Radiology-Aware Model-Based Evaluation Metric for Report Generation

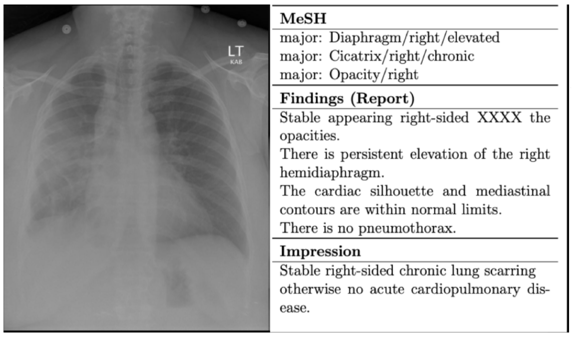
Anonymous EMNLP submission

Abstract

We propose a new automated evaluation metric for machine-generated radiology reports using the successful COMET architecture adapted for the radiology domain. We train and publish four medically-oriented model checkpoints, including one trained on RadGraph, a radiology knowledge graph. Our results show that our metric correlates moderately to high with established metrics such as BERTscore, BLEU, and CheXbert scores. Furthermore, we demonstrate that one of our checkpoints exhibits a high correlation with human judgment, as assessed using the publicly available annotations of six board-certified radiologists, using a set of 200 reports. We also performed our analysis gathering annotations with two radiologists on a collection of 100 reports. The results indicate the potential effectiveness of our method as a radiology-specific evaluation metric.[1]

1 Introduction

Evaluation metrics are essential to assess the performance of Natural Language Generation (NLG) systems. Although traditional metrics are widely used due to their simplicity, they have limitations in their correlation with human judgments, leading to the need for newer evaluation metrics (Blagec et al., 2022; Sai et al., 2022; Novikova et al., 2017). However, newer metrics have not been widely adopted in the literature due to poor explainability and lack of benchmarking (Leiter et al., 2022). In the medical image report generation domain, several new metrics have been developed, including medical abnormality terminology detection (Li et al., 2018), MeSH accuracy (Huang et al., 2019), medical image report quality index (Zhang et al., 2020b), and anatomical relevance score (Alsharid et al., 2019). These metrics aim to establish more relevant evaluation measures than traditional metrics such as BLEU. However, despite their existence, newer publications still rely on traditional metrics, leading to less meaningful evaluations of specialized tasks (Messina et al., 2022).



MeSH
major: Diaphragm/right/elevated
major: Cicatrix/right/chronic
major: Opacity/right

Findings (Report)
Stable appearing right-sided XXXX the opacities.
There is persistent elevation of the right hemidiaphragm.
The cardiac silhouette and mediastinal contours are within normal limits.
There is no pneumothorax.

Impression
Stable right-sided chronic lung scarring otherwise no acute cardiopulmonary disease.

Figure 1: An example report showing the two images and the MeSH, findings and impression columns. Image constructed by the authors with the data from Demner-Fushman et al. (2016).

Radiology reports are narratives that should accurately reflect important properties of the entities depicted in the scan. These reports consist of multiple sentences, including the position and severity of abnormalities and concluding remarks summarizing the most prominent observations (see Figure 1 for an example report). The task of generating radiological reports is challenging due to their unique characteristics and the need for accurate clinical descriptions (Langlotz, 2015). However, current metrics like BLEU do not capture these specific properties, highlighting the need for domain-specific metrics that consider the unique requirements of radiology reports (Chen et al., 2020). At a high level of abstraction, we attempt to answer the following main research questions in this work: (1) Can an existing successful metric model architecture be adapted and optimized to develop a novel radiology-specific metric for evaluating the quality and accuracy of automatically generated radiology reports? and (2) To what extent does the integration of radiology-aware knowledge, impact the precision and dependability of the assessment

[1] The code, data, and model checkpoints to reproduce our findings will be publicly available.

metric in evaluating the efficacy and accuracy of automatically generated radiology reports?

To this end, we suggest an automated measurement for assessing radiology report generation models. It aims to enhance existing metrics designed for different domains, including both automated metrics like COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020) and traditional metrics like SPIDEr (Semantic Propositional Image Description Evaluation) (Liu et al., 2017) or BLEU (Papineni et al., 2002). This improvement involves incorporating a radiology-specific knowledge graph known as RadGraph (Jain et al., 2021). Our contributions are as follows: (i) We design an evaluation model (RadEval) tailored explicitly for assessing radiology reports generated by generative models. By incorporating domain-specific knowledge from RadGraph, we aim to enhance the accuracy and relevance of the assessment., (ii) We evaluate the proposed strategy by applying it to a set of radiology reports generated by two models. We use the IU X-Ray dataset of ground truth radiology reports and compare the automated scores obtained using our framework with the scores of other established metrics. and (iii) We perform an error analysis study with radiology experts that examine the discrepancies between the generated and the ground truth reports. This analysis allows us to further identify the quality of our metric compared with human judgment.

