models. For LR, we used Elastic-Net regularization [30, 42], which regulates model weights with the mixture of L1- and L2-norm regularizations. Elastic-Net takes two parameters, C and the L1 ratio. C is the reciprocal strength to regularize the model weights, and the L1 ratio is the degree of dominance of L1-norm regularization. The C and the L1 ratio were determined with the grid search and five-fold cross-validation, whose candidates and choices are shown in Table 2. For GBDT, the tree depth was set to 6. The number of iterations was determined by grid search and five-fold cross-validation in the same way as LR.

We used the scikit-learn 0.22.2post1 implementation for LR and the CatBoost 0.25.1 [43] implementation for GBDT.

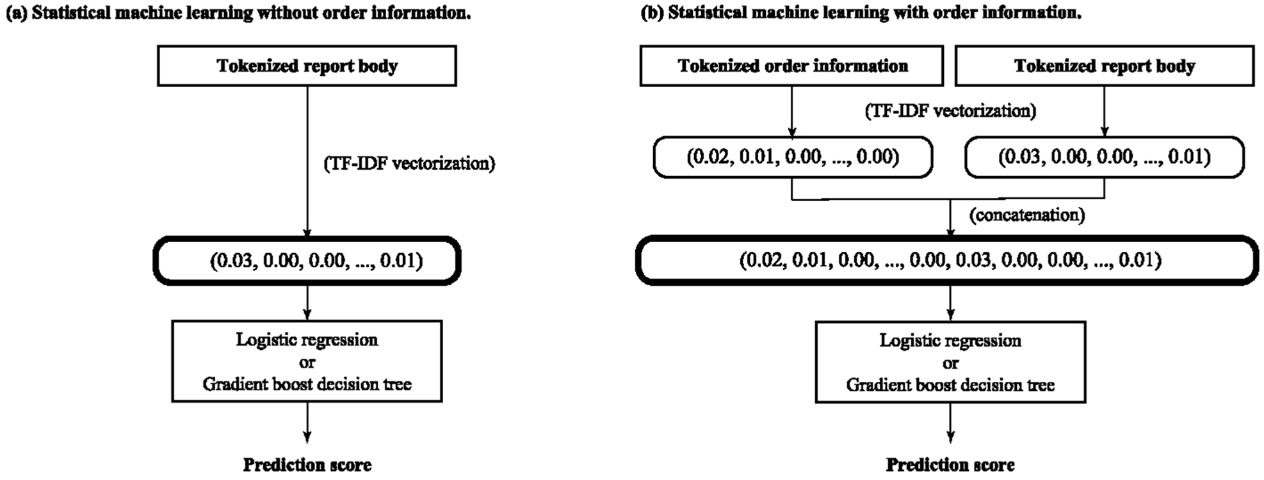


Fig. 3 Detection of actionable reports with statistical machine learning (**a**) without and (**b**) with order information

Performance evaluation

Since this experiment is under a highly imbalanced setting, the performance of each method was mainly evaluated with the AUPRC [37, 38], along with the average precision score.

We statistically compared the AUPRC and average precision among LR, GBDT, LSTM, and BERT using Welch's t-test with Bonferroni correction [44]. The bootstrapping approach was applied, where 2000 replicates were made, and 2000 AUPRCs and average precisions were calculated for LR, GBDT, LSTM, and BERT. Using the same approach, we also statistically compared the AUPRC and average precision in the experiments without and with order information for each method.

The area under the receiver operating characteristics (ROC) curve (AUROC) was also calculated [45, 46]. The recall, precision, specificity, and F1 score were also calculated at the optimal cut-off point of the ROC curve. The optimal cut-off point was chosen using the minimum distance between the ROC curve and the upper left corner of the plot.

Scikit-learn 0.22.2.post1 implementation was used for calculation of the evaluation metrics, bootstrapping, and statistical analysis.

For a more detailed analysis, we divided the truly actionable reports in the test set into explicit actionable reports (those with expressions recommending follow-up imaging, further clinical investigations, or treatments) and implicit ones (those without such expressions) by manual review by one radiologist (Y. Nakamura, four years of experience in diagnostic radiology). We also calculated recalls for the mass and non-mass subsets of the truly actionable reports in the test set since some previous studies have focused on actionable reports that point out incidental masses or nodules [15–22]. Each of the reports was included in the mass subset when its actionable findings were determined to involve masses or nodules by manual review, otherwise reports were included in the non-mass subset.

Oversampling

We mainly used the training set mentioned in the previous section, but its significant class imbalance may affect the performance of the automated detection of actionable reports. Oversampling positive data can be one of the methods to minimize the negative impact of the class imbalance [47].

To examine the effectiveness of oversampling, we additionally performed experiments using the oversampled training set. The oversampled training set was created by resampling each actionable radiology report ten times and each non-actionable radiology report once from the original training set. Hyperparameters for each method (LR, GBDT, LSTM, and BERT) and for each input policy (using and not using order information) were determined using the same strategy as that in the experiments without oversampling. The chosen hyperparameters are shown in Table 2.

Note that we did not oversample the validation datasets during the five-fold cross-validation because we intended to search optimal hyperparameters for the same positive class ratio as the test set.

To examine the effect of oversampling, we statistically compared the AUPRC and average precision obtained without and with oversampling in the same way as aforementioned.

