

## **Title Page**

### Title

Fully automatic summarization of radiology reports using natural language processing with language models.

### Authors

Mizuho Nishio, MD, PhD<sup>a,\*</sup>, Takaaki Matsunaga, MD<sup>a</sup>, Hidetoshi Matsuo, MD, PhD<sup>a</sup>,  
Munenobu Nogami, MD, PhD<sup>a,b</sup>, Yasuhisa Kurata, MD, PhD<sup>c</sup>, Koji Fujimoto, MD, PhD<sup>d</sup>,  
Osamu Sugiyama, PhD<sup>e</sup>, Toshiaki Akashi, MD, PhD<sup>f</sup>, Shigeki Aoki, MD, PhD<sup>f</sup>, Takamichi  
Murakami, MD, PhD<sup>a</sup>

<sup>a</sup> Department of Radiology, Kobe University Graduate School of Medicine, 7-5-2 Kusunoki-cho, Chuo-ku, Kobe 650-0017, Japan

<sup>b</sup> Division of Medical Imaging, Biomedical Imaging Research Center, University of Fukui, 23-3 Matsuokashimoaizuki, Eiheiji, Yoshida, Fukui 910-1193, Japan

<sup>c</sup> Department of Diagnostic Imaging and Nuclear Medicine, Kyoto University Graduate School of Medicine, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan.

<sup>d</sup> Advanced Imaging in Medical Magnetic Resonance, Kyoto University Graduate School of Medicine, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan.

<sup>e</sup> Department of Informatics, Kindai University, 3-4-1 Kowakae, Higashiosaka City 577-8502, Japan

<sup>f</sup> Department of Radiology, Juntendo University Graduate School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo, 113-8421, Japan.

\*Corresponding author:

Mizuho Nishio, MD, PhD

Department of Radiology

Kobe University Graduate School of Medicine  
7-5-2 Kusunoki-cho, Chuo-ku, Kobe 650-0017, Japan  
Tel.: +81-78-382-6104  
Fax: +81-78-382-6129  
E-mail: nishiomizuho@gmail.com

#### Funding information

This work was supported by JSPS KAKENHI (Grant Numbers: 22K07665, 23K07154, 23K17229, and 23KK0148).

#### Acknowledgement

We thank JMID project and the following 10 academic medical centers.

- Juntendo University
- Kyushu University
- Keio University
- The University of Tokyo
- Okayama University
- Kyoto University
- Osaka University
- Hokkaido University
- Ehime University
- Tokushima University

#### Abbreviations

- Chest x-ray = CXR
- Japan Medical Image Database = JMID
- Natural language processing = NLP

- Recall-Oriented Understudy for Gisting Evaluation = ROUGE
- Text-to-Text Transfer Transformer = T5

## **Abstract**

Natural language processing using language models has yielded promising results in various fields. The use of language models may help improve the workflow of radiologists. This retrospective study aimed to construct and evaluate language models for the automatic summarization of radiology reports. Two datasets of radiology reports were used: MIMIC-CXR and the Japan Medical Image Database (JMID). MIMIC-CXR is an open dataset comprising chest radiograph reports. JMID is a large dataset of CT and MRI reports comprising reports from 10 academic medical centers in Japan. A total of 128,032 and 1,101,271 reports from the MIMIC-CXR and JMID, respectively, were included in this study. Four Text-to-Text Transfer Transformer (T5) models were constructed. Recall-Oriented Understudy for Gisting Evaluation (ROUGE), a quantitative metric, was used to evaluate the quality of text summarized from 19,205 and 58,043 test sets from MIMIC-CXR and JMID, respectively. The Wilcoxon signed-rank test was utilized to evaluate the differences among the ROUGE values of the four T5 models. In addition, subsets of automatically summarized text in the test sets were manually evaluated by two radiologists. Based on the Wilcoxon signed-rank test, the best T5 models were selected for the automatic summarization. The quantitative metrics of the best T5 models were as follows: ROUGE-1 =  $57.75 \pm 30.99$ , ROUGE-2 =  $49.96 \pm 35.36$ , and ROUGE-L =  $54.07 \pm 32.48$  in MIMIC-CXR; ROUGE-1 =  $50.00 \pm 29.24$ , ROUGE-2 =  $39.66 \pm 30.21$ , and ROUGE-L =  $47.87 \pm 29.44$  in JMID. The radiologists' evaluations revealed that 86% (86/100) and 85% (85/100) of the texts automatically summarized from MIMIC-CXR and JMID, respectively, were clinically useful. The T5 models constructed in this study were capable of automatic summarization of radiology reports. The radiologists' evaluations revealed that most of the automatically summarized texts were clinically valuable.

## **Introduction**

Radiology reports are a valuable source of information for improving clinical practice and supporting research. A multitude of radiology reports have been written in recent years owing to the advances in the field of radiology. However, manually processing a large number of unstructured reports is difficult as radiology reports are often recorded as unstructured data.

Natural language processing (NLP) has enabled computers to process natural languages (1,2) by facilitating the extraction of structured information from electronic medical records and radiology reports. Consequently, NLP has been used for text classification, text summarization, and text generation in the field of radiology (3–5). Recent advances in NLP have been accompanied by the application of deep learning.

NLP has the potential to reduce the workload of radiologists by extracting structured information from radiology reports. This would aid clinicians and radiologists in the decision-making process and in identifying patients for research. However, unlike computer vision (6), NLP has not received significant attention in the field of artificial intelligence, and reviews on the application of NLP in radiology have been limited (5).

The development of language models has been a promising advancement in NLP. Language models are neural networks trained using a large amount of text, and the number of parameters in the model can be used as a measure of performance. Several types of language models have been developed in recent years, such as Bidirectional Encoder Representations from Transformers (7), Text-to-Text Transfer Transformer (T5) (8,9), and Generative Pre-Training-1 (10), Generative Pre-Training-2 (11), Generative Pre-Training-3 (12) and so on (13). These models have achieved state-of-the-art performance in NLP tasks.

Radiology reports are divided into two sections: findings and impression. Automatic summarization of the impression section from the findings section would reduce the workload of radiologists. Thus, this study aimed to investigate the effectiveness of using a language model to summarize radiology reports automatically. The contributions of this study are threefold: (i) The T5 language model was used to summarize radiology reports automatically.