2 Metric Architecture

We use COMET, an evaluation architecture framework developed for machine translation scoring by Unbabel AI (2020); Rei et al. (2020) and train our own metric on radiology data, focusing on the technicalities of radiology reports as outlined before. COMET offers training different types of architectures: Estimator models and Translation Ranking models. The fundamental difference between them is the training objective. While the Estimator is trained to regress directly on a quality score, the Translation Ranking model is trained to minimize the distance between a "better" hypothesis and both its corresponding reference and its original source. We use the referenceless mode of the Estimator model as our input data consists of only two inputs - one ground truth report (the source) and one model-generated report (the hypothesis). The source s, and hypothesis h are independently encoded using a pretrained language encoder (here: XLM-R by Conneau et al. (2020)). The resulting embedding vectors are then passed through a pooling layer to create a sentence embedding for each input. Given a sentence embedding for the hypothesis $\hat{h}$, and the source $\hat{s}$, it extracts the following combined features: (i) Element-wise source product: $\hat{h} * \hat{s}$, and (ii) Absolute element-wise source difference: $|\hat{h} - \hat{s}|$. These combined features are then concatenated to the source embedding $\hat{s}$ and hypothesis embedding $\hat{h}$ into a single vector $[\hat{h}; \hat{s}; \hat{h} * \hat{s}; |\hat{h} - \hat{s}|]$ that serves as input to a feed-forward regressor. Figure 2 depicts the COMET Estimator model architecture. The entire model is trained by minimizing the Mean Squared Error (MSE) between the predicted scores and quality assessments (target values) as a loss function.

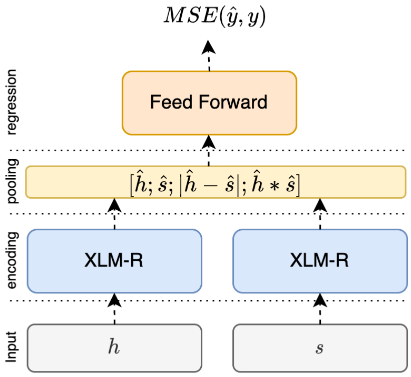


Figure 2: Model architecture for the referenceless metric in the COMET Estimator model (Image provided by Unbabel AI (2020)). The source s, and hypothesis h are independently encoded using a pretrained language encoder. The resulting embedding vectors are then passed through a pooling layer to create a sentence embedding for each input as $\hat{h}$ and $\hat{s}$. Finally, the resulting sentence embeddings are combined and concatenated into one single vector that is passed to a feed-forward regressor. The entire model is trained by minimizing the Mean Squared Error (Rei et al., 2020).

3 Dataset Curation

Because the COMET architecture is built for assessing the quality of machine translation it requires a parallel corpus of source (i.e. the original text), hypothesis (i.e. the machine translation), and reference (i.e. the correct translation of the source) as input to train the model. In the radiology domain, this corresponds to the source being the ground truth report and the hypothesis being the model-generated one. We do not have the notion of a *correct* version of a generated report in our case and are therefore using the referenceless architecture. To ensure the reliability of our model, we require a sufficiently large number of reports for training. We construct the training data for our metric, by creating a corpus of similar reports using the IU X-Ray report collection (Demner-Fushman et al., 2016), a widely utilized dataset within the radiology domain. The IU X-Ray dataset contains chest X-Ray images, along with accompanying reports of actual findings, brief summaries of these findings (referred to as the *impression*), and assigned Medical Subject Headings (MeSH) labels. MeSH is a controlled vocabulary used by the National Library of Medicine database to index and organize biomedical information (National Library of Medicine, 2023). These terms are used to categorize medical articles based on their content and encompass a broad range of medical topics, including anatomy, diseases, drugs, and procedures.

We concatenated major MeSH labels and removed irrelevant MeSH values (i.e. "no indexing" and "technical quality of image unsatisfactory") for each report and performed K-Means clustering on the MeSH terms to group reports containing similar topics. We achieved the best results with 6 clusters. This clustering process allowed us to then take the cross-product of each cluster individually to create the report pairs. Figure 5 shows the final clusters. The most prominent values for each cluster can be seen in Appendix B and the scores to determine the number of clusters in Appendix D.

Next, we scored the similarity of the reports in relation to all other reports in the same cluster using the RadCliQ Metric (Yu et al., 2023a), which is a novel evaluation measure for the similarity of clinical reports leveraging a combination of the BLEU-2 score and the RadGraph F1 metric. The latter ”computes the overlap in clinical entities and relations that RadGraph extracts from a machine- and human-generated reports” (Yu et al., 2023a, p.4).

We then generate two sets of comparative report pairs. The first one (referred to as *Best Match corpus* by selecting the top-scored (i.e. most similar) match for each report (based on RadCliQ metric), resulting in a set that encompasses all reports of the cleaned IU X-Ray dataset at least once (i.e. the set size is equal to the size of the cleaned IU X-Ray dataset and each report in the dataset has one corresponding report, which matches best in terms of similarity). The secondary one (referred to as *Top 10% corpus* allowed for multiple instances of single reports in the set if they had multiple best matches.

After having created two sets we divided them into two distinct subsets, a training set and a test set. We created a random split of 80/20 using to extract 20% of the data into the test set and keep the remaining 80% as the training set. This ensured that our model can be trained and evaluated on two distinct sets of data. With the training process in mind, we also split the training data set further into two subsets, the primary training set, and the validation subset, using the same 80/20 split to have the validation data out of the training set. This validation set is provided to the model trainer to fine-tune its hyper-parameters on each epoch.