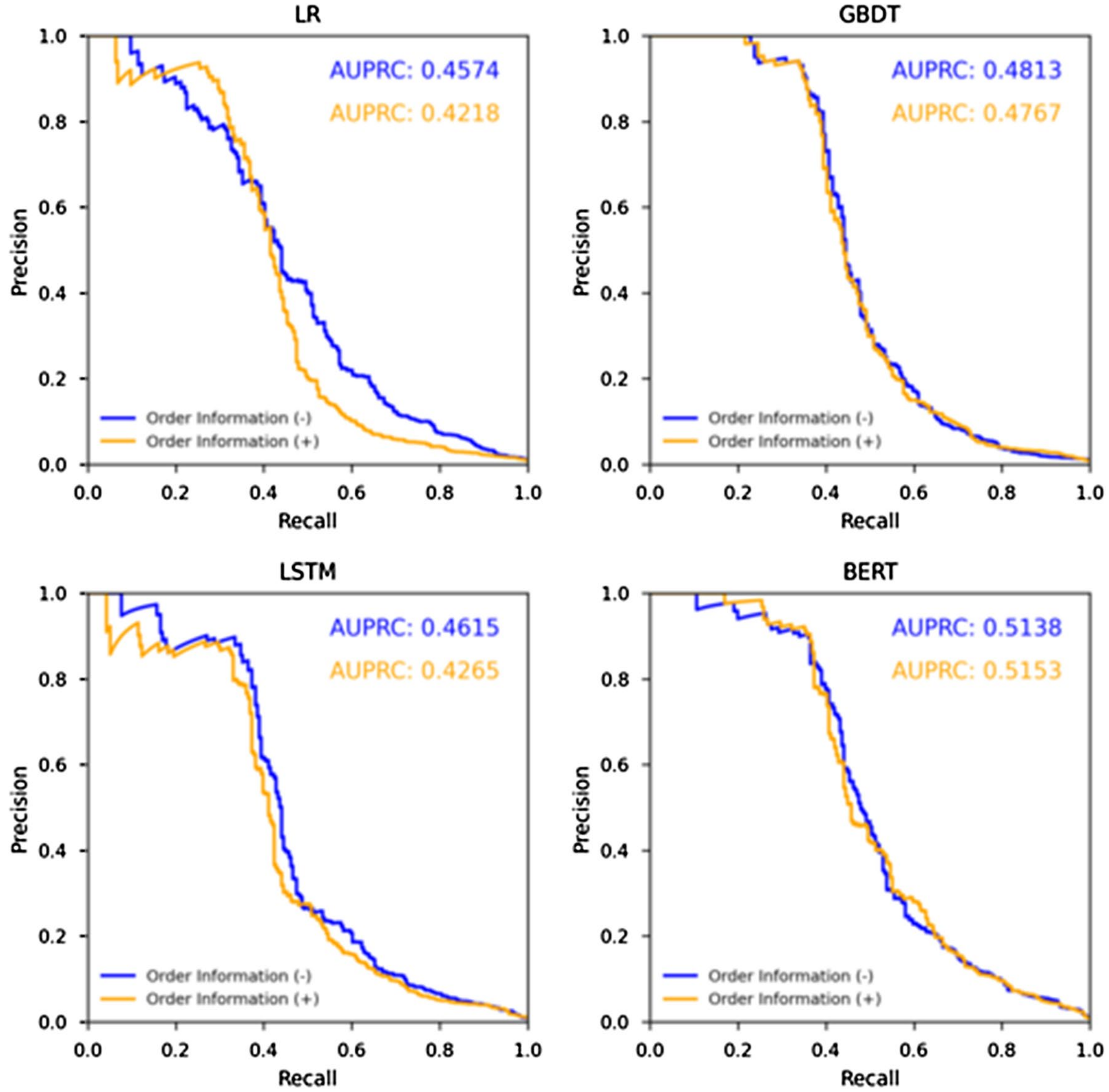


Fig. 4 Precision-recall curves for detection of actionable reports achieved by each method

Results

Figures 4 and 5 show the precision-recall curves and the ROC curves of each method. Table 3 presents the performance of each method calculated from precision-recall curves and optimal cut-off points of ROC curve. Table 4 shows the results of statistical analysis to compare the performance characteristics of LR, GBDT, LSTM, and BERT. In both of the experiments without and with order information, BERT achieved the highest AUPRC and average precision among the four methods, and it showed a statistically significant improvement over the other methods. In particular, the highest AUPRC of 0.5153 was achieved using BERT with order information. The F1 score tended to be higher for the methods with higher AUPRCs, average precisions, and AUROCs. The highest precision was 0.0634, considerably lower than that for recall.

The advantage of using order information was unclear. Tables 3 and 5 show that the use of order information markedly decreased AUPRC except for BERT. Only BERT slightly improved AUPRC with the use of order information, but the improvement was not statistically significant.

Oversampling showed a limited positive effect on the performance. As in Tables 6 and 7, oversampling positive samples in the training dataset ten times resulted

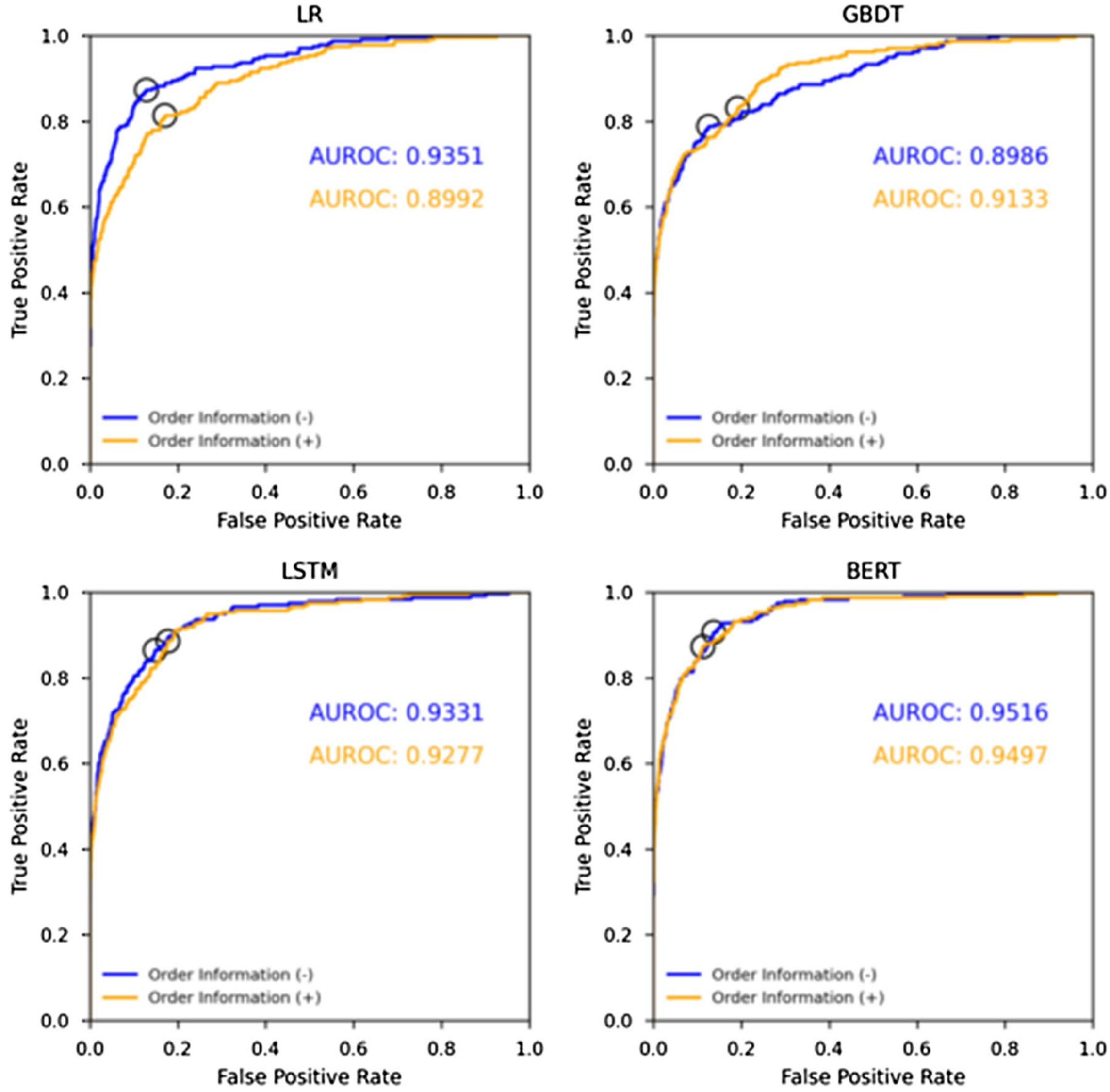


Fig. 5 ROC curves for detection of actionable reports achieved by each method, with optimal cut-off points shown as open circles

Table 3 Performance in detection of actionable reports with different use of order information for each method

Method	Use of order information	AUPRC	Average precision	AUROC	F1 score	Recall	Precision	Specificity
LR	(−)	0.4574	0.4224	0.9351	0.1052	0.8729	0.0559	0.8715
	(+)	0.4218	0.4580	0.8992	0.0763	0.8136	0.0400	0.8296
GBDT	(−)	0.4813	0.4816	0.8986	0.0975	0.7881	0.0520	0.8746
	(+)	0.4767	0.4771	0.9133	0.0699	0.8305	0.0365	0.8087
LSTM	(−)	0.4617	0.4620	0.9331	0.0916	0.8644	0.0484	0.8516
	(+)	0.4265	0.4272	0.9277	0.0797	0.8856	0.0417	0.8226
BERT	(−)	0.5138	0.5142	**0.9516**	0.1030	**0.9068**	0.0546	**0.9271**
	(+)	**0.5153**	**0.5157**	0.9497	**0.1183**	0.8729	**0.0634**	0.9140

Maximum values are shown in bold

Table 4 Results of statistical analysis to examine the performance of each detection method