(ii) Automatic summarization of radiology reports was performed in two languages: chest radiograph (CXR) reports in English and CT and MRI reports in Japanese. (iii) Recall-Oriented Understudy for Gisting Evaluation (ROUGE), a quantitative evaluation metric, and a semi-quantitative evaluation performed by radiologists were used to evaluate the automatically summarized sentences, and the relationship between the ROUGE metrics and radiologists' evaluations was investigated.

## **Material and method**

This retrospective study was approved by the Institutional Review Boards of the Japan Medical Image Database (JMID) project and Kobe University Hospital. The requirement for informed consent was waived. This study was conducted in accordance with the Checklist for Artificial Intelligence in Medical Imaging (14).

### **Dataset**

Two datasets were used in the study: MIMIC-CXR and JMID. The MIMIC-CXR dataset comprises chest radiographs and the corresponding reports (15), whereas the JMID dataset was obtained from the JMID project, a project wherein 10 academic medical centers in Japan collaborated to create a large radiology database with de-identified patient data. All reports dated from 8/4/2010 to 3/31/2023 were collected from the JMID database. Reports with missing data in the findings or impression sections were excluded. Pairs of findings and impression sections were collected from each report in the two datasets. Figure 1 presents the flow diagrams of the study.

### **Ground truth**

MIMIC-CXR and JMID include clinical radiology reports; thus, the impression sections of actual reports were used as the ground truth.

### Dataset partition

The MIMIC-CXR images were randomly divided into training, validation, and test sets at a 16:1:3 ratio. The reports acquired from JMID were divided into three sets according to the date of the reports: training set, 8/4/2010–11/30/2022 and 12/10/2022–12/31/2022; validation set, 12/1/2022–12/9/2022; and test set, 1/1/2023–3/31/2023. Date-based dataset partitioning was not possible in MIMIC-CXR as the MIMIC-CXR reports were not dated.

### Language model

T5 is a transformer-based neural network model comprising an encoder and decoder (8,9). Fine-tuning the pre-trained T5 model can improve the performance of text summarization, question answering, and text classification (16,17). The present study focused on text summarization (summarization of radiology reports). Consequently, the input was the text of the findings section, whereas the output was that of the impressions section. The T5 model was trained to summarize the findings section automatically via fine-tuning. Pre-trained T5 models were obtained from Hugging Face (<https://huggingface.co/models>) for fine-tuning. Two pre-trained T5 models (“t5-base” (18) and “google/mt5-base” (19)) were obtained for MIMIC-CXR. Two pre-trained T5 models (“megagonlabs/t5-base-japanese-web” (20) and “google/mt5-base” (19)) were obtained for JMID. “t5-base,” “megagonlabs/t5-base-japanese-web,” and “google/mt5-base” are the pre-trained English, Japanese, and multilingual models, respectively.

### Model training

Figure 2 summarizes the process of model development and prediction using the T5 model. Fine-tuning of the T5 models is detailed in the Appendix 1. Batch sizes of 2 and 8 were used for fine-tuning. There were two choices of pre-trained models and two types of batch sizes for MIMIC-CXR and JMID, resulting in a total of four combinations.

### Evaluation

The summarized text obtained from the fine-tuned T5 models was evaluated quantitatively using ROUGE metrics (21,22) and semi-quantitatively by radiologists. First, the ROUGE metrics were calculated between the impression text of the actual report and the predicted text. The ROUGE metrics indicate the summarization quality by measuring the alignment between the model-generated and original radiologist-generated summaries. The ROUGE metrics are detailed in the Appendix 2. Next, 100 reports were randomly selected from the test sets to undergo a semi-quantitative evaluation. Two radiologists with 17 and 7 years of experience in clinical radiology independently rated the predicted impression of the 100 reports on a 5-point scale as the semi-quantitative evaluation. The following text was evaluated: (i) the pairs of actual and predicted impression sections and (ii) the finding sections of the actual reports. The 5-point scores were defined as follows: 1, the predicted impression could not be used clinically without rewriting; 2, most of the predicted impressions requires rewriting to be clinically useful; 3, approximately half of the predicted impression requires rewriting to be clinically useful; 4, the predicted impression is clinically usable with minor modifications; and 5, the predicted impression is clinically usable without modification. The summarized text with scores of 4 and 5 was considered clinically useful. A consensus was reached through discussion in case of disagreements between the two radiologists.

### Statistics

The differences among the ROUGE metrics of the four T5 models were compared using the Wilcoxon signed-rank test. Quadratic-weighted kappa values were calculated for the scores of the two radiologists. The kappa values were interpreted using the following criteria: 0.00–0.20, none to slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect agreement. Spearman's correlation coefficients between the ROUGE-2 values and radiologists' consensus scores were determined for the 100 reports of MIMIC-CXR and JMID, and the coefficients were statistically evaluated. Statistical significance was set at a p-value of 0.05. Statistical analyses were performed using R (version 4.2.2), Python (version 3.8.8), and Scipy (version 1.10.1).



## **Results**

### **Dataset**

Figure 1 presents the flowchart of the study. Table 1 presents the characteristics of the two datasets used in this study. Among the 227,781 reports present in the MIMIC-CXR dataset, 128,032 reports had no missing data. The findings or impression sections were frequently missing in the reports of the MIMIC-CXR dataset. The MIMIC-CXR dataset was divided into three subsets: a training set comprising 102,425 reports; a validation set comprising 6,402 reports; and a test set comprising 19,205 reports. The modality was radiography, and the location was specified as the chest in the MIMIC-CXR dataset, and the reports were written in English. Among the 1,120,311 reports in the JMID dataset, 1,101,271 had no missing data. The training, validation, and test sets comprised 1,035,574, 7,654, and 58,043 reports, respectively. The modalities in the JMID dataset were CT and MRI, with locations spanning different parts of the body, and the reports were written in Japanese.