During this process, we ensured an appropriate share of normal and abnormal reports are included in both train/validation/test datasets and to not bias the data towards normal reports too much (see Table 6).

4 Model training

During model training, we optimized the Kendall Tau value between predicted and ground truth rankings. Increasing the maximum number of training epochs from 20 to 40 resulted in higher Kendall Tau values. We also compared the performance of using BioClinical BERT (Alsentzer et al., 2019) instead of XLM-R and training on the RadCliQ Score versus the RadGraph F1 score for the Top 10% corpus.

Our motivation for providing comparative report pairs is to assist future researchers in training their own metrics using a ’Source - Hypothesis’ model architecture in their research. To ensure the quality of our corpus, we have compared the exact overlap on MeSH labels among source and reference reports (i.e. the number of overlapping tokens). Our analysis of the Top 10% corpus revealed that 80.2% of the rows had overlap in their MeSH labels, with 46.9% having one token overlapping and 33.3% having more than one. Only 19.83% of rows had no exact overlaps in MeSH tokens. Similarly, when we examined the extent of overlap between MeSH labels in the complete corpus (i.e. among all scores), we found that 34.20% of rows had no overlap between their MeSH labels. In contrast, 31.69% of rows had only one overlap, and 34.11% had more than one overlap between their MeSH labels. We, therefore, see that the scores in the Top

Checkpoint Name	**Encoder**	**Report pairs**	**Training Target Value**	**Max**($Kendall_\tau$)
Match XLM-R RadCliQ	XLM-R	Best Match	RadCliQ-score	0.696 (Epoch 3)
Match Clinic RadCliQ	BioClinical BERT	Best Match	RadCliQ-score	0.714 (Epoch 10)
Top Clinic RadCliQ	BioClinical BERT	Top 10%	RadCliQ-score	0.830 (Epoch 24)
Top Clinic RadGraph	BioClinical BERT	Top 10%	RadGraph F1-score	0.714 (Epoch 18)

Table 1: The specifications of final model checkpoints: **Match XLM-R RadCliQ:** Based on the Best Match corpus, with XLM-R as the encoder layer and RadCliQ as the quality assessments (target values). **Match Clinic RadCliQ:** Based on the Best Match corpus, with BioClinical BERT as the encoder layer and RadCliQ as the quality assessments (target values). **Top Clinic RadCliQ:** Based on the Top 10% corpus, with BioClinical BERT as the encoder layer and RadCliQ as ts the quality assessments (target values). **Top Clinic RadGraph:** Based on the Top 10% corpus, with BioClinical BERT as the encoder layer and RadGraph F1 as the quality assessments (target values). Max ($Kendall_\tau$) is evaluated on the Validation set.

10% corpus reflect the contents of the reports well.

5 Final model checkpoints

During our experiments with different clustering and similarity score methods, we have generated many comparative report pairs and also already trained several models to benchmark their performance. Out of all models, we have decided to focus on a couple of best-performing checkpoints (based on the highest Kendall τ value while training). We used our two corpora (Best Match and Top 10%) and combined them each once with the XLM-R encoder layer and once with the medical-specific BioClinical BERT (Alsentzer et al., 2019). Also, we trained the models on two scores: Once on the plain Radgraph F1 score, and once on the combined RadCliQ metric score to compare how they differ in correlation performance.

It is important to notice, that the RadCliQ score is a measure of how many errors a report will contain (i.e. lower is better) and RadGraph F1 is a measure of graph similarity (i.e. higher is better, Yu et al. 2023a). Our model checkpoints will behave accordingly when giving their predicted scores. For all checkpoints the scores are unbounded but we provide the typical range. The names of our checkpoints are based on the type of corpus (best match or Top 10%), the encoder (XLM-R or BioClinical BERT), and the type of score they output (RadCliQ or RadGraph F1).

We trained the following checkpoints (see also Table 1): (i) **Match XLM-R RadCliQ:** Based on the Best Match corpus, with XLM-R as the encoder layer and RadCliQ as the quality assessments (target values). A lower score indicates a better report. Scores typically fell within -3.5 and +0.5 in our tests., (ii) **Match Clinic RadCliQ:** Based on the Best Match corpus, with BioClinical BERT as the encoder layer and RadCliQ as the quality assessments (target values). A lower score indicates a better report. Scores typically fell within -3.5 and +0.5 in our tests., (iii) **Top Clinic RadCliQ:**Based on the Top 10% corpus, with BioClinical BERT as the encoder layer and RadCliQ as ts the quality assessments (target values). A lower score indicates a better report. Scores typically fell within -3.0 and +1.5 in our tests., and (iv) **Top Clinic RadGraph:** Based on the Top 10% corpus, with BioClinical BERT as the encoder layer and RadGraph F1 as the quality assessments (target values). A higher score indicates a better report. Scores typically fell within -0.2 and +1.5 in our test.