Use of order information	Metrics	Method	*p* values		
			versus GBDT	versus LSTM	versus BERT
(−)	AUPRC	LR	$p<0.0001$	$p<0.0001$	$p<0.0001$
		GBDT	–	$p<0.0001$	$p<0.0001$
		LSTM	–	–	$p<0.0001$
	Average precision	LR	$p<0.0001$	$p=0.0002$	$p<0.0001$
		GBDT	–	$p<0.0001$	$p<0.0001$
		LSTM	–	–	$p<0.0001$
(+)	AUPRC	LR	$p<0.0001$	$p<0.0001$	$p<0.0001$
		GBDT	–	$p<0.0001$	$p<0.0001$
		LSTM	–	–	$p<0.0001$
	Average precision	LR	$p<0.0001$	$p<0.0001$	$p<0.0001$
		GBDT	–	$p<0.0001$	$p<0.0001$
		LSTM	–	–	$p<0.0001$

Table 5 Results of statistical analysis to examine the impact of use of order information

Method	Metrics	*p* values
		Use of order information (−) versus (+)
LR	AUPRC	$p<0.0001$
	Average precision	$p<0.0001$
GBDT	AUPRC	$p<0.0001$
	Average precision	$p<0.0001$
LSTM	AUPRC	$p<0.0001$
	Average precision	$p<0.0001$
BERT	AUPRC	$p=0.0972$
	Average precision	$p=0.1143$

in statistically significant improvements of AUPRC and average precision only for GBDT.

We analyzed further how predictions were made by each method. For LR and GBDT, each of the available *n*-grams (i.e., uni-, bi-, and trigrams) were scored using coefficients assigned by the LR models or feature importance assigned by the GBDT models, which reflected the *n*-grams that the LR and GBDT models placed importance during prediction. *N*-grams consisting only of either Japanese punctuations or Japanese postpositional particles were excluded because they were assumed to be of little value. The results are shown in Figs. 6 and 7, which suggest that the LR and GBDT models tended to predict radiology reports as actionable if they contained such expressions as "is actionable," "investigation," "cancer," or "possibility of cancer." This suggests that the models picked up explicit remarks by radiologists recommending clinical actions or pointing out cancers. In contrast, patterns in keywords used by the LR model for non-actionable radiology reports were less clear, although some negations such as "is absent" or "not" are observed in Fig. 6b. The word "apparent", which is frequently accompanied by negative findings in Japanese radiology reporting, is also present in the top negative *n*-grams in Fig. 6b. These imply that the LR model might deduce that radiology reports are non-actionable when negative findings predominate. Order information may not be used much by the LR and GBDT models because few of the *n*-grams in order information are present in Figs. 6 and 7.

Figure 8 is a visualization of the self-attention of the LSTM and BERT classifier, highlighting tokens on which large importance was placed by each model during prediction. For LSTM, tokens attracting more attention than others are shown in red. The attention scores were calculated by averaging the row vectors of the attention matrix generated by the self-attention layer. The attention matrix has the length × length size, whose (i, j) element of the attention matrix stands for the degree of the i-th token attending the j-th token. Thus, averaging the row vectors can clarify which token is attracting more attention overall than others. For BERT, tokens directing intensive attention toward the [CLS] special token are shown in red. The attention scores were calculated by averaging all of the attention weight matrices in each of the 12 attention heads in the last Transformer encoder layer of the BERT classifier. In Fig. 8, attention scores tended to be higher in expressions such as recommendations or suspicions than in anatomical, radiological, or pathological terms.

Table 8 shows the recalls of each method for the explicit and implicit actionable reports in the test set. 111

Table 6 Performance characteristics of methods in detection of actionable reports without and with oversampling of positive samples in the training data

Method	Use of order information	Oversampling	AUPRC	Average precision	AUROC	F1 score
LR	(−)	(−)	**0.4574**	0.4224	**0.9351**	0.1052
		(+)	0.3166	0.3167	0.8036	0.0474
	(+)	(−)	0.4218	**0.4580**	0.8992	0.0763
		(+)	0.4214	0.4221	0.9277	**0.1089**
GBDT	(−)	(−)	0.4813	0.4816	0.8986	0.0975
		(+)	0.4854	0.4858	**0.9335**	0.0841
	(+)	(−)	0.4767	0.4771	0.9133	0.0699
		(+)	**0.4874**	**0.4878**	0.9307	**0.0920**
LSTM	(−)	(−)	**0.4617**	**0.4620**	**0.9331**	**0.0916**
		(+)	0.4188	0.4194	0.9262	0.0818
	(+)	(−)	0.4265	0.4272	0.9277	0.0797
		(+)	0.4086	0.4066	0.9255	0.0795
BERT	(−)	(−)	0.5138	0.5142	**0.9516**	0.1030
		(+)	0.4256	0.4273	0.9464	**0.1190**
	(+)	(−)	**0.5153**	**0.5157**	0.9497	0.1183
		(+)	0.4549	0.4559	0.9441	0.0953

Maximum values for each method are shown in bold

Table 7 Results of statistical analysis to examine the impact of oversampling

Method	Use of order information	Metrics	***p* values**
			Oversampling (−) versus (+)
LR	(−)	AUPRC	$p<0.0001$
		Average precision	$p<0.0001$
	(+)	AUPRC	$p=0.7971$
		Average precision	$p=0.9280$
GBDT	(−)	AUPRC	$p=0.0001$
		Average precision	$p<0.0001$
	(+)	AUPRC	$p<0.0001$
		Average precision	$p<0.0001$
LSTM	(−)	AUPRC	$p<0.0001$
		Average precision	$p<0.0001$
	(+)	AUPRC	$p<0.0001$
		Average precision	$p<0.0001$
BERT	(−)	AUPRC	$p<0.0001$
		Average Precision	$p<0.0001$
	(+)	AUPRC	$p<0.0001$
		Average Precision	$p<0.0001$

truly actionable reports (47%) were implicit in the test set. Although Figs. 6, 7 and 8 imply that all four methods tended to detect actionable findings mainly on the basis of the existence of specific expressions, Table 8 shows that our methods were able to identify actionable reports even if they did not explicitly recommend further medical procedures.

Five of the implicit actionable reports were detected only by BERT and not detected by other methods without order information. Figure 9 shows the BERT attention visualizations towards three of the reports, all of which point out pneumothorax. Although none of the three reports include explicit recommendations or emphatic expressions to highlight actionable findings, BERT successfully predicted them as actionable. Moreover, Figure 9 shows that BERT has assigned high attention scores to a part of the involved disease name "pneumothorax."

In short, although Figs. 6, 7 and 8 suggest that all four methods mainly relied on whether radiology reports contain specific expressions of recommendation, suspicion, or negation, Fig. 9 implies further the capability of BERT to consider characteristics of diseases.

Table 9 shows the recall for truly actionable reports in the test set. The results in Table 9 suggest that our methods detected actionable reports regardless of the pathological entity of their actionable findings.

As in Table 10, actionable reports accounted for 0.41% of brain, head, and neck; 1.1% of body; and 0.51% of musculoskeletal CT radiology reports in the test set. Table 10 also shows that the recall scores for the actionable musculoskeletal CT reports were greater than those for brain, head, and neck CT reports.

(See figure on next page.)

Fig. 6 Top *n*-grams with positive and negative coefficients with the largest absolute values of the LR models (**a**) without and (**b**) with order information. Only the top 25 *n*-grams are shown when more than 25 *n*-grams had non-zero coefficients. *N*-grams in order information are marked with [Order]. The translation is not given for *n*-grams too short to make sense. Negation appears among *n*-grams with the smallest negative coefficient

Discussion

The results show that our method based on BERT outperformed other deep learning methods and statistical machine learning methods in distinguishing various actionable radiology reports from non-actionable ones. The statistical machine learning methods used only limited features, because the radiology reports were converted into the vectors of the frequency of words as the standard feature extraction method [40]. In contrast, BERT and LSTM presumably captured various features of each radiology report including the word order, lexical and syntactic information, and context [28, 29]. Moreover, the superiority of BERT over LSTM was probably brought about by leveraging knowledge from a large amount of pre-training data.