### **Model performance**

Table 2 presents the ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L) results of the fine-tuned T5 models in the MIMIC-CXR and JMID test sets. The results of the fine-tuned T5 models in the validation sets of MIMIC-CXR and JMID datasets are presented in the Appendix 3. Table 2 presents the ROUGE results for the four combinations of fine-tuned T5 models in the MIMIC-CXR and JMID datasets. The fine-tuned “t5-base” model with a batch size of 2 demonstrated the highest ROUGE-1, ROUGE-2, and ROUGE-L values (ROUGE-1 =  $57.75 \pm 30.99$ , ROUGE-2 =  $49.96 \pm 35.36$ , and ROUGE-L =  $54.07 \pm 32.48$ ) in the MIMIC-CXR dataset. In contrast, the fine-tuned “google/mt5-base” model with a batch size of 8 achieved the lowest values (ROUGE-1 =  $52.35 \pm 31.36$ , ROUGE-2 =  $43.97 \pm 35.48$ , and ROUGE-L =  $49.05 \pm 32.53$ ). The fine-tuned “google/mt5-base” model with a batch size of 2 demonstrated the highest ROUGE-1, ROUGE-2, and ROUGE-L values (ROUGE-1 =  $50.00 \pm 29.24$ , ROUGE-2

=  $39.66 \pm 30.21$ , and ROUGE-L =  $47.87 \pm 29.44$ ) in the JMID dataset. In contrast, the fine-tuned “megagonlabs/t5-base-japanese-web” model with a batch size of 8 achieved the lowest values (ROUGE-1 =  $44.30 \pm 28.56$ , ROUGE-2 =  $34.01 \pm 28.44$ , and ROUGE-L =  $42.11 \pm 28.54$ ). These results underscore the interaction between the type of pre-trained model, batch size, and dataset characteristics.

The differences in the ROUGE-1, ROUGE-2, and ROUGE-L values were statistically evaluated between each pair of the four fine-tuned models. The p-values of ROUGE-1, ROUGE-2, and ROUGE-L were  $<.001$  in the pairs of the four models in the MIMIC-CXR dataset, except for the p-values of ROUGE-1 between “google/mt5-base” with a batch of 2 and “t5-base” with a batch of 8. The p-values of ROUGE-1 between “google/mt5-base” with a batch of 2 and “t5-base” with a batch of 8 was 0.12. As the number of test set was larger in JMID dataset than that in MIMIC-CXR dataset, the p-values of ROUGE-1, ROUGE-2, and ROUGE-L were  $<.001$  for each pair of the four models in JMID. Thus, we focused on the optimal fine-tuned models (the fine-tuned “t5-base” model with a batch size of 2 for MIMIC-CXR and the fine-tuned “google/mt5-base” model with a batch size of 2 for JMID) based on the p-values.

Table 3 presents the results of the radiologists' semi-quantitative scores for the predicted summaries of the 100 reports generated by the optimal fine-tuned models of MIMIC-CXR and JMID. The kappa values of the semi-quantitative scores between the two radiologists were 0.785 (95% confidence interval = 0.669–0.900) and 0.736 (95% confidence interval = 0.590–0.883) for the 100 reports acquired from the MIMIC-CXR and JMID test sets, respectively, indicating substantial agreement between the two radiologists. The number of reports for the consensus scores was as follows: score 1 = 1, score 2 = 2, score3 = 11, score 4 = 15, and score 5 = 71 in MIMIC-CXR; score 1 = 2, score 2 = 3, score3 = 10, score 4 = 25, and score 5 = 60 in JMID. These results indicate that 86% (86/100) and 85% (85/100) of the automatically summarized texts were clinically useful in the MIMIC-CXR and JMID datasets, respectively.

Figure 3 presents the scatter plots illustrating the relationships between the

ROUGE-2 values and semi-quantitative scores of the two radiologists. Significant positive correlations were observed between the ROUGE-2 values and semi-quantitative scores in both datasets. The calculated correlation coefficients were 0.446 (95% confidence interval = 0.274–0.591) and 0.261 (95% confidence interval = 0.0681–0.435) for the MIMIC-CXR and JMID datasets, respectively. The corresponding p-values of the correlation coefficients were  $<.001$  for both datasets, indicating that the positive correlations observed between the ROUGE-2 values and the semi-quantitative scores were statistically significant. Scatter plots of the ROUGE-2 values and semi-quantitative scores for each of the two radiologists are presented in the Appendix 4 and 5. Figure 4 presents representative examples of radiology reports and the summary text predicted by the fine-tuned T5 models.

## **Discussion**

An automatic summarization model of NLP was constructed using the T5 model. The fine-tuned T5 model was capable of summarizing radiology reports automatically. The present study revealed that the fine-tuned “t5-base” model with a batch size of 2 for MIMIC-CXR and the fine-tuned “google/mt5-base” model with a batch size of 2 for the JMID were the best T5 models. The scores of the radiologists' semi-quantitative evaluation of the 100 reports were  $\geq 4$  for 86% in the MIMIC-CXR test set and 85% in the JMID test set. This indicates that most of the predicted text in the impression section was clinically useful. These results demonstrate the usefulness of the T5 summarization models in the automated summarization of radiology reports. In addition, statistically significant correlations were observed between the radiologists' semi-quantitative scores and the quantitative evaluation using ROUGE metrics for the MIMIC-CXR and JMID datasets.

MIMIC-CXR is a dataset comprising reports and CXR images (15); however, the anatomical locations and diseases included in MIMIC-CXR are relatively limited. In contrast, the JMID dataset comprises CT and MR images of all locations, and the anatomical locations and diseases involved are broader than those in MIMIC-CXR. Consequently, report variation

was greater in the JMID dataset, and report summarization was more difficult in JMID than in MIMIC-CXR. Table 3 demonstrates that the radiologists' scores were comparable for the two datasets. JMID comprised a greater number of reports, approximately ten times more than MIMIC-CXR (Table 1). The results presented in Tables 1 and 3 suggest that the number of radiology reports influenced the performance of the automatic summarization models. Thus, the dataset size for constructing language models may be important in NLP, similar to computer vision (23,24).

The two different pre-trained T5 models were used for MIMIC-CXR and JMID in the present study: “t5-base” and “google/mt5-base” for MIMIC-CXR; “megagonlabs/t5-base-japanese-web” and “google/mt5-base” for JMID. The best ROUGE values were obtained with “t5-base” for MIMIC-CXR and “google/mt5” for JMID (Table 2). This result indicates that the English T5 model outperformed the multilingual T5 model in the English task, whereas the multilingual T5 model outperformed the Japanese T5 model in the Japanese task. In general, considering the variation in dataset sizes based on the languages, English stands out owing to its larger size. The results of the present study and the dataset size of English datasets suggest that utilizing models pre-trained in English datasets would be more effective for English tasks because the pre-trained model specialized for English has sufficient generalizability. However, the dataset size used for pre-training was smaller for languages other than English. Consequently, the multilingual model would be more effective for non-English tasks because the multilingual model was pre-trained using both English and non-English datasets (8).

