Division of Radiology and Biomedical Engineering, Graduate School of Medicine
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

July 19, 2021

Professor Mike Conway
Editorial Board Member
BMC Medical Informatics and Decision Making

Dear Professor Conway:

Thank you for your email regarding our manuscript, "Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers" (3963b05a-0ec1-4d28-8c13-87dc2efd7388), and the valuable comments from the reviewers. We enclose our revised manuscript and point-by-point responses to the two reviewers' comments.

Changes have been made to our manuscript in response to the reviewers' comments, and the suggestions that we received from native English proofreaders have also been reflected in the revised manuscript.

Some of the figures have been updated. Please refer to the figures included in our revised manuscript, not the figures remaining on the upload system.

We hope that the revisions made to the manuscript have satisfactorily addressed the reviewers' concerns and that the manuscript is now suitable for publication in *BMC Medical Informatics and Decision Making*. Thank you for your consideration of this paper.

Sincerely,
Yuta Nakamura

Response to Reviewer #1

We would like to express our appreciation to the reviewer for reading our manuscript and the insightful comments.

1. Isn't the objective variable as actionable 0.62% too small? Can you build a model?

Response:

- We highly appreciate the comment. We have had the following two concerns about the dataset, which is partly shared by the reviewer: (i) The number of radiology reports in the positive class may be too small for the machine learning models to perform well. (ii) We collected radiology reports that had been issued only in the first three months since the deployment of the *actionable* tagging function, only 0.62% of which were actionable. However, the ratio might drastically change as the operation of the *actionable* tagging function continues for a longer time.
- To address the issue, we have collected radiology reports further by extending the period from about five months to twenty months. Consequently, we have obtained around four times as many radiology reports as in our initial submission. We have carried out again all of the necessary experiments, annotations, and analyses using the new dataset.
- We have revised "Clinical data" and "Class labeling and data split" sections to reflect the change. We have also explained more clearly that we excluded radiology reports for CT-guided biopsy and radiation therapy planning:

  > - 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: (i) eight radiology reports whose findings and impressions were both registered as empty, (ii) 254 reports for CT-guided biopsies, and (iii) 2,030 reports for CT scans for radiation therapy planning. The remaining 90,923 radiology reports corresponded to 18,388 brain, head, and neck; 64,522 body; 522 cardiac; and 5,673 musculoskeletal reports; and 3,209 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.

- The increase in the amount of data has resulted in the improvement of model performance characteristics compared with our initial submission. We show Table 3 (renumbered from Table 2) containing the updated results:

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

- Similarly, we have updated the main result mentioned in the "Results" section as below:
  - 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.
- Please note that the change in the ratio of actionable reports was small (from 0.62% to 0.87%).

2. This should be Table.2 not Table.3 in line 6 of Result section, shouldn't it?

Response:

- We apologize that our explanation was not clear enough. We surely intended to refer to the original Table 3, because it showed that the difference among the performance characteristics of BERT and other methods was statistically significant. We have revised the corresponding part as below. Please note that the original Tables 2 and 3 have been renumbered to 3 and 4:
  - 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.

3. Did five-fold cross-validation apply to BERT only?

Response:

- We highly appreciate this comment which helps our manuscript more scientifically strict.
- We have carried out hyperparameter tuning all over again using five-fold cross-validation not only for BERT but also for all of the other methods.
- First, we have updated the "BERT" subsection with a more detailed description of hyperparameter tuning:
  - 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.
  -

-

- Second, we have added Table 2 showing candidates and choices of hyperparameters:

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

- Third, we have added descriptions of hyperparameter tuning for methods other than BERT in the "Methods" section. Please note that we have included a long short-time memory (LSTM) model in baselines:
  - 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.
  - Here, we describe the details of hyperparameters of the LR and GBDT models. For LR, we used Elastic-Net regularization [30], 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.

4. The Figure.6a and 6b should be described in detail.

Response:

- Thank you for the comment. We agree that a more detailed explanation is required for how Figures 6a and 6b were created and what is described in them.
- First, we have added a description on how we obtained the visualizations in Figures 6a and 6b in the "Results" section:
  - 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.
- Second, we have also added a description of Figures 6 and 7 in the "Results" section, part of which has been moved from their captions:
  - The results are shown in Figures 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 Figure 6b. The word "apparent", which is frequently accompanied by negative findings in Japanese radiology reporting, is also present in the top negative words in Figure 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 Figures 6 and 7.