As in Tables 8 and 9, our BERT-based approach was effective in identifying actionable reports regardless of the explicitness or the targeted abnormality. The probable reasons were that (1) implicit actionable reports often emphasized the abnormality that was considered actionable (e.g., “highly suspected to be primary lung cancer” for lung nodules) and that (2) the BERT classifiers were alert to such emphatic expressions in addition to explicit recommendations for follow-up, investigations, or treatment. Furthermore, Figure 9 shows that BERT could still identify implicit actionable reports without emphatic expressions for the actionable findings, and it could assign high attention scores to the names of the actionable findings. This implies that BERT is capable of learning to distinguish disease names that are likely to be often reported as actionable findings.

As in Table 10, the detection performance was affected by the body part of the radiology reports. This is probably caused by the difference in the proportion of explicit and mass actionable reports for each body part. The actionable musculoskeletal CT reports were more often explicit and targeting mass abnormality than the brain, head, and neck CT reports. Tables 8 and 9 suggest that explicit and mass actionable reports were comparatively easier to identify than implicit and non-mass ones. This was probably why all four methods achieved higher recalls scores for musculoskeletal actionable reports than brain, head, and neck ones.

Order information did not necessarily improve the performance. This may be because the truly actionable reports had a too diverse relationship between the order information and the report body. We found that the *actionable* tags were not only used to caution about findings that were irrelevant to the main purpose of ordering (e.g., lung nodules found in a CT examination to diagnose fracture). Rather, the *actionable* tags were also given to the radiology reports to highlight unusual clinical courses (e.g., liver metastases from colon cancer first appeared five years after the surgery of the primary lesion) or to prompt immediate treatments (e.g., hemorrhage in the nasal septum associated with nasal fracture). These complex situations may have not been recognized well from our small dataset, even with the ability of BERT to capture the relationship between the report body and order information.

The low precision (0.0365–0.0634) was another problem in this study. It was probably mainly due to the low positive case ratio (0.87%). Generally, an imbalance of occurrences between positive and negative samples strongly hampers a binary classification task [48]. This negative impact of low positive case ratio was not alleviated by simple oversampling, probably because it did not provide bring new information to learn characteristics of actionable reports to the models. To overcome this limitation, obtaining a larger amount of positive data by collecting more radiology reports or data augmentation [49] may be an effective solution. Other approaches such as cost-sensitive learning [50] or the use of dice loss function [51] can also be worth trying in future studies.

An important advantage of the proposed approach in this study is that the radiology reports were labeled with tags provided in actual radiological practice. Generally, radiologists determine whether specific findings are actionable or not on the basis of not only radiological imaging but also a comparison with a prior series of images, order information, and electronic health records. The actionable tag can consequently reflect such clinical decisions. Therefore, there is probably room for improvement in the performance of automated detection of actionable reports by using the imaging data themselves and the information in electronic health records. This benefit may not be obtained by independent class labeling, referring only to the sentences in the radiology reports.

Using the *actionable* tag as the label has another merit: to identify implicit actionable reports. The results of this study suggest that the radiologists may have sometimes thought that actionable findings were

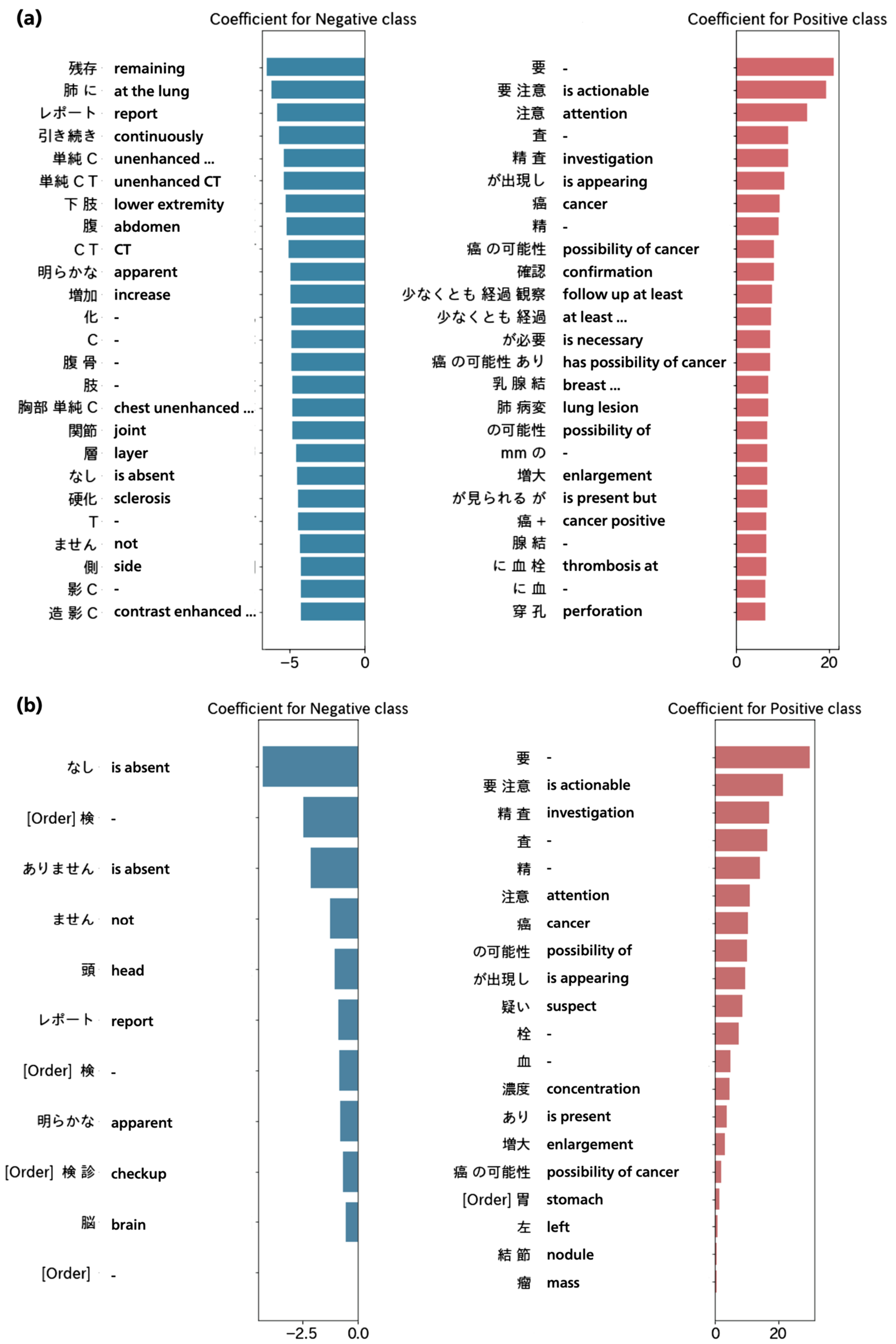


Fig. 6 (See legend on previous page.)

