

Multi-Dimensional Evaluation Metrics for Chest X-Ray Reports

by

Saumya Rawat

S.B Computer Science and Engineering, Massachusetts Institute of
Technology (2021)

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
May 6, 2022

Certified by.....
Peter Szolovits
Professor
Thesis Supervisor

Accepted by
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Multi-Dimensional Evaluation Metrics for Chest X-Ray Reports

by

Saumya Rawat

Submitted to the Department of Electrical Engineering and Computer Science
on May 6, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In the past few years, there has been abundant research in using machine learning to generate high quality radiology reports using the large MIMIC-CXR chest x-ray dataset. However, there has been little work focused on evaluating the quality of generated reports from a clinical perspective, where accuracy is the most important factor. Current evaluation metrics evaluate reports in one dimension. This work proposes the use of multiple dimensions (factual correctness, comprehensiveness, style, and overall quality) to better capture evaluation preferences of a clinical text generating model where preferences can differ based on the use case. This work also presents a dataset of radiologist rating annotations for generated and reference chest x-ray radiology reports. Lastly, it also creates an improved metric for the readability dimension by adding context awareness of frequent and acceptable medical terminology.

Thesis Supervisor: Peter Szolovits

Title: Professor

Contents

1	Introduction	13
1.1	Background and Related Work	15
1.1.1	Radiology Report Generation	15
1.1.2	Evaluation Metrics	15
1.1.3	Readability Metrics	16
2	Experiments	17
3	Radiologist Ranking Task	19
3.1	Methods	19
3.2	Results	22
3.2.1	Evaluating Metrics	22
3.2.2	Evaluating Generated Reports	23
3.2.3	Radiologist Decision Process	24
3.3	Discussion	24
3.3.1	Understanding Radiologist Decisions	24
3.3.2	Improved Metrics	25
3.3.3	Study Improvements	26
3.4	Conclusion	26
4	Radiologist Rating Task — Multi-Dimensional Evaluation	27
4.1	Methods	27
4.1.1	Report Generation Models	28

4.1.2	Selecting Diverse Images	28
4.1.3	Generated Report Captions	29
4.1.4	Multi-Dimensional Analysis	29
4.1.5	Rating Task	31
4.1.6	Evaluation Techniques	31
4.2	Results	33
4.2.1	Radiologist Agreement	33
4.2.2	Evaluating Metrics	34
4.2.3	Evaluating Generated Reports	34
4.2.4	Multi-Dimensional Analysis	36
4.3	Discussion	38
4.3.1	Generation Methods	38
4.3.2	Need for Improved Metrics	38
4.3.3	Improved Study	38
4.4	Conclusion	39
5	Improved Readability Metric	41
5.1	Methods	42
5.1.1	Dataset	42
5.1.2	Designing the Improved Dale-Chall Readability Metric	42
5.1.3	Evaluation Techniques	44
5.2	Results	44
5.2.1	Evaluating Metrics	44
5.2.2	Multi-dimensional Analysis	46
5.3	Discussion	47
5.3.1	Improved Metric Performance	47
5.3.2	Readability Metric Improvements	48
5.4	Conclusion	48
6	Conclusion	51
6.1	Main Contributions	51

6.2 Future Work 52

List of Figures

3-1	Three different annotation tasks we considered for the radiologists to perform.	20
3-2	An example of the chosen annotation task.	21
4-1	Scoring Guidelines provided to radiologists.	30
4-2	Task presented to radiologists, where all captions are shown at the same time.	32
4-3	Overview of radiologist inter-annotator agreement	33
4-4	Overview of different evaluation metrics and their agreement with radiologist overall rating	34
4-5	This graph shows the average overall rating (range of 1 to 7 (best)) radiologists scored for each caption generation method along with one standard deviation error bars.	35
4-6	This graph shows how metrics and the averaged radiologists' dimensional ratings ranked the different types of generative models. Lower bars indicate a better ranking and better performing report.	37
4-7	This heat map shows the Pearson correlation for four dimensions based on the average of the three radiologists' ratings for each dimension.	37
5-1	Average Dale-Chall readability scores for each report caption generation method where lower values represent the metric scoring the report as more readable.	45

5-2 This heat map shows the Pearson correlation between the radiologist's rated dimensions, standard Dale-Chall readability metric, and the Improved Dale-Chall readability metric. Note that for the four rated dimensions (Factual Correctness, Comprehensiveness, Presentation/Style, and Overall Rating) a higher score means a better report. However, for the Standard Dale-Chall and Improved Dale-Chall a higher score means a harder to read or worse report. 46