6 Model performance

We assessed the performance of our model's metric using our test dataset (see Section 3) and the IU X-Ray dataset's test set (i.e. 590 sets containing the ground truth and generated reports by two state-of-the-art radiology report generation methods[2]: R2Gen (Chen et al., 2020) and M2Tr (Cornia et al., 2020)). To provide a comprehensive comparison, we calculated the performance of each radiology generation method using five metrics. These involve BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020a), CheXbert Similarity (Smit et al., 2020), RadGraph F1 and RadCliQ (Yu et al., 2023a).

BLEU and BERTScore have commonly used metrics in natural language generation tasks to assess the similarity between machine-generated and

[2] We used the following implementations: M2Tr: https://github.com/ysmiura/ifcc and R2Gen: https://github.com/cuhksz-nlp/R2Gen

Model	**BLEU-4**	**BLEU-2**	**BERTscore**	**CheXbert**	**RadGraph F1**	**RadCliQ**
			Our Top 10% test data set			
Match XLM-R RadCliQ	-	**86.26%**	66.98%	27.75%	*71.38%*	95.37%
Match Clinic RadCliQ	-	**87.99%**	67.80%	27.80%	*71.05%*	96.52%
Top Clinic RadCliQ	-	**88.76%**	67.03%	27.45%	*67.22%*	95.51%
Top Clinic RadGraph	-	41.35%	*48.86%*	24.45%	87.92%	**67.57%**
			R2Gen reports			
Match XLM-R RadCliQ	78.08%	**86.85%**	*79.54%*	52.69%	24.74%	66.37%
Match Clinic RadCliQ	81.84%	**88.94%**	*80.95%*	51.95%	19.36%	63.03%
Top Clinic RadCliQ	77.17%	**85.81%**	*76.63%*	47.36%	14.52%	58.00%
Top Clinic RadGraph	**61.37%**	*66.33%*	65.09%	39.09%	5.17%	40.96%
			M2Tr reports			
Match XLM-R RadCliQ	74.71%	*84.88%*	76.58%	47.66%	**85.90%**	95.28%
Match Clinic RadCliQ	79.72%	**87.60%**	*79.83%*	45.54%	71.73%	87.70%
Top Clinic RadCliQ	73.50%	**83.51%**	*74.55%*	43.90%	60.46%	78.58%
Top Clinic RadGraph	58.12%	*64.29%*	64.01%	33.64%	65.60%	**71.76%**

Table 2: Spearman rank correlation between the RadEval score of our model checkpoints and the other metrics based on the generated reports by M2Tr and R2Gen. The **highest correlation is marked in bold** and *the second highest in italics*. The score on which the specific model checkpoint was trained is printed in light grey.

human-generated texts. BLEU measures the overlap of n-grams and is representative of text overlap-based metrics. On the other hand, BERTScore captures contextual similarity beyond exact textual matches. It uses a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to encode the two pieces of text and measure their similarity based on their contextualized embeddings.

CheXbert vector similarity and RadGraph F1 are metrics specifically designed to evaluate the accuracy of clinical information. CheXbert vector similarity calculates the cosine similarity between the indicator vectors of 14 pathologies extracted from machine-generated and human-generated radiology reports using the CheXbert automatic labeler. This metric focuses on evaluating radiology-specific information but is limited to pathologies. To address this limitation, Yu et al. (2023a) propose the utilization of the report's knowledge graph to represent a wide range of radiology-specific information. Introducing a novel metric called RadGraph F1, they measure the overlap in clinical entities and relations extracted by RadGraph from both machine-generated and human-generated reports.

RadCliQ is a combined metric introduced by Yu et al. (2023a), which combines the BLEU and RadGraph F1 metrics through a linear regression model. The purpose is to estimate the total number of errors that radiologists would assign to a generated report. This metric requires the BLEU and RadGraph F1 scores computed for the generated report as input. According to Yu et al. (2023a), RadGraph F1 is the most comparable metric to human judgment, followed by BERTScore, BLEU-2, and CheXbert. We evaluated our model checkpoints trained on RadCliQ and RadGraph F1 to explore their performance difference. We performed the inference using the model checkpoints to obtain the predicted "RadEval" scores. We then calculated the Spearman correlation value between our RadEval Score and the other metrics' scores for the different checkpoints.

In the test dataset we constructed (Top 10%), the data in Table 2 demonstrates that all RadCliQ-trained models (Match Clinic RadCliQ, Match XLM-R RadCliQ, and Top Clinic RadCliQ) exhibited a high correlation of over 85% with the BLEU-2 score, which was according to our anticipation as described above. Additionally, these model checkpoints showed the second-highest correlation of approximately 69% with the RadGraph F1 score, which was also in line with our initial expectations.

Interestingly, we found that our RadCliQ-trained models also displayed a reasonably high correlation of approximately 67% with the BERTscore metric. The RadGraph F1-trained checkpoint (Top Clinic RadGraph) on the other hand showed the highest correlation with the RadCliQ score at 67.57% and the second highest correlation with BERTscore at 48.86%, with BLEU-2 following at 41.35%. It is worth noting that none of our model checkpoints exhibited a high correlation with the CheXbert score, with correlations ranging between 24% and 28%. Even though the correlation with BLEU-2 for the RadGraph F1-trained checkpoint was much lower compared to the RadCliQ-trained checkpoints (-45 percentage points), the RadGraph F1-trained checkpoint also showed a lower correlation with BERTscore (-19 percentage points) and CheXbert score (-3 percentage points) at the same time, albeit less drastic than the drop in BLEU-2 correlation.