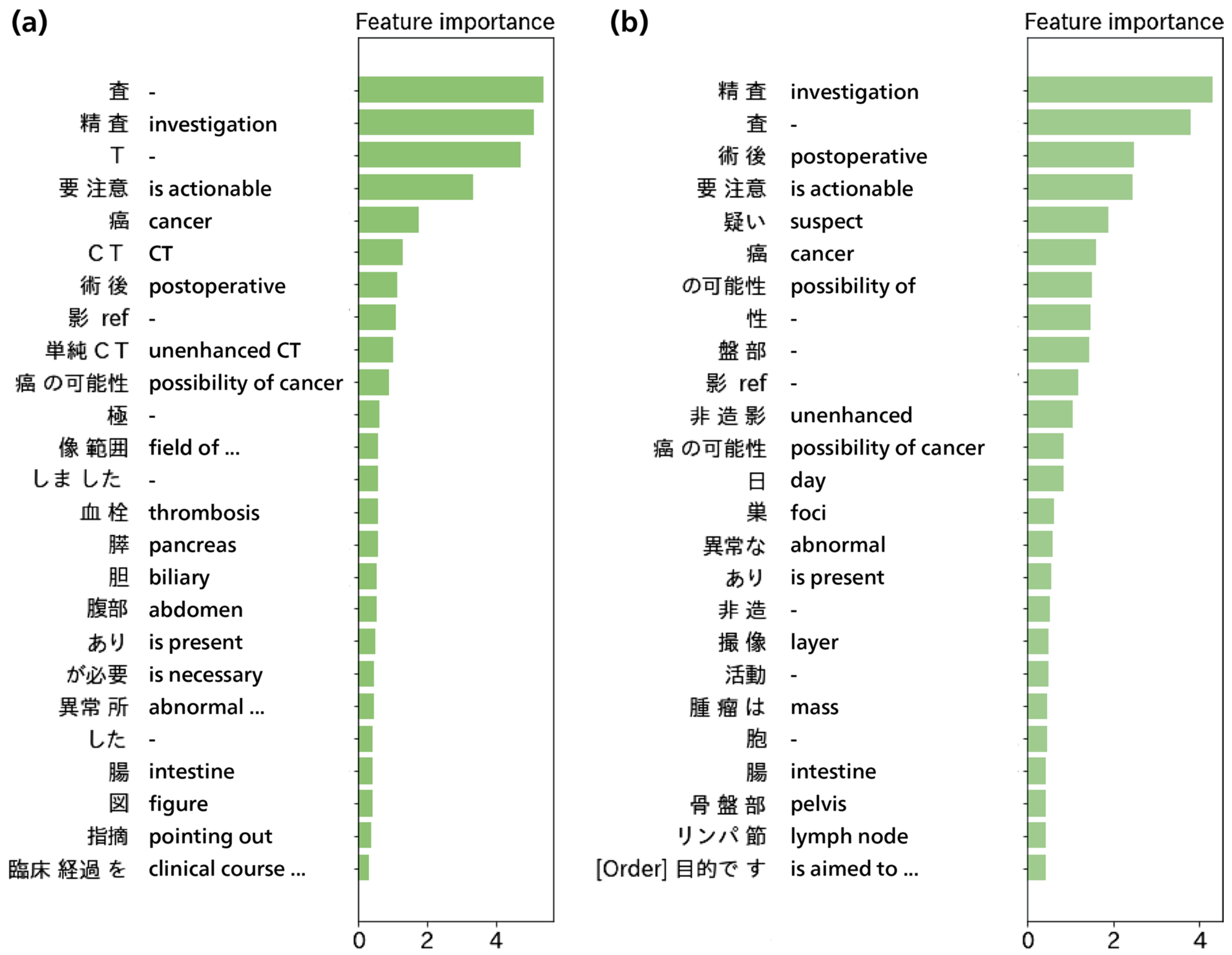


Fig. 7 Top 25 *n*-grams with the largest feature importance of GBDT (**a**) without and (**b**) with order information. *N*-grams in order information are marked with [Order]. The translation is not given for *n*-grams too short to make sense

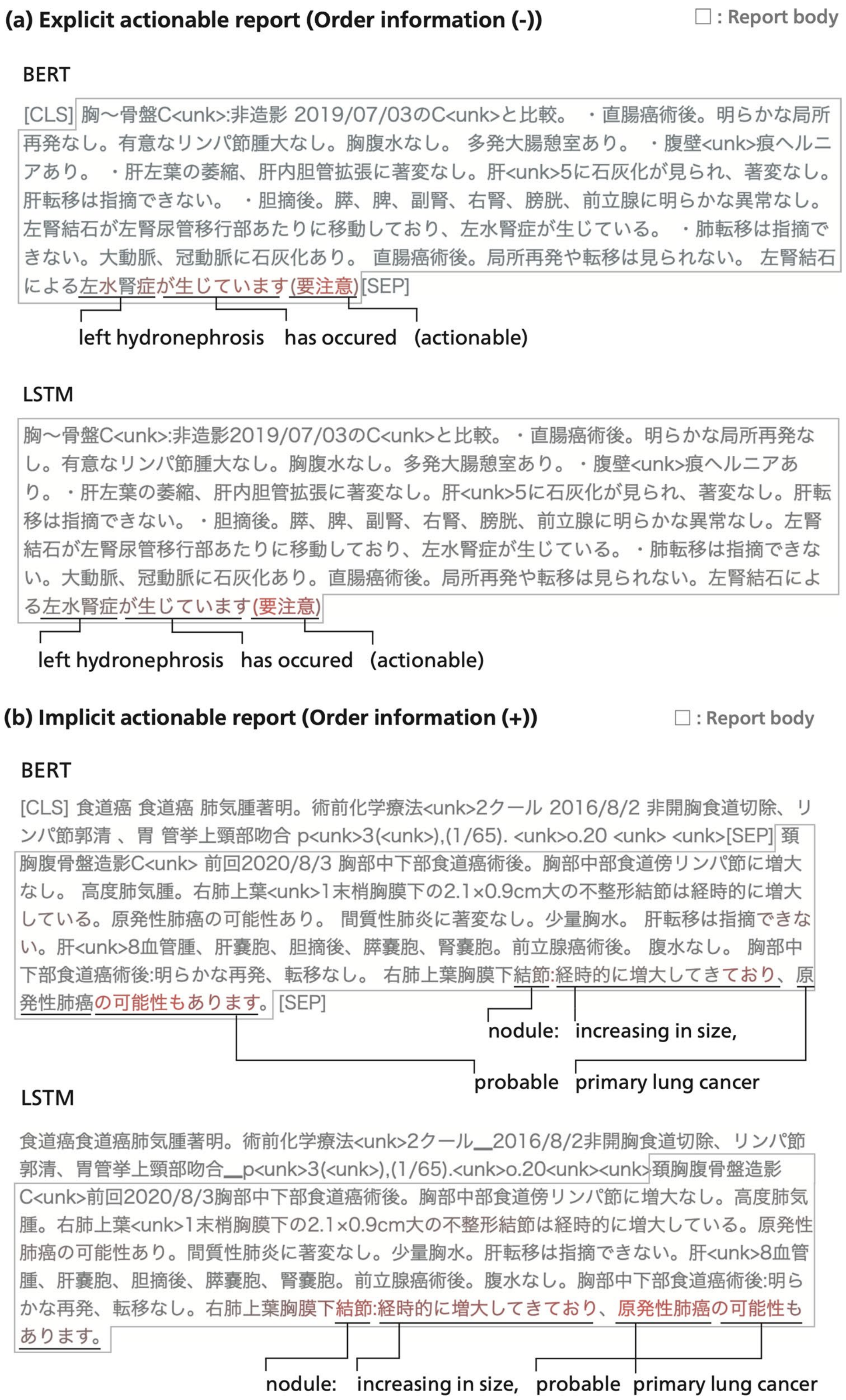
present in the radiological images without explicitly urging further clinical examinations or treatments in the radiology report. The labeling and detection methods in this study identified such implicit actionable reports, though with lower performance than those for explicit ones.

Another advantage of the approach of this study is that actionable findings for any pathological entity were dealt with, thereby realizing comprehensive detection. Since various diseases appear as actionable findings in radiological imaging [1, 7–15], this wide coverage is considered essential for better clinical practice.

The *actionable* tagging itself can play a certain role in the clinical management of actionable reports. Nonetheless, introducing an automated detection system for actionable findings can make further contributions by providing decisions complementary to those of the radiologists. This is because different radiologists have been shown to act differently to actionable findings [52], and there have been no specific criteria for actionable tagging in our hospital thus far.