Similar to a previous study (21), statistically significant correlations were observed between the ROUGE-2 values and the radiologists' semi-quantitative scores for the 100 reports selected from the test sets. The Spearman's correlation coefficients were 0.446 for MIMIC-CXR and 0.261 for JMID, indicating weak correlations. The ROUGE-2 values exhibited significant variability even when the radiologists' score was 5 (Figure 3). Therefore, evaluating individual summarized sentences by relying solely on the ROUGE metrics may not be reliable. However, the average ROUGE values could potentially serve as a surrogate for the average

radiologist scores when the test set (evaluation dataset) is sufficiently large. These two evaluation methods should be used complementarily, particularly when the evaluation dataset is limited.

JMID is a large dataset with a size of more than one million. Creating a dataset larger than that of the JMID is difficult. Conversely, to achieve significant improvements in the performance of automatic summarization, significant improvements are required in the summarization model architecture or the pre-trained model.

This study had certain limitations. First, we speculate that the usefulness of a multilingual or language-specific pre-trained model depends on the language of the radiology reports. However, this may depend on the dataset characteristics other than the language. As only two datasets were utilized, this aspect could not be adequately evaluated in this study. Second, only two languages were used in this study. Other languages should be investigated in future studies. Third, the dataset sizes of MIMIC-CXR and JMID were relatively large for medical NLP. Smaller datasets were not used in this study.

In conclusion, this study demonstrated the feasibility of automatic report summarization, and the majority of the automatically summarized text was clinically useful. Significant correlations were observed between the semi-quantitative evaluations performed by radiologists and the quantitative evaluations by the ROUGE metrics. The use of quantitative assessments provided by the ROUGE metrics in conjunction with the semi-quantitative scores provided by radiologists could be complementary. These results could lead to further advances in NLP radiology research.

## **References**

1. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl* 2023;82:3713-3744. doi: [10.1007/s11042-022-13428-4](https://doi.org/10.1007/s11042-022-13428-4).
2. Chowdhary KR. Natural Language Processing. *Fundamentals of Artificial Intelligence*. New Delhi: Springer India, 2020; 603-649. doi: [10.1007/978-81-322-3972-7\\_19](https://doi.org/10.1007/978-81-322-3972-7_19).
3. Pons E, Braun LM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. *Radiology* 2016;279:329-343. doi: [10.1148/radiol.16142770](https://doi.org/10.1148/radiol.16142770).
4. Linna N, Kahn CE Jr. Applications of natural language processing in radiology: A systematic review. *Int J Med Inform* 2022;163:104779. doi: [0.1016/j.ijmedinf.2022.104779](https://doi.org/10.1016/j.ijmedinf.2022.104779).
5. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, Grover C, Suárez-Paniagua V, Tobin R, Whiteley W, Wu H, Alex B. A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak* 2021;21:179. doi: [10.1186/S12911-021-01533-7](https://doi.org/10.1186/S12911-021-01533-7).
6. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018;9:611-629. doi: [10.1007/s13244-018-0639-9](https://doi.org/10.1007/s13244-018-0639-9).
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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; 4171-4186. doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
8. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

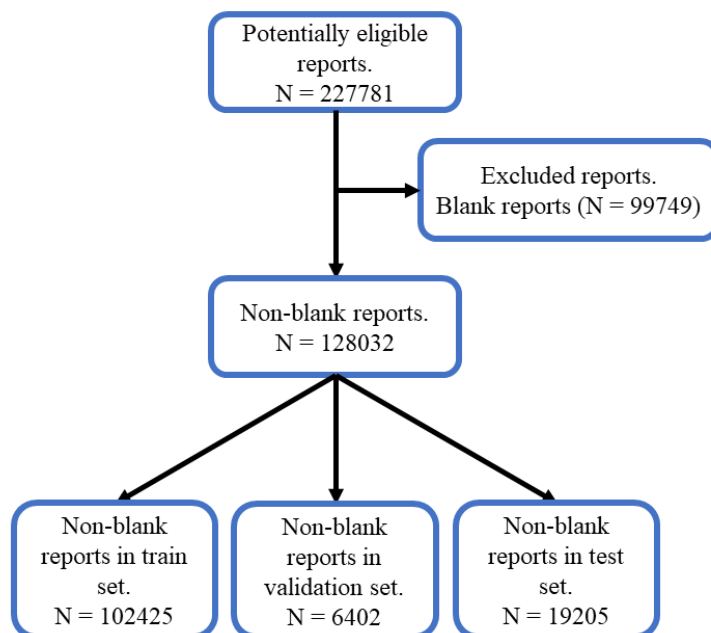
- Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021;483-498. doi: [10.18653/v1/2021.naacl-main.41](https://doi.org/10.18653/v1/2021.naacl-main.41).
9. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J Mach Learn Res* 2020;21:5485-5551.
  10. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. 2018. [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf)
  11. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. 2019. <https://d4mucfpxyww.cloudfront.net/better-language-models/language-models.pdf>
  12. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language Models are Few-Shot Learners. *ArXiv* 2020;abs/2005.14165. <https://arxiv.org/abs/2005.14165>.
  13. He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *Proceedings in ICLR2021*. 2021.
  14. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* 2020;2:e200029. doi: [10.1148/RYAI.2020200029](https://doi.org/10.1148/RYAI.2020200029).
  15. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019;6:317. doi: [10.1038/s41597-019-0322-0](https://doi.org/10.1038/s41597-019-0322-0).
  16. Mastropaolo A, Scalabrino S, Cooper N, Palacio DN, Poshyvanyk D, Oliveto R, Bavota G. Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks. 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). 2021;336-347. doi: [10.1109/ICSE43902.2021.00041](https://doi.org/10.1109/ICSE43902.2021.00041).