List of Tables

3.1	Of the 199 consensus comparisons, how often would each metric rank the two reports the same way the radiologists did?	22
3.2	Percent of instances that annotators ranked each method a given ranking.	23
3.3	Breakdown of how often one method was ranked higher than another method in the 199 consensus pairs.	24
3.4	Top n-grams from the explanations provided by annotators for decision-making. Phrases containing stop words were removed.	24

Chapter 1

Introduction

There is a strong, growing interest in applying machine learning towards clinical applications to improve radiologist efficiency and accuracy as well as provide additional support in underserved communities. Deep learning models have been trained, fine-tuned, and applied to evaluate radiology x-ray images to generate radiology reports [8, 18]. To further improve the quality of these reports, it is important to evaluate them in a meaningful manner. However, current evaluation metrics are standard NLP metrics (BLEU and CIDEr) and unsuitable for generated clinical text [22, 2, 6, 3, 13] as they focus on the fluency of the language more than the accuracy of the content.

In the clinical domain, radiologists believe clinical correctness and accuracy is the most important factor [4] to consider when evaluating the quality of a chest x-ray report. It's more important for a radiology report to state correct information than it is for the report to stylistically sound good. Likewise, it's more important for every detail stated to be accurate and to include all important details than it is to include many details where some may be inaccurate. Given this importance of clinical information accuracy, there has been some work extracting clinical information from reports [10, 19, 24] to measure clinical accuracy [18, 4]. However, these works only identify one axis of report quality and there is still more work to be done with respect to other axes and the combination of quality factors to produce a gold standard evaluation metric for clinical text.

Current standard NLP metrics are focused on one dimension and the overall quality of a candidate generated text when compared to a reference text. However, there are many ways candidate texts can differ from reference texts, and these modes of difference are not captured in the metric and are instead combined into an overall score. By grouping various factors radiologists evaluate with into one overall score, we lose specificity and information that could otherwise help us identify specific areas of improvement for candidate texts. In particular, text can differ in terms of accuracy, comprehensiveness, style, understanding, and more. However, current metrics have a strong focus on the quality of the English language used and less of a focus on the actual correctness of the content, which radiologists consider heavily when evaluating reports.

Readability is an important factor that measures how easy text is to read and enables clinicians to quickly read and understand a finding or report. However, current readability metrics are generic and rely on the length of words and if they are common words. [15, 16, 9] This causes medical terms, which are generally longer in length, to be negatively weighted resulting in lower readability. In the clinical domain, readability should factor in longer, common medical terminology and aim for understandable specificity.

This work focuses on understanding multidimensional evaluation of chest x-rays and producing a new metric for one of the dimensions, readability. It builds a dataset of radiologist annotations for measuring quality across factual correctness, comprehensiveness, style, and overall quality through a radiologist rating task involving MIMIC-CXR [11] chest x-rays and generated radiology reports. It creates a new readability metric that doesn't weigh common large clinical words negatively.

1.1 Background and Related Work

1.1.1 Radiology Report Generation

Radiology report generation has increased in prominence over the last few years. Generally, a computer vision model takes in a given x-ray image and returns text associated with the predicted x-ray findings report. The abundant access to x-ray datasets and reports has led to many machine learning generated report studies [7, 11, 5]. In particular, MIMIC-CXR, which we use in our study, is a large dataset of chest x-rays which holds more than 300,000 images from over 60,000 patients [11]. Our studies use the anteroposterior (AP) chest radiographs from the MIMIC-CXR dataset.

To accurately generate reports, models must be able to identify small, key differences in an image as most images typically look similar and have small meaningful finding identifiers. In the second and third experiments (Radiologist Rating Task — Multi-Dimensional Evaluation, Improved Readability Metric), we incorporate the Show, Attend, and Tell [26] and TieNet [25] models. Show, Attend, and Tell uses a convolutional neural network model to encode the image and a recurrent neural network with attention to generate the sequence of words. TieNet uses Text-Image Embedding network to create image and text representations and incorporates multi-level attention models integrated into CNN-RNN architecture to highlight meaningful text words and image regions.

1.1.2 Evaluation Metrics

Evaluation metrics allow us to evaluate model performance and quality of generated reports in a scalable and efficient manner. Current metrics can be grouped into four categories, n-gram matching, embedding matching, learned metrics, and radiology-specific metrics. BLEU and ROUGE [20, 17] are common NLP metrics that use precision and recall over n-gram overlaps between candidate and reference text. BERTScore, NUBIA, and BLEURT [27, 23, 12] are newer transformer archi-

texture based evaluation metrics that will also be used in this study. CheXpert [10] is a radiology-specific tool which evaluates the presence of 14 specific chest x-ray diagnoses. We can extract these diagnoses for reference and candidate reports, and we can use them to compute metrics such as accuracy, precision, and recall. We will apply these evaluation metrics on the generated and reference reports to determine how they differ against our human, radiologist annotators.