(See figure on next page.)

Fig. 8 Examples of LSTM and BERT predictions for two truly actionable reports with visualization of attention scores. (**a**) is an explicit actionable report detected without order information, and (**b**) is an implicit actionable report detected using order information. (**a**) Points out hydronephrosis due to ureteral calculus in the postoperative CT examination of rectal cancer, and (**b**) points out a lung nodule pointed out in the CT examination more than four years after the operation of esophageal carcinoma. "< unk >" stands for out-of-vocabulary subwords that were not recognized by the LSTM and BERT classifiers. Subwords with relatively high attention scores are colored red. For luminous visualization, Japanese periods are not colored



(a) Explicit actionable report (Order information (-))

BERT

[CLS] 胸～骨盤C<unk>:非造影 2019/07/03のC<unk>と比較。 ・直腸癌術後。明らかな局所再発なし。有意なリンパ節腫大なし。胸腹水なし。 多発大腸憩室あり。 ・腹壁<unk>痕ヘルニアあり。 ・肝左葉の萎縮、肝内胆管拡張に著変なし。肝<unk>5に石灰化が見られ、著変なし。肝転移は指摘できない。 ・胆摘後。膵、脾、副腎、右腎、膀胱、前立腺に明らかな異常なし。左腎結石が左腎尿管移行部あたりに移動しており、左水腎症が生じている。 ・肺転移は指摘できない。大動脈、冠動脈に石灰化あり。 直腸癌術後。局所再発や転移は見られない。 左腎結石による左水腎症が生じています(要注意)[SEP]

LSTM

胸～骨盤C<unk>:非造影2019/07/03のC<unk>と比較。・直腸癌術後。明らかな局所再発なし。有意なリンパ節腫大なし。胸腹水なし。多発大腸憩室あり。・腹壁<unk>痕ヘルニアあり。・肝左葉の萎縮、肝内胆管拡張に著変なし。肝<unk>5に石灰化が見られ、著変なし。肝転移は指摘できない。・胆摘後。膵、脾、副腎、右腎、膀胱、前立腺に明らかな異常なし。左腎結石が左腎尿管移行部あたりに移動しており、左水腎症が生じている。・肺転移は指摘できない。大動脈、冠動脈に石灰化あり。直腸癌術後。局所再発や転移は見られない。左腎結石による左水腎症が生じています(要注意)

(b) Implicit actionable report (Order information (+))

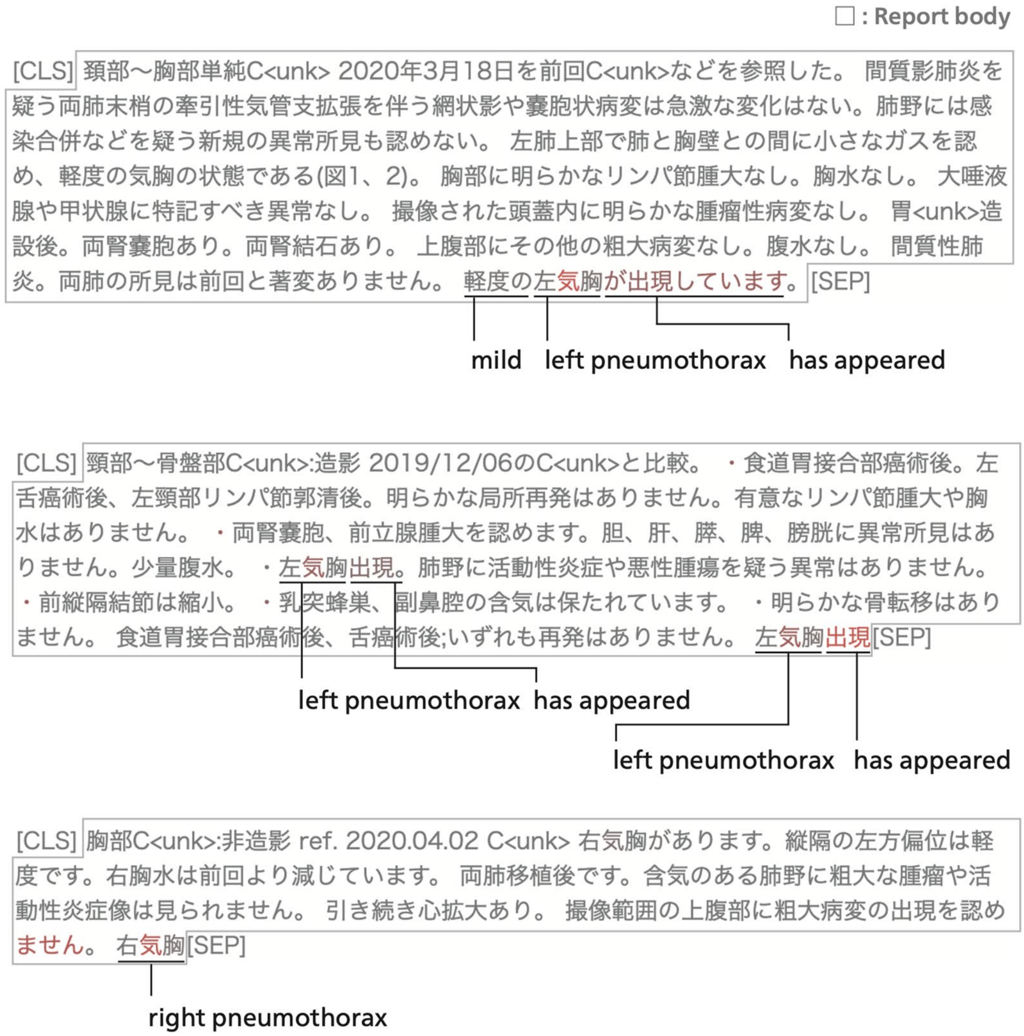
BERT

[CLS] 食道癌 食道癌 肺気腫著明。術前化学療法<unk>2クール 2016/8/2 非開胸食道切除、リンパ節郭清 、胃 管挙上頸部吻合 p<unk>3(<unk>),(1/65). <unk>o.20 <unk> <unk>[SEP] 頚胸腹骨盤造影C<unk> 前回2020/8/3 胸部中下部食道癌術後。胸部中部食道傍リンパ節に増大なし。 高度肺気腫。右肺上葉<unk>1末梢胸膜下の2.1×0.9cm大の不整形結節は経時的に増大している。原発性肺癌の可能性あり。 間質性肺炎に著変なし。少量胸水。 肝転移は指摘できない。肝<unk>8血管腫、肝嚢胞、胆摘後、膵嚢胞、腎嚢胞。前立腺癌術後。 腹水なし。 胸部中下部食道癌術後:明らかな再発、転移なし。 右肺上葉胸膜下結節:経時的に増大してきており、原発性肺癌の可能性もあります。 [SEP]

LSTM

食道癌食道癌肺気腫著明。術前化学療法<unk>2クール__2016/8/2非開胸食道切除、リンパ節郭清、胃管挙上頸部吻合__p<unk>3(<unk>),(1/65).<unk>o.20<unk><unk>頚胸腹骨盤造影C<unk>前回2020/8/3胸部中下部食道癌術後。胸部中部食道傍リンパ節に増大なし。高度肺気腫。右肺上葉<unk>1末梢胸膜下の2.1×0.9cm大の不整形結節は経時的に増大している。原発性肺癌の可能性あり。間質性肺炎に著変なし。少量胸水。肝転移は指摘できない。肝<unk>8血管腫、肝嚢胞、胆摘後、膵嚢胞、腎嚢胞。前立腺癌術後。腹水なし。胸部中下部食道癌術後:明らかな再発、転移なし。右肺上葉胸膜下結節:経時的に増大してきており、原発性肺癌の可能性もあります。

Fig. 8 (See legend on previous page.)