17. Abacha AB, Yim WW, Adams G, Snider N, Yetisgen-Yildiz M. Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations. Proceedings of the 5th Clinical Natural Language Processing Workshop. Toronto, Canada: Association for Computational Linguistics, 2023;503-513. doi: [10.18653/v1/2023.clinicalnlp-1.52](https://doi.org/10.18653/v1/2023.clinicalnlp-1.52).
18. t5-base · Hugging Face. <https://huggingface.co/t5-base>. Accessed September 22, 2023.
19. google/mt5-base · Hugging Face. <https://huggingface.co/google/mt5-base>. Accessed September 22, 2023.
20. megagonlabs/t5-base-japanese-web · Hugging Face. <https://huggingface.co/megagonlabs/t5-base-japanese-web>. Accessed September 22, 2023.
21. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004; 74-81. <https://aclanthology.org/W04-1013>.
22. Lin CY, Hovy E. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003;150-157. <https://aclanthology.org/N03-1020>.
23. Luo C, Li X, Wang L, He J, Li D, Zhou J. How Does the Data set Affect CNN-based Image Classification Performance? 2018 5th International Conference on Systems and Informatics (ICSAI). 2018;361-366. doi: [10.1109/ICSAI.2018.8599448](https://doi.org/10.1109/ICSAI.2018.8599448).
24. Dawson HL, Dubrule O, John CM. Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification. Comput Geosci 2023;171:105284. doi: [10.1016/j.cageo.2022.105284](https://doi.org/10.1016/j.cageo.2022.105284).

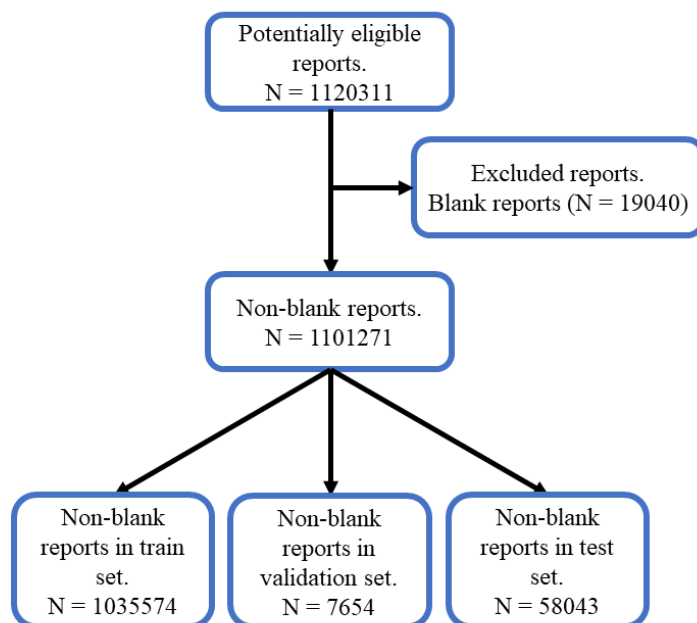


## Figures and Figure Captions

### (A) MIMIC-CXR



### (B) JMID



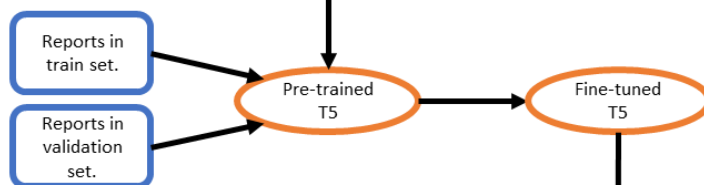
**Figure 1.** Flowcharts of including radiology reports. (A) MIMIC-CXR, (B) JMID. Abbreviations: JMID, Japan Medical Image Database

(A)

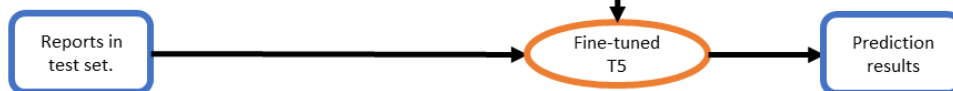
Phase I: Obtaining pre-trained model



Phase II: Fine-tuning



Phase III: Prediction



(B)