When we analyzed the correlation scores of our model on the two model-generated datasets. we observed different correlation patterns than the report pairs test dataset. The correlation values between our model and both BLEU scores were high, ranging from 73% to 86% on both R2Gen and M2Tr reports for RadCliQ-trained checkpoints. However, for the RadGraph F1-trained checkpoint, the correlation was low, ranging from 14% to 25% for R2Gen and 60% to 85% for M2Tr. R2Gen had the lowest correlation (5.17%) with RadGraph F1. The correlation with BERTscore and CheXbert scores was generally higher than the parallel corpus test dataset, ranging from 33% to 53% and 64% to 80%, respectively.

We found that the RadGraph F1-trained checkpoint for both generation models had better correlation values than the other RadCliQ-trained model checkpoints with BLEU-2 and BERTscore, being at most 19 percentage points away from the highest value for BLEU-2 and at most 11 percentage points for BERTscore. The maximal drop for the CheXbert score was 10 percentage points, compared to 3 percentage points for our corpus test dataset.

7 Automated metric/radiologist alignment

7.1 Alignment analysis with ReXVal dataset

In our first experimental alignment study, we make use of the Radiology Report Expert Evaluation (ReXVal) Dataset (Yu et al., 2023b) [3]. The ReXVal Dataset is a collection of assessments made by radiologists regarding errors found in automatically generated radiology reports. This dataset includes evaluations from six board certified radiologists. The assessments cover clinically significant and clinically insignificant errors, categorized into six different error types. The reports being evaluated are compared to ground-truth reports from the MIMIC-CXR dataset (Johnson et al., 2019). Each of the 50 studies in the dataset contains one ground-truth report and four reasonably accurate generated reports by selecting candidate reports that score highly according to each of four automated metrics (i.e. BLEU, BERTscore, CheXbert and RadGraph F1), referred to as oracle-metric reports, resulting in 200 pairs of candidate and ground-truth reports that radiologists have annotated.

We utilized this dataset to assess the correlation between our proposed metric and radiologists' evaluations. To do so, we employed the approach proposed by the authors to calculate the mean values of significant, insignificant, and total errors for each oracle report, considering the input from their six annotators. Then, we compute RadEval and RadCliQ scores for each metric-oracle report and determine the level of alignment between the radiologists and the metrics (i.e.RadEval and RadCliQ) using the Spearman rank correlation coefficient. The results (Figure 3) demonstrate that our proposed metrics perform better than the RadCliQ metric compared on all oracle reports other than BLEU. Our RadGraph **Top Clinic RadGraph** checkpoint surpasses RadCliQ in terms of human correlation up to 10 percentage points (in the BERTscore oracle reports). Also our other checkpoint **Match Clinic RadCliQ** surpasses the RadCliQ Metric by up to 5 percentage points.

To fairly compare and analyze whether the improvements in the human study are statistically significant, we performed a significance test using CoCor (Diedenhofen and Musch, 2015).

Following the CoCor method, we define the following groups in which the groups are dependent and overlap. (i) JK (Correlation RadCliQ - Human),, (ii) JH (Correlation RadEval - Human), and (iii) KH (Correlation RadEval - RadCliQ). We set `Alpha = 5%`, `Confidence Level = 95%`, `Null-Value = 0` and `Sample size = 50 samples`

[3] The Dataset is available on PhysioNet (Goldberger et al., 2000) at https://physionet.org/content/rexval-dataset/1.0.0/

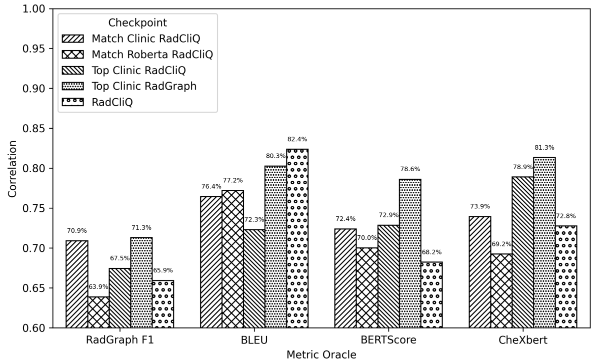


Figure 3: Spearman rank correlation between the RadCliQ (Yu et al., 2023a) and our RadEval model checkpoints, and the human error scores assigned by radiologists in four oracle-metric reports datasets.

Dataset name	Correlation		p-value
	Human-RadCliQ(%)	Human-RadEval(%)	
BertScore-Oracle set	68.2	78.6	0.0084
CheXbert-Oracle set	72.8	81.3	0.0131

Table 3: Statistically significant test using CoCor (Diedenhofen and Musch, 2015) on the ReXVal dataset. Showing that the high correlation measured by our Top Clinic RadGraph Checkpoint compared to the original RadCliQ metric is statistically significant.

per. We have collected the p-values for the dependent, overlapping model according to Hendrickson et al. (1970) and report the results for the following set of datasets from ReXVal; where the predictions (generated reports) have been selected based on one of the following Oracle metrics: (1) **BertScore-Oracle**: The null hypothesis can be rejected with p–values 0.0084, and (2) **CheXbert-Oracle**: The null hypothesis can be rejected with p-values 0.0131. Confirming the alternative hypothesis: r.jk is less than r.jh (one-sided) with r.jk being the correlation of RadCliQ with the human and r.jh being the correlation of RadEval (ours) with the human.