1.1.3 Readability Metrics

Readability metrics measure how understandable and readable text is for the average person. Common readability metrics include Flesch-Kincaid Grade Level [15], Dale-Chall [16], and Gunning-Fog Index [9]. The Flesch-Kincaid Grade Level determines the reading grade level of text based on the average number of words per sentence and the average number of syllables per word. Dale-Chall determines reading grade level by considering a list of 3000 common words to be understandable and all words not in that list to be difficult. Gunning-Fog Index determines reading grade level by the average number of words in a sentence and the average number of complex words in the given text. These metrics work well in generic scenarios but are not as meaningful in the clinical setting where accuracy is important and often words are longer. Further work to improve readability metrics has been conducted for applications in clinical text simplification [14] and complex term explanation [2]. These studies indicate a need for a new metric focused on health-related texts; currently used generic readability metrics rate reports that include longer but accurate clinical words as harder to read despite the language being common among clinicians.

Chapter 2

Experiments

Previous research suggested a need to further investigate what is important for radiology reports so that we can better automatically generate reports using machine learning. Machine learning generated reports are only as good as the examples they're trained on and the parameters and metrics that we set to tell the program if the generated report is good or not. In my research, I focus on determining how we determine if a generated radiology report is good or not; in other words does it accurately and correctly represent what an actual radiologist would write.

The first step of this process is confirming that current metrics are inadequate to develop state of the art generated chest x-ray reports. In this initial study, we conducted a radiologist ranking task where radiologists rank chest x-ray reports and we compared the radiologists' rankings to those that current metrics determined. The results of that experiment showed how radiologists' rankings differed from the current state of the art metrics. In the next step, we chose to further investigate why there was a difference between how metrics and radiologists evaluated the quality of the report. We did this through conversation with a radiologist and developing a second experiment. In this second experiment, we asked radiologists to rate reports across multiple dimensions to understand the importance level of various factors radiologists consider. Based on knowledge and further learning from these experiments, I chose to take a closer look at the readability aspect of clinical text by developing an improved metric. This metric provided better guidance on whether the generated report was

readable and understandable given the context of being clinical text where there are generally longer, less common but necessary words present.

In summary, I will cover these experiments in the following chapters:

- Radiologist Ranking Task to understand how current evaluation metrics perform for clinical radiology report text.
- Radiologist Rating Task — Multi-Dimensional Evaluation to understand how radiologists evaluate the quality of a report over multiple dimensions and how they relates to current evaluation metrics.
- Improved Readability Metric to better address the presentation/style dimension of evaluating radiology reports.

Chapter 3

Radiologist Ranking Task

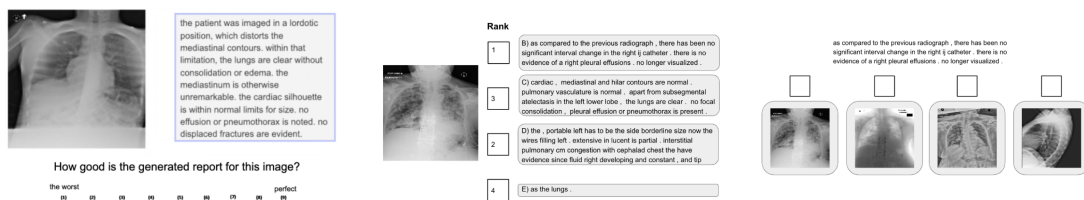
While there has been a plethora of research on machine learning generated text for images, the general and domain specific evaluation metrics are not beneficial for clinical text where correctness is the most important factor. In this initial study, we developed a radiology ranking task and qualitative discussions to understand how radiologists' ranking of a given set of generated reports compares to those that metrics would provide. Throughout the study we had frequent discussions with radiologists to understand how the task would be received and how to best extract insights.

3.1 Methods

For each image, we generated four different captions: reference, 3-gram, nearest neighbor, and random report. The reference report is the actual report written by a radiologist found in the CXR dataset. The 3-gram report is generated by finding the 100 closest images and fitting a tri-gram model with their reports. The nearest neighbor report is generated by finding the report of the most similar image in the DenseNet-induced feature space training set. Lastly, the random report is a report randomly selected from the training set (not including the actual reference report for the image). These were presented in a randomized order for each image.

We presented two radiologists with 100 diverse anteroposterior chest x-ray images and asked them to rank the associated 4 captions for each image based on how well

Figure 3-1: Three different annotation tasks we considered for the radiologists to perform.