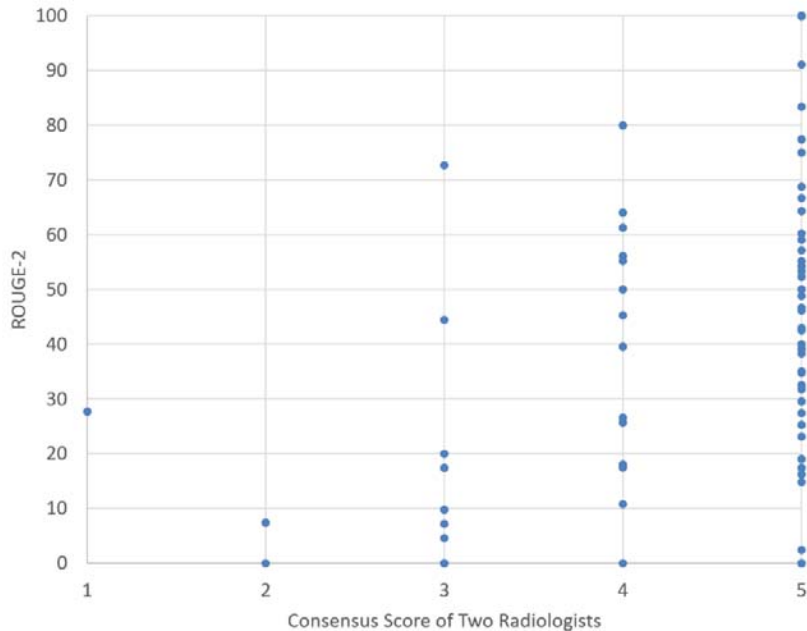
Input text



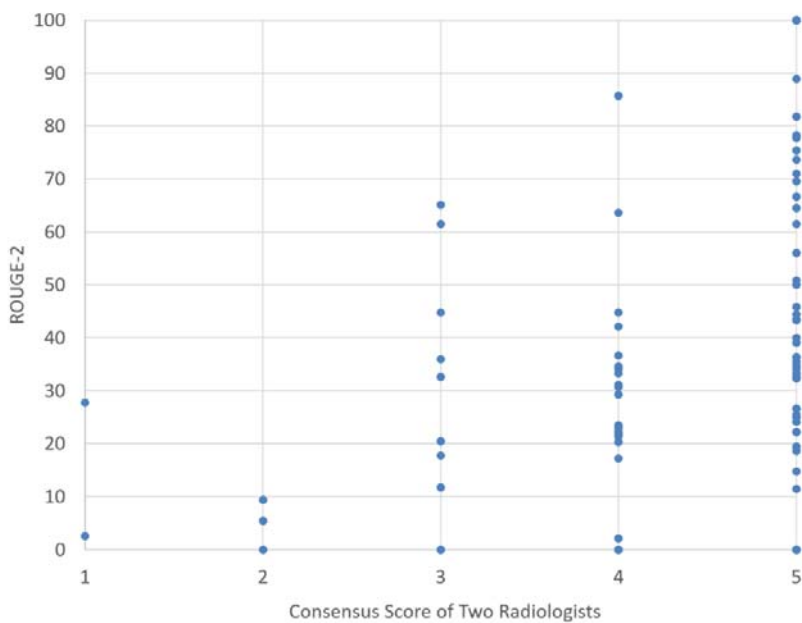
**Figure 2.** Outline of model development and prediction. (A) Flow of obtaining a pre-trained T5 model, fine-tuning the T5 model from the pre-trained model, and predicting the text of the impression section with the fine-tuned model. (B) Examples of the summary text predicted from the findings section.

Abbreviations: T5, Text-to-Text Transfer Transformer

(A) MIMIC-CXR



(B) JMID



**Figure 3.** Scatter plots between the ROUGE-2 values and semi-quantitative consensus scores by two radiologists. (A) Scatter plot for MIMIC-CXR, (B) Scatter plot for JMID. Note for (A): Correlation coefficient and p-value between ROUGE-2 values and consensus scores were 0.446 and  $<.001$ , respectively. Note for (B): Correlation coefficient and p-value between ROUGE-2 values and consensus scores were 0.261 and  $<.001$ , respectively.

Abbreviations: JMID, Japan Medical Image Database; ROGUE, Recall-Oriented Understudy for Gisting Evaluation

(A)

### Finding

---

The cardiomediastinal and hilar contours are within normal limits. Lungs are well expanded and clear. There is no focal consolidation, pleural effusion or pneumothorax.

### Impression

---

No acute cardiopulmonary process.

### Predicted impression

---

No acute cardiopulmonary process.

(B)

### Finding

---

Compared to chest radiographs from \_\_\_\_, right infrahilar, right basilar, retrocardiac and left basilar opacities have increased, concerning for worsening atelectasis or aspiration. There is increasing central vascular congestion and new moderate pulmonary edema. Small right pleural effusion has worsened. No appreciable effusion on the left. No pneumothorax. Mildly tortuous and calcified thoracic aorta is stable. ETT in standard placement, unchanged. Left subclavian central line terminates in the upper right atrium. Enteric tube descends below the diaphragm and out of the field-of-view.

### Impression

---

1. Increased central vascular congestion with new moderate pulmonary edema. 2. Worsening small right pleural effusion. 3. Increased right infrahilar, right basilar, retrocardiac and left basilar opacities, suggestive of worsening atelectasis or aspiration.

### Predicted impression

---

1. Worsening right infrahilar, right basilar, retrocardiac and left basilar opacities concerning for worsening atelectasis or aspiration. 2. Worsening central vascular congestion and new moderate pulmonary edema. 3. Worsening small right pleural effusion.

(C)

### Finding

---

Dialysis catheter terminates in the right atrium but has been perhaps retracted slightly. The cardiac, mediastinal and hilar contours appear stable including moderate cardiomegaly. There is no trace pleural effusions are suspected. Streaky retrocardiac opacity suggests atelectasis. The lungs appear otherwise clear.

### Impression

---

Perhaps slight retraction of dialysis catheter. No evidence of acute cardiopulmonary disease. Stable cardiomegaly.

### Predicted impression

---

Dialysis catheter terminates in the right atrium but has been perhaps retracted slightly. No evidence of acute cardiopulmonary disease.

(D)

### Finding

---

AP upright and lateral views of the chest provided. Mild basal atelectasis noted. Lungs are otherwise clear though hyperinflated and somewhat lucent likely reflecting known emphysema. Cardiomedial silhouette is stable. Bony structures are intact. No free air below the right hemidiaphragm. Calcifications are prominent in the right and left neck likely reflecting prominent carotid bulb calcification. Bony defect at the right distal clavicle may reflect acute or chronic injury for which clinical correlation advised.

### Impression

---

1. Emphysema with mild bibasilar atelectasis. 2. Prominent carotid bulb calcifications and a carotid ultrasound may be considered to further assess. 3. Bony defect at the right distal clavicle, correlate for focal pain as an acute fracture difficult to exclude.

### Predicted impression

---

Emphysema. No free air below the right hemidiaphragm.

**Figure 4.** Representative example of radiology reports and predicted summary text from MIMIC-CXR. Note: The radiologists' scores are as follows: A = 5, B = 5, C = 4, D = 2.



## **Tables**

Table 1: Characteristics of datasets.

Item	MIMIC-CXR	JMID
Number of reports	227781	1120311
Number of non-blank reports	128032	1101271
Number of reports in train set	102425	1035574
Number of reports in validation set	6402	7654
Number of reports in test set	19205	58043
Age (year)		
train set	Not available	62.33 ± 18.49
validation set	Not available	62.78 ± 18.16
test set	Not available	62.46 ± 18.82
Sex (male:female)		
train set	Not available	555975:479599
validation set	Not available	4002:3652
test set	Not available	30463:27580
Modality	X-ray	CT, MRI
Location	Chest	Various locations
Language	English	Japanese
Private/Public datasets	Public dataset	Private dataset

Note: Non-blank reports indicate that neither the findings nor impression sections are blank in the radiology reports. Age was not available for 16754, 726, and 6210 reports in the training, validation, and test sets, respectively.

Abbreviations: JMID, Japan Medical Image Database

Table 2: ROUGE values in the test sets of MIMIC-CXR and JMID.