5. Please define "the 12 attention heads" in Result section.

Response:

- The definition of "the 12 attention heads" is the same as that given by Devlin J et al., 2019 (DOI: 10.18653/v1/N19-1423). We used a pre-trained BERT model whose architecture is equivalent to that of the $\text{BERT}_{\text{BASE}}$ model referred to by Devlin J et al. The $\text{BERT}_{\text{BASE}}$ model has 12 self-attention heads in each layer, as explained by Devlin J et al., 2019 (DOI: 10.18653/v1/N19-1423) as follows:
    - we denote the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A. 3 We primarily report results on two model sizes: $\text{BERT}_{\text{BASE}}$ (L=12, H=768, A=12, Total Parameters=110M) and $\text{BERT}_{\text{LARGE}}$ (L=24, H=1024, A=16, Total Parameters=340M).
- Apart from the above, the description about how attention scores were calculated was indeed quite unclear. We have revised the corresponding section with a more detailed explanation:
    - 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.

6. You should show which of the sixteen truly actionable reports (38%) were implicit in the test set.

Response:

- Thank you for the comment. First, we have updated the portion and number of the implicit actionable reports in the "Results" section:
    - 111 truly actionable reports (47%) were implicit in the test set.
- We are afraid that it is quite difficult to list all the implicit actionable reports. Instead, we have added a more detailed analysis of implicit actionable reports. We focused on the implicit actionable reports that do not contain any emphatic expression for the actionable findings (e.g., "highly suspected to be primary lung cancer" for lung nodules). Such radiology reports cannot be identified by looking at the presence or absence of particular expressions. We have found that BERT is still capable of identifying such implicit actionable reports, and in such cases, BERT has assigned high attention scores to disease names. We have added Figure 9 showing results of this analysis:
    - **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 Figure 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."
- We have also added the description of this analysis in the "Results" section:
    - 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."

7. Clearly mention the reason of methods detected actionable reports in Table 5 in Discussion section.

Response:

- We appreciate the comment. In the "Discussion" section in our original manuscript, we mentioned the capability of our method to identify implicit actionable reports, but not for non-mass actionable reports.
- Our analysis suggests that all of the methods, including the BERT-based one, tend to mainly rely on the presence or absence of expressions for recommendations, possibilities, or suspicions, rather than particular disease names. However, we suppose that the capability of BERT is not limited to considering such superficial clues. We have added Figure 9 explaining why. Figure 9 implies that BERT can identify implicit actionable reports by focusing on specific disease names. We have newly discussed this viewpoint in the "Discussion" section:
    - 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 (i) 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 (ii) 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.

Response to Reviewer #2

We would like to express our appreciation to the reviewer for reading our manuscript and the insightful comments.

1. The author compared BERT with statistical machine learning methods to prove the advantage of BERT. I think more comparison with other deep learning methods was needed.

Response:

- We highly appreciate this comment, which helped to improve our manuscript. We agree that the performance of BERT should be compared with that of another deep learning method, not only with those of statistical machine learning methods, to properly examine the impact of introducing a transformer-based model.
- We have additionally performed an experiment with a bidirectional long short-term memory (LSTM) model followed by a self-attention layer. The reason for introducing a self-attention layer was that a vanilla LSTM model alone performed far poorer than other methods in empirical experiments.
- Accordingly, we have added a new subsection, "Baselines: LSTM", in the "Methods" section as below:
    - 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, 29]. 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 × 1,536 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.
-

- We have revised the description of our methods in Table 1 by changing "SML, BERT" to "SML, LSTM, BERT":

This study	Japanese	Yes	Yes	Reports with an *actionable* tag	Order Information, Findings, Impression	SML, LSTM, BERT

- We have also revised Table 3 (renumbered from 2), describing the main result including the metrics of LSTM classifiers. Please note that there have been changes in the scores, although the main conclusions have been affected slightly. This is because we have carried out again all of the experiments with a larger amount of data by extending the targeted period to obtain radiology reports from our hospital. Below is the revised Table 3:

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

-

2. Result in Figure 6 to 8 suggested that all three methods mainly relied on whether radiology reports contain specific expressions of recommendation, suspect, or negation. The author should discuss the clinical meanings of these top expressions, since I didn't find clinical significance of expressions like "no","apparent" etc.