(a) **Direct Assessment.** Radiologist would be asked to select how good the generated caption is for the image from 1-10.

(b) **Caption Ranking.** Radiologist would be asked to rank 4 proposed captions based on how well each describes the given image.

(c) **Image Selection.** Radiologist would be asked to select which image is the one being described by a given caption.

they described and presented the findings in the image. The image caption pairs were presented to both radiologists in the same order.


As this task involves asking radiologists to perform a subjective task, we explored various methods to determine what’s the best way to understand how they evaluate report quality. We considered direct assessment, caption ranking, and image selection tasks as shown in 3-1. We ultimately chose to use a caption ranking task as it limits bias when all captions for an image are presented at the same time, and annotators can be more consistent with each other by ranking the reports against each other.

Two radiologists were presented with 100 images and their associated 400 reports (4 report captions per image). As shown in figure 3-2, for each image in the task, the radiologist was presented with 3 statements:

- “The four following reports are all trying to describe this image (some of them might be factually incorrect). Please rank them from best (1) to worst (4).”
- “Briefly describe how you arrived at this ordering (a few simple bullet points is fine).”
- “Confidence that another radiologist would arrive at the same choice for best report (1=Not confident at all, 5=Very confident).”

Figure 3-2: An example of the chosen annotation task.

The four following reports are all trying to describe this image (some of them might be factually incorrect). Please rank them from best (1) to worst (4).



there has been interval placement of a right picc line, this traverses the mediastinum and the tip is positioned in the left brachiocephalic vein the lunate in the svc. no pneumothorax. there is unchanged left lower lobe a atelectasis. infection cannot be excluded. no pleural effusion seen.

1 2 3 4

BEST WORST

ap semiupright and lateral views of the chest provided. picc line intervaly removed. top normal heart size again noted. there is a NAME residual right pleural effusion. retrocardiac linear density likely represents residual mild atelectasis. difficult to exclude a developing pneumonia. no convincing signs of edema mediastinal contour appears normal. no pneumothorax. bony structures appear intact.

1 2 3 4

BEST WORST

since the prior study, the cardiac silhouette is enlarged, there is more central vascular congestion, and there is mild interstitial edema. no large pleural effusion. no pneumothorax.

1 2 3 4

BEST WORST

single ap view of the field of view. diffuse pulmonary opacities bilaterally have mildly increased, likely atelectasis, consider pneumonitis in the right apical pneumothorax remains relatively NAME, but is persistent. no free air is seen ending within the renal pelves bilaterally.

1 2 3 4

BEST WORST

In this task, we wanted to evaluate how well current metrics agree with the radiologist judgements we collected, and we wanted to evaluate how well the generated captions (1-NN, 3-gram, random) performed in terms of clinical correctness and compared to the radiologists' judgements. We focused on evaluating these current metrics: baselines (random-score, length based), readability scores (Dale-Chall), classic (BLEU-1, BLEU-4, BELU-Uniform-Smoothed, ROUGE-1, CIDEr), neural (BERTScore), and CheXpert-derived (accuracy, precision, recall, CIDEr to break ties). Our ranking task produced 6 pairwise report caption comparisons between the four captions (1st and 2nd, 1st and 3rd, 1st and 4th, 2nd and 3rd, 2nd and 4th, 3rd and 4th). We use these pairwise comparisons to more simply evaluate the performance of two captions against each other. When we evaluate given text with an evaluation metric, we get a metric score in the range of 0.0 to 1.0 where a higher metric score indicates more similarity to the reference caption. For each metric, we calculated the metric score for the 3 non-reference captions as the score would be 1.0 for the reference caption. Then for each of our 6 pairwise caption comparisons, we determined how many times the given metric scores for the captions agreed with the ranking the radiologists noted.

Table 3.1: Of the 199 consensus comparisons, how often would each metric rank the two reports the same way the radiologists did?

Metric	Percent Agree	Percent Ties
random-score	50.0%	0
choose shorter report	54.3%	0.5%
Dale-Chall Readability Index	58.3%	0
BLEU-1	53.3%	0
BLEU-4	50.8%	0
BLEU-Uniform-Smoothed	61.3%	0
ROUGE-1	56.3%	1%
CIDEr	58.8%	0
BERTScore	61.3%	0
CheXpert-accuracy	43.7%	24.6%
CheXpert-precision	39.2%	27.6%
CheXpert-recall	29.6%	41.7%
CheXpert-accuracy + .001*CIDEr	57.3%	0.5%
CheXpert-precision + .001*CIDEr	54.8%	0.5%
CheXpert-recall + .001*CIDEr	54.3%	1.0%

3.2 Results

Overall for the task, we evaluated the 600 pairwise comparisons formed from ranking 4 report captions for each of the 100 images. Our two radiologists’ pairwise rankings agreed with each other for 459 of the 600 comparisons (76.5%). When we removed the 300 comparisons that involved the reference, the radiologists agreed with each other for 199 of the remaining 300 comparisons (66.3%). Given these results, we evaluated how well current metrics compared to radiologists’ assessments, how well the generated reports performed against the actual reference report, and how radiologists decided the report rankings for each chest x-ray image.