Dataset	T5 model	ROUGE-1	ROUGE-2	ROUGE-L
MIMIC-CXR	t5-base, B=2	57.75 ± 30.99	49.96 ± 35.36	54.07 ± 32.48
MIMIC-CXR	t5-base, B=8	55.27 ± 30.55	46.98 ± 34.79	51.34 ± 31.99
MIMIC-CXR	google/mt5-base, B=2	55.84 ± 31.26	47.86 ± 35.61	52.54 ± 32.59
MIMIC-CXR	google/mt5-base, B=8	52.35 ± 31.36	43.97 ± 35.48	49.05 ± 32.53
JMID	megagonlabs/t5-base-japanese-web, B=2	48.99 ± 29.49	38.80 ± 30.20	46.91 ± 29.64
JMID	megagonlabs/t5-base-japanese-web, B=8	44.30 ± 28.56	34.01 ± 28.44	42.11 ± 28.54
JMID	google/mt5-base, B=2	50.00 ± 29.24	39.66 ± 30.21	47.87 ± 29.44
JMID	google/mt5-base, B=8	46.18 ± 28.39	35.69 ± 28.65	43.91 ± 28.46

Note: Values represent the mean ± standard deviation. The numbers of reports are 19205 and 58043 for the test sets of MIMIC-CXR and JMID, respectively.

Abbreviations: T5, Text-to-Text Transfer Transformer; B, batch size; ROUGE, Recall-Oriented Understudy for Gisting Evaluation.



Table 3: Results of semi-quantitative evaluation by radiologists in 100 radiology reports of test sets.

Item	MIMIC-CXR	JMID
Number of reports evaluated by radiologists	100	100
Number of Score 1	1	2
Number of Score 2	2	3
Number of Score 3	11	10
Number of Score 4	15	25
Number of Score 5	71	60

Note: Scores were determined by the consensus of the two radiologists.  
Abbreviations: JMID, Japan Medical Image Database

## **Appendix 1**

### **Details of Text-to-Text Transfer Transformer (T5) model**

Text-to-Text Transfer Transformer (T5) is a transformer-based deep learning model that uses a text-to-text approach. Most NLP tasks, including translation, question answering, and classification, comprise giving input sentences to models and training them to generate target sentences. This facilitates the use of the same model, loss function, and hyperparameters for a variety of tasks. The T5 model has achieved state-of-the-art results on many NLP benchmarks, while maintaining sufficient flexibility to be fine-tuned for a variety of important downstream tasks.

A large amount of unlabeled text and an objective analogous to BERT's "masked language modeling" were used to pre-train the T5 model. Tokens of input text were randomly corrupted with special tokens during pre-training. The pre-trained T5 reconstructed the corrupted tokens of the input text after pre-training.

The pre-trained T5 model can be used for a variety of text-to-text tasks; however, text summarization of radiology reports was performed as a downstream task in this study. The following pre-trained T5 models were used to summarize the reports in the MIMIC-CXR and JMID datasets.

- MIMIC-CXR: t5-base or google/mt5-base
- JMID: megagonlabs/t5-base-japanese-web or google/mt5-base

The t5-base, megagonlabs/t5-base-japanese-web, and google/mt5-base were pre-trained with unlabeled English, Japanese, and multilingual text, respectively.

A character encoding conversion to UTF-8 was performed for the Japanese text of JMID as a preprocessing step before fine-tuning the T5 models. No preprocessing was performed for MIMIC-CXR.

The following hyperparameters were used to fine-tune the pre-trained T5 models.

- Number of tokens in input: 1024
- Number of tokens in output: 128
- Number of training epochs: 5
- Batch size: 2 or 8
- Number of steps for gradient accumulation: 32
- Learning rate: 5e-5
- Learning scheduler: cosine annealing
- Warmup ratio in cosine annealing: 0.05

The following parameter was used to generate the predicted impression using the fine-tuned T5 models.

- Number of tokens in beam search: 6

Python (version, 3.8.8), pytorch (version, 1.8.0), and transformers (version, 4.22.2) were used for model development. The T5 models were developed and evaluated on

a workstation with NVIDIA(R) RTX(TM) A6000.

For the model development and summary prediction, run\_summarization.py of transformers was used in the present study ([https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run\\_summarization.py](https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run_summarization.py)).

## **Appendix 2**

### **Details of ROUGE**

ROUGE is a commonly used metric that is used to evaluation of text summarization. This metric measures the alignment between human-generated summaries (reference summaries) and summaries generated by a model. ROUGE has several variants. ROUGE-1, ROUGE-2, and ROUGE-L were used in this study. ROUGE-1 and ROUGE-2 are basic metrics that measure the alignment on an n-gram basis. ROUGE-1 and ROUGE-2 use unigrams and bigrams, respectively. The original definition of ROUGE-1 and ROUGE-2 is as follows:

$$ROUGE - n = \frac{\sum_{S \in references} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in references} \sum_{gram_n \in S} Count(gram_n)},$$

where  $n$  represents the length of the n-gram and  $Count_{match}(gram_n)$  represents the maximum number of n-gram co-occurring in a model-generated summary and a reference summary. Three options can be used to calculate ROUGE-1 and ROUGE-2: recall basis, precision basis, and F-measure basis. F-measure-based ROUGE-1 and ROUGE-2 were used in this study.

ROUGE-1 and ROUGE-2 use the frequency of n-gram co-occurring; however, ROUGE-L uses common subsequence between human-generated summaries and model-generated summaries. Given two sequences  $X$  and  $Y$  (where  $X$  is a reference summary sentence, and  $Y$  is a model-generated summary sentence), the longest common subsequence (LCS) of  $X$  and  $Y$  is defined as a common subsequence with maximum length. To estimate the similarity between two summaries  $X$  of length  $m$  and  $Y$  of length  $n$ , ROUGE-L ( $F_{lcs}$ ) is defined as LCS-based F-measure according to the following equations:

$$R_{lcs} = \frac{LCS(X, Y)}{m}$$
$$P_{lcs} = \frac{LCS(X, Y)}{n}$$
$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

Summaries must be tokenized for the evaluation of the alignment in calculating ROUGE metrics. BERT's tokenizer was used to perform tokenization in this study.