In these two datasets (50 examples each), the correlations are shown in Table 3. The results show that both RadEval and RadCliQ have a strong correlation with human judgments (i.e., > 60%) on the ReXval dataset.

7.2 Alignment analysis with internal dataset

To further investigate the alignment of the automated evaluation metrics with radiologists, we created a balanced dataset of 100 reports for human annotation from an initial set of 590 reports generated using M2Tr (Cornia et al., 2020; Nooralahzadeh et al., 2021). The dataset balance was achieved by categorizing the reports into low, average, and high groups based on the 0.33 quantiles of the RadCliQ metric score. Random sampling was then performed to select 150 reports from each category. The reports were further filtered to separate normal and abnormal categories, excluding those labeled as normal in the 'mesh-0' column and removing reports with empty 'mesh-1' values. The remaining abnormal reports were then filtered based on the 'IMPRESSION' column, removing those containing specific phrases associated with normal reports [4]. The resulting dataset comprised 80 abnormal and 20 normal reports.

In this regard, we are inspired by the work of Yu et al. (2023a), in which the authors asked a radiologist to count the number of clinically significant and insignificant errors observed in the predicted report for each prediction pair and the ground truth and categorize them into one of the following categories (Yu et al., 2023a, p.4-5) (the categories with † are added by us). (1) False prediction of finding, (2) Omission of finding, (3) Incorrect location/position of finding, (4) Incorrect severity of the finding, (5) Mention of comparison that is not present in the reference impression, (6) Omission of comparison describing a change from a previous study, (7) † Mention of uncertainty that is not present in the reference, and (8) † Omission of uncertainty that is present in the reference.

To accomplish the study, initially, two board certified radiologists independently identified and extracted the positive findings from the ground truth reports. The positive findings were then classified into significant and insignificant ones. A comparison was made between the findings extracted from the ground truth reports and the generated reports. Using the eight predefined error categories, the number of errors for each category was counted on the basis of the results of the comparison. Ultimately, both radiologists engaged in discussions with each other and reached a consensus for each report. After receiving the evaluations of two annotators, we evaluated the level of alignment between our metric and their evaluations employing the Spearman rank correlation coefficient. This allows us to quantify the relationship between the metric scores and the count of errors identified by radiologists in the reports. We establish the align-

[4] i.e., variations of the phrases *no acute cardiopulmonary abnormalities*, *no evidence of active disease*, *no acute findings*. The complete list of filters can be found in Appendix A.

ment between the metric and radiologists' evaluations for our checkpoints by conducting this analysis on a selected set of 100 studies. We examine the total number of errors and specifically focus on the number of errors that are clinically significant, as indicated by the radiologists' annotations. In this analysis, Table 4, we found that the correlation for the RadCliQ model is 33.49% for total errors and 19.29% for significant mistakes. It shows a slightly higher positive correlation than our two models, indicating a more substantial alignment between the model's predictions and the human-annotated errors in our dataset. The correlation between Match XLM-R RadCliQ and human annotation is 28.71% for total errors and 18.37% for significant errors. These values suggest a moderate positive correlation between the model's predictions and the human-annotated total and significant errors. However, it performed up to 11% better than the compared metrics (BLEU and RadGraph F1) on the total sum of errors and up to 4% for the significant errors.

To analyze the quality of our metric, we looked at reports with a higher occurrence of errors (referred to as reports with "noisy generation", where our annotators have identified more than 3 errors in total). There are 30 such reports among our 100 studies[5]. When looking only at the noisy reports we can see for the **Match Roberta RadCliQ** and **Top Clinic RadCliQ** checkpoints, that we outperform the comparison metrics by up to 19% on the sum of errors and up to 6% for the significant errors. For this set of reports, we even perform better than RadCliQ in both categories.

Model	**Human Annotation Correlation**	
	#total errors (%)	#sig. errors (%)
Complete Human Annotation Dataset: `100 Examples`		
BLEU	17.70	13.85
RadGraph F1	28.44	16.33
RadCliQ	**33.49**	**19.29**
Match XLM-R RadCliQ	28.71	18.37
Noisy generation (> 3 errors in the prediction) : `30 Examples`		
BLEU	27.36	1.39
RadGraph F1	19.10	0.91
RadCliQ	33.80	1.69
Match XLM-R RadCliQ	34.48	6.42
Top Clinic RadCliQ	**37.35**	**7.97**

Table 4: Spearman rank correlation between the RadEval score of our two best performing model checkpoints and the human error scores assigned by radiologists. As a comparison, we include BLEU and two recent radiology-specific metrics and report their correlation scores with our annotators. Correlations for RadGraph F1 are multiplied with -1 as these scores estimate the report quality (i.e., higher is better), and the human annotators provide the error score (i.e., lower is better).