Response:

- First, we agree that the original Figures 6 and 7 were not sufficiently informative because symbols and function words other than content words frequently appeared in the top *n*-grams. Thus, we have improved Figures 6 and 7 by excluding *n*-grams from the visualization of model behaviors if they were Japanese punctuations, Japanese postpositional particles, or the concatenations of both. We have also revised the corresponding part in the "Results" section as below:
    - 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.
- Next, let us discuss the top expressions for negations. Although it is difficult to interpret most of the top negative words in the updated Figures 6a and 6b, some words for negations including "not" and "absent" are found there. We assume that such words often co-occur with negative findings. The word "apparent" is also present in Figure 6b, which can also be a key to detecting negative findings. This is because, in Japanese radiology reporting, the word "apparent" is frequently used to describe negative findings (e.g., "pulmonary metastasis is not apparent.") but is rarely accompanied by positive findings. Given that negative findings may rarely be actionable, these negative words and the word "apparent" may contribute as keys for the models to detecting non-actionable radiology reports. Thus, we have revised the corresponding part of the "Discussion" section as below:
    - 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 Figure 6b. The word "apparent", which is frequently accompanied by negative findings in Japanese radiology reporting, is also present in the top negative words in Figure 6b. These imply that the LR model might deduce that radiology reports are non-actionable when negative findings predominate.
- Third, our additional analysis suggests that BERT models can probably make a decision on the basis of characteristics of diseases rather than by simply looking at affirmative or negative words. We have added Figure 9, which illustrates the additional analysis:
    - **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 Figure 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."
- We have also discussed this viewpoint in the "Results" section as below:
    - 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."

3. In the classification problem, the low positive case ratio (0.62%) of data needs to be noted. In this study, oversampling or under sampling methods were not applied to solve the data imbalance problem. In the Discussion section, the author mentioned the method by Madabushi et al. However, the author didn't apply this method to solve the data imbalance problem.

Response:

- We highly appreciate this comment.
- We have carried out oversampling to include 10 times as many positive samples in the training set. We did not re-use the same hyperparameters as in the experiment without oversampling, but carefully tuned models again to select the best hyperparameters with oversampling.
- We have added a new subsection, "Oversampling", at the end of the "Methods" section:
  - 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.
  -
- Contrary to our expectations, the oversampling did not contribute to the model performance except for GBDT. We have added Tables 6 and 7 showing the results:

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

Method	Use of order information	Metrics	p values
			Oversampling (-) vs (+)
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	$p < 0.0001$

- Accordingly, we have updated the "Results" section with the explanation of the results of oversampling:
  - Oversampling showed a limited positive effect on the performance. As in Tables 6 and 7, oversampling positive samples in the training dataset ten times brought statistically significant improvement of AUPRC and average precision only for GBDT.
  - We assume that the reason for the poor contribution of simple oversampling was that it did not bring additional information to the models.

  We also recognize that other various methods are worth trying to alleviate the class imbalance including different oversampling ratios, undersampling, cost-sensitive learning, or better loss functions, although it is difficult to try all of them owing to time limitation. We have revised the corresponding part of the "Discussion" section with new references 49 and 51 (reference 46 in the original manuscript has been renumbered to 50):
  - This negative impact of low positive case ratio was not alleviated by simple oversampling, probably because it did 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.
  - 49. Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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.

- 51. Li X, Sun X, Meng Y, Liang J, Wu F, Li J. Dice Loss for Data-imbalanced NLP Tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. p. 465–76.

4. The dataset contained radiology reports corresponded to brain, head and neck, body and musculoskeletal. Writing emphasis and content in these different kinds of reports had big difference. The author should discuss the impact on the result since these reports were simply mixed together.

Response:

- Thank you for this comment, which helped us make a more detailed analysis. We have calculated the ratio of actionable reports in the test set and recall scores for each body part, which is shown in Table 10:

Body part	#Actionable reports			Recall							
				Method	LR		GBDT		LSTM		BERT	
				Use of order information	(-)	(+)	(-)	(+)	(-)	(+)	(-)	(+)
	Total	Implicit	Non-mass									
Brain, Head & Neck	23/5,584 (0.41%)	10/23 (43.5%)	16/23 (69.6%)		0.739	0.652	0.609	0.783	**0.870**	0.826	**0.870**	0.696
Body	206/19,256 (1.1%)	101/206 (49.0%)	91/206 (44.2%)		0.879	0.835	0.801	0.825	0.854	0.883	**0.903**	0.888
Cardiac	0/151 (0%)	-	-		-	-	-	-	-	-	-	-
Skeletal	9/1,758 (0.51%)	1/9 (11.1%)	6/9 (66.7%)		**1.000**	0.667	0.889	**1.000**	**1.000**	**1.000**	**1.000**	0.889
Other	0/959 (0%)	-	-		-	-	-	-	-	-	-	-

- We have also added the description in the "Results" section as below:
  - 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.
- The analysis above has revealed the different proportions of subsets of actionable reports among body parts, which may result in better performance for musculoskeletal actionable reports than for brain, head, and neck ones. We have added this discussion at the end of the "Results" section as below:
  - 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.