### **Appendix 3**

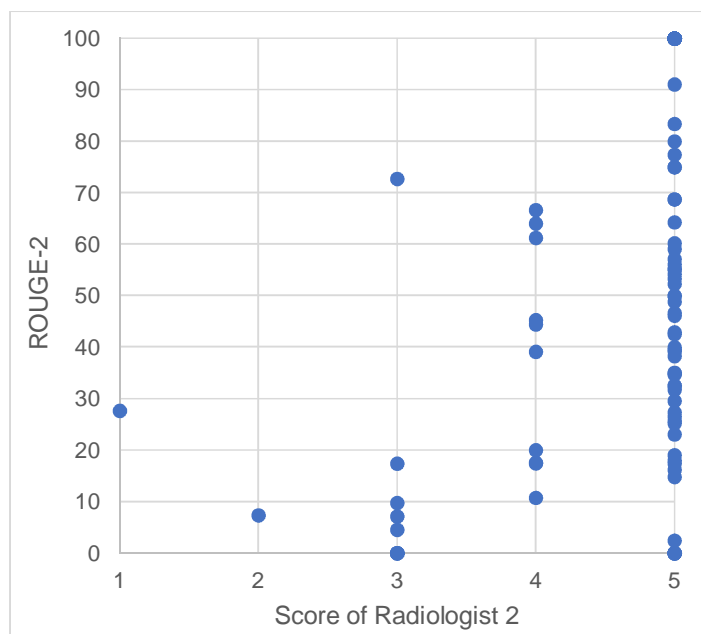
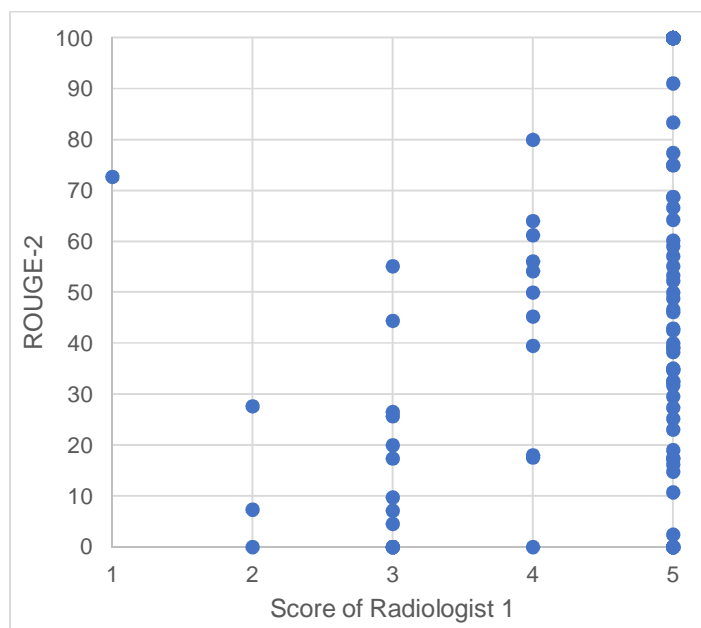
Validation loss of MIMIC-CXR and JMID in the fine-tuning of the T5 models.

Dataset	T5 model	Validation loss
MIMIC-CXR	t5-base, B=2	1.0537
MIMIC-CXR	t5-base, B=8	1.1476
MIMIC-CXR	google/mt5-base, B=2	1.0431
MIMIC-CXR	google/mt5-base, B=8	1.1693
JMID	megagonlabs/t5-base-japanese-web, B=2	0.9893
JMID	megagonlabs/t5-base-japanese-web, B=8	1.098
JMID	google/mt5-base, B=2	0.9422
JMID	google/mt5-base, B=8	1.0344

Note: Validation sets of MIMIC-CXR and JMID comprised 6402 and 7654 reports, respectively.

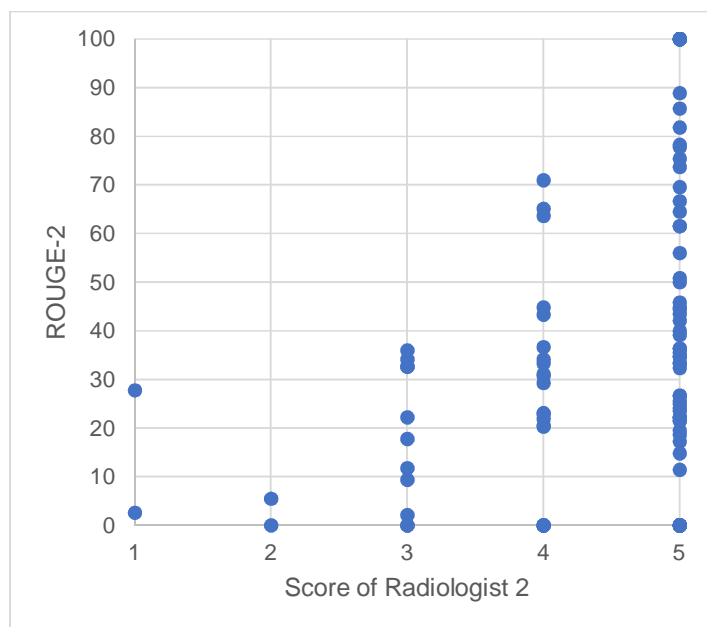
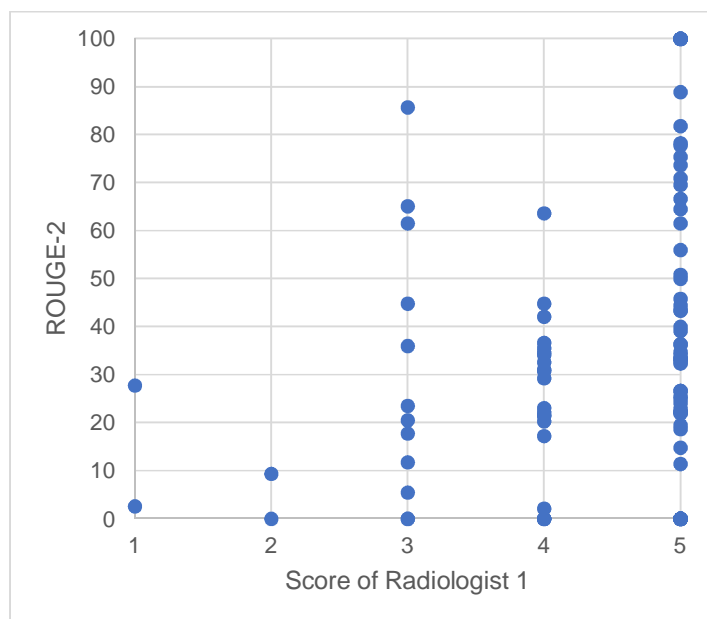
Abbreviations: T5, Text-to-Text Transfer Transformer; B, batch size.

## Appendix 4



Scatter plots of the ROUGE-2 values and radiologist' semi-quantitative scores in MIMIC-CXR.

## Appendix 5



Scatter plots of the ROUGE-2 values and radiologist' semi-quantitative scores in JMID.