8 Conclusion

Our work focuses on developing a novel evaluation metric to evaluate the quality and precision of automatically generated radiology reports. We propose an evaluation model called RadEval that incorporates domain-specific knowledge from a radiology-aware knowledge graph. We train the RadEval model using two corpora, the Best Match corpus and the Top 10% corpus, which contain pairs of ground truth reports that are similar in terms of their RadGraph representation. We evaluate the performance of the RadEval model on a test set and compare it to other established metrics such as BLEU, BERTScore, CheXbert, RadGraph F1, and RadCliQ. We find that the RadEval model performs well and correlates highly with these metrics. Additionally, when using the new ReXVal dataset of human annotations to compare our alignment with human judgment, we find a high correlation that even surpasses RadCliQ for most report pairs. When conducting our own human annotation study, we did not find a direct high correlation with our human annotators. Still, when comparing with the other metrics' agreement with the same human scores, we also performed better in some cases. Furthermore, it should be noted that although we have demonstrated relatively strong correlations between automated evaluation metrics and human judgment, additional research is still required to develop an appropriate evaluation metric that aligns with radiologists' expectations and has clinical validity.

9 Limitations and Ethical Considerations

Our proposed method has certain limitations and ethical considerations that merit discussion. One limitation of our study is that different radiologists evaluating the reports often gave different scores, even though the effort was to make the evaluation scheme objective and consistent. This variability

[5] We provided the statistics of the error categories in these 30 examples Appendix E and two noisy examples of what our dataset looks like and the corresponding error categories in Appendix F

among radiologists is a common issue when using subjective ratings from clinicians. It suggests that our evaluation scheme may have limitations and it might be challenging to evaluate radiology reports objectively. Another limitation is that we only considered a specific set of metrics in our study. There are other metrics available that could behave differently than the ones we examined. This means that there could be additional metrics that might provide different insights into evaluating radiology reports.

Regarding the datasets used in our study, we exclusively utilized publicly available datasets that are properly anonymized and de-identified, addressing privacy concerns. However, it is crucial to emphasize that if datasets containing comparison exams become available in the future, additional precautions must be taken to ensure that no personally identifiable information is inadvertently disclosed or used in a manner that could identify individual patients The public MIMIC-CXR and IU-X-ray datasets are employed in this work, in which all protected health information was de-identified. De-identification was performed in compliance with Health Insurance Portability and Accountability Act (HIPAA) standards in order to facilitate public access to the datasets. Deletion of protected health information (PHI) from structured data sources (e.g., data fields that provide patient name or date of birth) was straightforward. All necessary patient/participant consent has been obtained, and the appropriate institutional forms have been archived. We used the datasets for RadGraph and ReXVal, which are under the PhysioNet license. Therefore, as required, we will release our code and data to PhysioNet.

By acknowledging these limitations and ethical considerations, we aim to encourage future research and discussions in the field, driving advancements in radiology report generation while prioritizing patient privacy, accuracy, and fairness.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mohammad Alsharid, Harshita Sharma, Lior Drukker, Pierre Chatelain, Aris T. Papageorghiou, and J. Alison Noble. 2019. Captioning ultrasound images automatically. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 338–346, Cham. Springer International Publishing.

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. 2022. A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52–63, Dublin, Ireland. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Birk Diedenhofen and Jochen Musch. 2015. cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, 10:1–12.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Gerry F Hendrickson, Julian C Stanley, and John R Hills. 1970. Olkin's new formula for significance of r13 vs. r23 compared with hotelling's method. *American Educational Research Journal*, 7(2):189–195.

Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. 2019. Multi-attention and incorporating background information model for chest x-ray image report generation. *IEEE Access*, 7:154808–154817.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis P Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *PhysioNet*.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042.

Curtis P Langlotz. 2015. *Radiology report: a guide to thoughtful communication for radiologists and other medical professionals*. Independent Publishing Platform, San Bernardino, CA.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1537–1547, Red Hook, NY, USA. Curran Associates Inc.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of SPIDEr. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.*, 54(10s).

National Library of Medicine. 2023. Medical subject headings files 2023. [Online; accessed March 10th 2023].

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2).

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Unbabel AI. 2020. Comet: High-quality machine translation evaluation. [Online; accessed December 10th 2022].

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023a. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, page 100802.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, Curtis Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023b. Radiology report expert evaluation (rexval) dataset.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020b. When radiology report generation meets knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12910–12917.

Appendix

A Filter phrases used to get only abnormal reports

While working on our metric model, we made sure that the input data is balanced in terms of abnormal and normal reports using the following filters. In the first step, we removed all reports with a mesh-0 label of "normal". In addition, we employed a set of predefined phrases that indicate normal impressions in the radiology reports. These phrases are: (i) "No acute cardiopulmonary abnormality", (ii) "No acute cardiopulmonary abnormalities", (iii) "Negative for acute abnormality", (iv) "No evidence of active disease", (v) "No acute cardiopulmonary process", (vi) "No acute cardiopulmonary disease", (vii) "No acute cardiopulmonary findings", (viii) "No acute pulmonary findings", (ix) "No acute findings", (x) "No acute cardiopulmonary abnormality identified", (xi) "No acute cardiopulmonary abnormality seen", (xii) "No acute cardiopulmonary abnormality detected", (xiii) "No acute cardiopulmonary finding", (xiv) "No active disease", and (xv) "No acute disease".