□ : Report body
mild
left pneumothorax
has appeared
left pneumothorax
has appeared
left pneumothorax
has appeared
right pneumothorax

[CLS] 頸部～胸部単純C<unk> 2020年3月18日を前回C<unk>などを参照した。 間質影肺炎を疑う両肺末梢の牽引性気管支拡張を伴う網状影や嚢胞状病変は急激な変化はない。肺野には感染合併などを疑う新規の異常所見も認めない。 左肺上部で肺と胸壁との間に小さなガスを認め、軽度の気胸の状態である(図1、2)。 胸部に明らかなリンパ節腫大なし。胸水なし。 大唾液腺や甲状腺に特記すべき異常なし。 撮像された頭蓋内に明らかな腫瘤性病変なし。 胃<unk>造設後。両腎嚢胞あり。両腎結石あり。 上腹部にその他の粗大病変なし。腹水なし。 間質性肺炎。両肺の所見は前回と著変ありません。 軽度の左気胸が出現しています。[SEP]

[CLS] 頸部～骨盤部C<unk>:造影 2019/12/06のC<unk>と比較。 ・食道胃接合部癌術後。左舌癌術後、左頸部リンパ節郭清後。明らかな局所再発はありません。有意なリンパ節腫大や胸水はありません。 ・両腎嚢胞、前立腺腫大を認めます。胆、肝、膵、脾、膀胱に異常所見はありません。少量腹水。 ・左気胸出現。肺野に活動性炎症や悪性腫瘍を疑う異常はありません。 ・前縦隔結節は縮小。 ・乳突蜂巣、副鼻腔の含気は保たれています。 ・明らかな骨転移はありません。 食道胃接合部癌術後、舌癌術後;いずれも再発はありません。 左気胸出現[SEP]

[CLS] 胸部C<unk>:非造影 ref. 2020.04.02 C<unk> 右気胸があります。縦隔の左方偏位は軽度です。右胸水は前回より減じています。 両肺移植後です。含気のある肺野に粗大な腫瘤や活動性炎症像は見られません。 引き続き心拡大あり。 撮像範囲の上腹部に粗大病変の出現を認めません。 右気胸[SEP]

Fig. 9 Three implicit actionable radiology reports pointing out incidental pneumothorax, all of which were successfully identified only by BERT. BERT attention scores are visualized in the same way as in Fig. 8. Although none of the three radiology reports emphasize urgency or explicitly recommend clinical actions, BERT has given high attention scores to the disease name "pneumothorax."

Table 8 Recall scores for explicit and implicit truly actionable reports in the test set

Method	LR		GBDT		LSTM		BERT	
Use of order information	(−)	(+)	(−)	(+)	(−)	(+)	(−)	(+)
Explicit actionable reports (n = 125)	0.960	0.848	0.872	0.912	0.920	0.936	**0.968**	0.928
Implicit actionable reports (n = 111)	0.766	0.766	0.685	0.730	0.793	0.820	**0.829**	0.802

The maximum score for each subset is shown in bold

There are several limitations of the approach of this study. First, the BERT model used in this study was not specialized in the biomedical domain. The BERT model failed to recognize about 1% of the words, most of which were abbreviations or uncommon Chinese characters of medical terms. Kawazoe et al. have recently provided a BERT model pre-trained with Japanese clinical records, which may improve the performance [53]. The pre-training of BERT with a large Japanese biomedical corpus is worthwhile as future work, although it can be costly from the viewpoint of computational resources. Second, the short period since the launch of *actionable* tagging in our

Table 9 Recall scores for truly actionable reports pointing out mass and non-mass abnormalities in the test set

Method	LR		GBDT		LSTM		BERT	
Use of order information	(−)	(+)	(−)	(+)	(−)	(+)	(−)	(+)
Mass subset (n = 124)	**0.935**	0.895	0.839	0.855	0.895	0.927	0.927	0.903
Non-mass subset (n = 112) †	0.795	0.714	0.723	0.795	0.821	0.830	**0.875**	0.830

The maximum value for each subset is shown in bold

† Vascular lesions (hemorrhage, thrombosis, infarction, and others) (n = 46), pneumonia (n = 17), pneumothorax (n = 11), hydronephrosis (n = 8), gastrointestinal perforation (n = 4), mediastinal emphysema (n = 3), hydrocephalus (n = 2), and other abnormalities (n = 21)

Table 10 Recall for truly actionable reports in the test set calculated for each body part

Body part	#Actionable reports			Recall							
	Total	Implicit	Non-mass	Order information (−)				Order information (+)			
				LR	GBDT	LSTM	BERT	LR	GBDT	LSTM	BERT
Brain, head and neck	23/5584 (0.41%)	10/23 (43.5%)	16/23 (69.6%)	0.739	0.609	**0.870**	**0.870**	0.652	0.783	0.826	0.696
Body	206/19,256 (1.1%)	101/206 (49.0%)	91/206 (44.2%)	0.879	0.801	0.854	**0.903**	0.835	0.825	0.883	0.888
Cardiac	0/151 (0%)	–	–	–	–	–	–	–	–	–	–
Skeletal	9/1758 (0.51%)	1/9 (11.1%)	6/9 (66.7%)	**1.000**	0.889	**1.000**	**1.000**	0.667	**1.000**	**1.000**	0.889
Other	0/959 (0%)	–	–	–	–	–	–	–	–	–	–

Maximum values for each body part are shown in bold

hospital meant that the amount of data was limited. Continuous *actionable* tagging operations can lead to larger datasets. Finally, since this study is a single-institution study, our classifiers may be adapted to the epidemiology, the style of reporting, and the principle on actionable findings unique to our hospital. Expanding this study to other institutions with similar systems of reporting and communication will be valuable future work.

Conclusions

We have investigated the automated detection of radiology reports with actionable findings using BERT. The results showed that our method based on BERT is more useful for distinguishing various actionable radiology reports from non-actionable ones than models based on other deep learning methods or statistical machine learning.

Abbreviations

AUPRC: Area under the precision-recall curve; AUROC: Area under the receiver operating characteristics curve; BERT: Bidirectional encoder representations from transformers; CNN: Convolutional neural network; GRU: Gated recurrent units; CT: Computed tomography; GBDT: Gradient boosting decision tree; LR: Logistic regression; LSTM: Long short-term memory; NLP: Natural language processing; RIS: Radiology information system; ROC: Receiver operating characteristics; SML: Statistical machine learning; TF-IDF: Term frequency-inverse document frequency.

Acknowledgements

The Department of Computational Radiology and Preventive Medicine, The University of Tokyo Hospital, wishes to thank HIMEDIC Inc. and Siemens Healthcare K.K.

Authors' contributions

Y. Nakamura implemented the model and analyzed the data. Y. Nakamura, SH, and Y. Nomura wrote the manuscript. TN, SM, TW, TY, NH, and OA helped revise the manuscript. SH and Y. Nomura contributed in providing the dataset and in improvement of study design and analysis. All authors have read and approved the final manuscript.

Funding

The Department of Computational Radiology and Preventive Medicine, The University of Tokyo Hospital, is sponsored by HIMEDIC Inc. and Siemens Healthcare K.K.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional review board at The University of Tokyo Hospital (No.:2561-(18), approval date: 25 May 2009, last renewal date: 22 January 2020). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1975 Declaration of Helsinki, as revised in 2008(5). The institutional review board above stated that formal consent was not required for this study.

Consent for publication

Not applicable.

Availability of data and materials

The radiology reports and order information involved in this study are not publicly available because publishing the dataset is not approved by the institutional review board of the University of Tokyo Hospital. For more information, please contact the corresponding authors.