B Cluster Terms

	Most prominent term
0	normal
1	lung/hypoinflation
2	granulomatous disease
3	thoracic vertebrae/degenerative/mild
4	calcified granuloma/lung/base/right
5	calcified granuloma/lung/upper lobe/left
	Second most prominent term
0	No value
1	lung/hypoinflation markings/bronchovascular
2	cardiomegaly/mild
3	thoracic vertebrae/degenerative
4	calcified granuloma/lung/base/left
5	calcified granuloma/lung/upper lobe/right

Table 5: The most common and second most common terms for each MeSH cluster by numeric cluster Identifier (ID). The principal component analysis of the clusters can be seen in Figure 5.

C Dataset distribution for report pair generation

Dataset	**# reports total**	**% abnormal**
	Top 10%	
Training	47’162	63.19%
Testing	14’738	62.19%
Validation	11’791	63.21%
	Best Match	
Training	471’689	84.64%
Testing	147’403	84.59%
Validation	117’922	84.68%

Table 6: Size of the different dataset types for the two sets of comparative report pairs and their respective percentage of abnormal reports (i.e. mesh-0 $\neq$ *normal*)

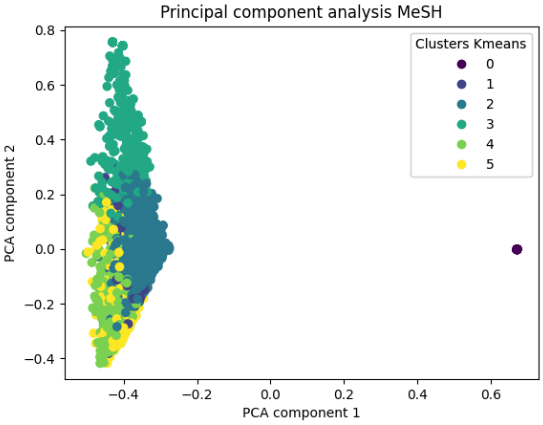


Figure 5: Visualization of the clusters generated by KMeans. The data has been reduced to two dimensions using PCA and the clusters are color-coded. **MeSH** $n = 6$: The outlier (cluster 0) are the normal reports. Clusters 2 and 3 are well defined, 4 and 5 have a lot of overlap. The most prominent values for each cluster can be seen in Appendix B.

D Scores for the MeSH clusters

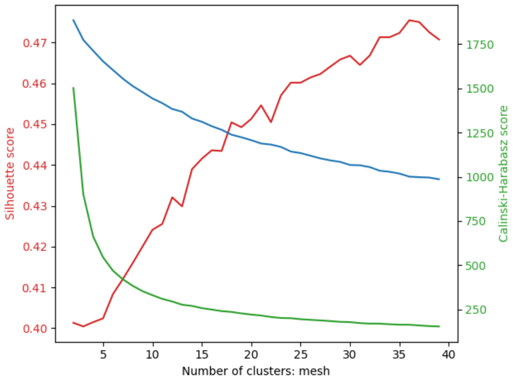


Figure 4: Silhouette score (red), Elbow score/inertia (blue), and Calinski-Harabasz score (green) for increasing amounts of clusters on the MeSH column

E The statistics of the error categories in the 30 noisy examples

Error Type	**Errors**	
	#Significant	#Insignificant (%)
(1) False prediction of finding	10	7
(2) Omission of finding	64	36
(3) Incorrect location/position of finding	1	1
(4) Incorrect severity of the finding	0	1
(5) Mention of comparison	0	6
(6) Omission of comparison	17	7
(7) Mention of uncertainty	1	0
(8) Omission of uncertainty	3	0
Total	96	58

Table 7

F Noisy Report Examples

Example 1.

- **Ground Truth Report**
  the heart is enlarged. there is pulmonary vascular congestion with diffusely increased interstitial and mild patchy airspace opacities. the <unk> xxxx pulmonary edema. there is no pneumothorax or large pleural effusion. there are no acute bony findings.

- **Predcition**
  there is a right upper lobe opacity. cardiomediastinal silhouette is normal. pulmonary vasculature and xxxx are normal. no pneumothorax or large pleural effusion. osseous structures and soft tissues are normal.

Error Type	**Category**	**Instances of failure**	**# Errors**
(2) Omission of Findings	Significant	heart size enlarged, vascular congestion, interstital, airspace opacities, pulmonary edema	5
(3) Incorrect location/position of finding	Insignificant	right upper lobe opacity	1

Example 2.

- **Ground Truth Report**
  stable enlargement of the cardiac stable mediastinal contours. increased interstitial markings in the central lungs and right greater than left. xxxx opacity on the lateral view over the heart also present on the previous exam suggesting chronic subsegmental atelectasis or scarring. no definite pleural effusion seen.

- **Predcition**
  the heart and cardiomediastinal silhouette are normal in size and contour. there is no focal air space pleural or pneumothorax. the osseous structures are intact.

Error Type	**Category**	**Instances of failure**	# Errors
(2) Omission of Findings	Significant	stable enlargement of the cardiac, stable mediastinal contour, increased interstital markings, xxx opacity - chronic atelectasis or scarring	5
(6) Omission of comparison	Significant	increased interstitial markings in central lung and right, on the previous exam ...	3