Competing interests

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Author details

[1]Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. [2]The Department of Radiology, The University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. [3]The Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan.

Received: 17 February 2021 Accepted: 23 August 2021
Published online: 11 September 2021

References

1. Larson PA, Berland LL, Griffith B, Kahn CE, Liebscher LA. Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group. J Am Coll Radiol. 2014;11(6):552–8.
2. Sloan CE, Chadalavada SC, Cook TS, Langlotz CP, Schnall MD, Zafar HM. Assessment of follow-up completeness and notification preferences for imaging findings of possible cancer: what happens after radiologists submit their reports? Acad Radiol. 2014;21(12):1579–86.
3. Baccei SJ, DiRoberto C, Greene J, Rosen MP. Improving communication of actionable findings in radiology imaging studies and procedures using an EMR-independent system. J Med Syst. 2019;43(2):30.
4. Cook TS, Lalevic D, Sloan C, Chadalavada SC, Langlotz CP, Schnall MD, et al. Implementation of an automated radiology recommendation-tracking engine for abdominal imaging findings of possible cancer. J Am Coll Radiol. 2017;14(5):629–36.
5. Langlotz CP. Structured radiology reporting: are we there yet? Radiology. 2009;253(1):23–5.
6. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology. 2016;279(2):329–43.
7. Meng X, Heinz MV, Ganoe CH, Sieberg RT, Cheung YY, Hassanpour S. Understanding urgency in radiology reporting: identifying associations between clinical findings in radiology reports and their prompt communication to referring physicians. Stud Health Technol Inform. 2019;264:1546–7.
8. Heilbrun ME, Chapman BE, Narasimhan E, Patel N, Mowery D. Feasibility of natural language processing-assisted auditing of critical findings in chest radiology. J Am Coll Radiol. 2019;16(9):1299–304.
9. Carrodeguas E, Lacson R, Swanson W, Khorasani R. Use of machine learning to identify follow-up recommendations in radiology reports. J Am Coll Radiol. 2019;16(3):336–43.
10. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform. 2013;46:354–62.
11. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. Automatic identification of critical follow-up recommendation sentences in radiology reports. AMIA Annu Symp Proc. 2011;2011:1593–602.
12. Dutta S, Long WJ, Brown DF, Reisner AT. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med. 2013;62:162–9.
13. Lau W, Payne TH, Uzuner O, Yetisgen M. Extraction and analysis of clinically important follow-up recommendations in a large radiology dataset. AMIA Jt Summits Transl Sci Proc. 2020;2020:335–44.
14. Dang PA, Kalra MK, Blake MA, Schultz TJ, Halpern EF, Dreyer KJ. Extraction of recommendation features in radiology with natural language processing: exploratory study. AJR Am J Roentgenol. 2008;191:313–20.
15. Imai T, Aramaki E, Kajino M, Miyo K, Onogi Y, Ohe K. Finding malignant findings from radiological reports using medical attributes and syntactic information. Stud Health Technol Inform. 2007;129:540–4.
16. Lou R, Lalevic D, Chambers C, Zafar HM, Cook TS. Automated detection of radiology reports that require follow-up imaging using natural language processing feature engineering and machine learning classification. J Digit Imaging. 2020;33(1):131–6.
17. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol. 2012;7:1257–62.
18. Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform. 2013;46:869–75.
19. Farjah F, Halgrim S, Buist DSM, Gould MK, Zeliadt SB, Loggers ET, et al. An automated method for identifying individuals with a lung nodule can be feasibly implemented across health systems. EGEMS. 2016. https://doi.org/10.13063/2327-9214.1254.
20. Gershanik EF, Lacson R, Khorasani R. Critical finding capture in the impression section of radiology reports. AMIA Annu Symp Proc. 2011;2011:465–9.
21. Oliveira L, Tellis R, Qian Y, Trovato K, Mankovich G. Identification of incidental pulmonary nodules in free-text radiology reports: an initial investigation. Stud Health Technol Inform. 2015;216:1027.
22. Pham A-D, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinform. 2014;15(1):266.
23. Mabotuwana T, Hall CS, Dalal S, Tieder J, Gunn ML. Extracting follow-up recommendations and associated anatomy from radiology reports. Stud Health Technol Inform. 2017;245:1090–4.
24. Morioka C, Meng F, Taira R, Sayre J, Zimmerman P, Ishimitsu D, et al. Automatic classification of ultrasound screening examinations of the abdominal aorta. J Digit Imaging. 2016;29:742–8.
25. Xu Y, Tsujii J, Chang EIC. Named entity recognition of follow-up and time information in 20 000 radiology reports. J Am Med Inform Assoc. 2012;19(5):792–9.
26. Fu S, Leung LY, Wang Y, Raulli A-O, Kallmes DF, Kinsman KA, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. JMIR Med Inform. 2019;7(2):e12109.
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.
28. 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). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86.
29. Lin Y, Tan YC, Frank R. Open Sesame: getting inside BERT's linguistic knowledge. In: Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP. Florence, Italy: Association for Computational Linguistics; 2019. p. 241–53.
30. Kuwabara R, Han C, Murao K, Satoh S. BERT-based few-shot learning for automatic anomaly classification from Japanese multi-institutional CT scan reports. Int J Comput Assist Radiol Surg. 2020;15(Suppl 1):S148–9.
31. Peng Y, Lee S, Elton DC, Shen T, Tang Y-X, Chen Q, et al. Automatic recognition of abdominal lymph nodes from clinical text. In: Proceedings of the 3rd clinical natural language processing workshop. Association for Computational Linguistics; 2020. pp. 101–10.
32. American College of Radiology. ACR practice parameter for communication of diagnostic imaging findings revised 2020. 2020. https://www.acr.org/-/media/ACR/Files/Practice-Parameters/CommunicationDiag.pdf?la=en. Accessed 10 Feb 2021.
33. Kudo T, Richardson J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 66–71.
34. Kikuta Y. BERT pretrained model trained on Japanese Wikipedia articles. 2019. https://github.com/yoheikikuta/bert-japanese. Accessed 10 Feb 2021.
35. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: 3rd International conference on learning representations, ICLR 2015. San Diego, CA, USA: 2015.
36. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: 3rd International conference on learning representations, ICLR 2015. San Diego, CA, USA: 2015.
37. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.

38. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on machine learning. New York, NY, USA: Association for Computing Machinery; 233–240, 2006. p. 233–40.
39. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80.
40. Dan J, James HM. Speech and language processing, 3rd edition in draft. 2020. https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf. Accessed 10 Feb 2021.
41. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232.
42. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22.
43. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 6639–49.
44. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34:502–8.
45. Obuchowski NA. ROC analysis. AJR Am J Roentgenol. 2005;184:364–72.
46. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27:861–74.
47. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 2018;106:249–59.
48. Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem: a review. Int J Adv Soft Comput Appl. 2015;7(3):176–204.
49. Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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). Hong Kong, China: Association for Computational Linguistics; 2019. p. 6382–8.
50. Madabushi HT, Kochkina E, Castelle M. Cost-sensitive BERT for generalisable sentence classification with imbalanced data. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda. Hong Kong, China: Association for Computational Linguistics; 2019. p. 125–34.
51. Li X, Sun X, Meng Y, Liang J, Wu F, Li J. Dice loss for data-imbalanced NLP tasks. In: Proceedings of the 58th annual meeting of the association for computational linguistics. 2020. p. 465–76.
52. Cochon LR, Kapoor N, Carrodeguas E, Ip IK, Lacson R, Boland G, et al. Variation in follow-up imaging recommendations in radiology reports: patient, modality, and radiologist predictors. Radiology. 2019;291(3):700–7.
53. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed with huge size of Japanese clinical narrative. medRxiv. 2020. https://doi.org/10.1101/2020.07.07.20148585.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.