3.2.1 Evaluating Metrics

As mentioned earlier, we evaluated commonly used metrics to determine how often their pairwise comparisons agreed with those of the radiologist. In particular, we focused on the 199 pairwise comparisons as they represent the subset where the

Table 3.2: Percent of instances that annotators ranked each method a given ranking.

	1st	2nd	3rd	4th	Average Ranking
reference	79.5%	13.0%	5.0%	2.5%	1.31
1-NN	9.5%	34.0%	37.5%	19.0%	2.66
random-report	6.0%	35.5%	39.0%	19.5%	2.72
3-gram	6.0%	17.0%	18.5%	58.5%	3.30

radiologists agree with each other and don't contain the reference caption. Table 3.1 presents the percentage of times where the metrics agreed with our expert radiologists. Of the presented metrics, BERTScore and BLEU-Uniform-Smoothed had the best performance by agreeing with experts 61.3% of the time. Many of the other metrics are around 50%, showcasing similar performance to random-score, which is essentially just picking the right answer half of the time.

3.2.2 Evaluating Generated Reports

When using machine learning to generate radiology reports, we want to be as close as possible to a clinically correct report. We evaluated how well our four captions (reference, nearest neighbor, random-report, and 3-gram) performed against each other. The two radiologists ranked 100 images each, and table 3.2 displays the average rank of each caption generation method. We see that reference is normally ranked the best followed by nearest neighbor, random-report, and lastly 3-gram. In table 3.3, we similarly see how nearest neighbor and random-report were more than three times more likely to be ranked higher than 3-gram. Additionally, nearest neighbor was chosen slightly more times as better than random-report. It is interesting to see the random-report perform much better than 3-gram and only slightly worse than the nearest neighbor. This may suggest the difficulty in discerning between which report is more bad than another or indicate that simple generation methods are not viable as images and their reports are unique.

Table 3.3: Breakdown of how often one method was ranked higher than another method in the 199 consensus pairs.

greater \ lesser	3-gram	1-NN	random-report
3-gram	0	14	16
1-NN	50	0	35
random-report	56	28	0

Table 3.4: Top n-grams from the explanations provided by annotators for decision-making. Phrases containing stop words were removed.

unigram	Count	bigram	Count	trigram	Count
“factually”	16	“even though”	6	“are factually wrong”	3
“all”	17	“most correct”	7	“one and two”	3
“wrong”	18	“hard to”	9	“not sure if”	4
“not”	21	“factually wrong”	10	“all but one”	5
“correct”	24	“all but”	13	“is hard to”	6

3.2.3 Radiologist Decision Process

Our study also asked radiologists to “Briefly describe how you arrived at this ordering (a few simple bullet points is fine),” such that we can qualitatively better understand the preformed rankings. Table 3.4 shows the top word occurrences from the radiologists’ answers to that question.

3.3 Discussion

3.3.1 Understanding Radiologist Decisions

It’s important to understand what factors radiologists consider when evaluating and writing reports such that we can better focus and align our generation techniques in a similar way.

With one of the radiologists, we discussed their thought process and how they approach writing and grading reports. The radiologist emphasized how the most important factor is the factual correctness: the findings presented in the report are completely correct. This emphasis on factual correctness is common among all clinical

texts. It does not matter how well the report is written; if it’s factually wrong, then the whole report is bad. The results in 3.4 also emphasize “factually”, “wrong”, “correct”, showcasing how both radiologists primarily prioritized identifying correct vs. wrong reports. After identifying correctness, radiologists would then evaluate which of the reports had fewer errors or more common errors. If reports had similar levels of correctness, radiologists evaluated which report is more complete and includes the relevant details. These discussions showed us how there are multiple factors radiologists consider when evaluating reports with a strong focus on clinical correctness. Thus demonstrating a need for evaluation metrics to combine these factors and emphasize clinical correctness. Further investigation into multidimensional evaluation regarding identifying the specific factors and determining how to best weigh these factors is necessary to confirm and refine this hypothesis.

3.3.2 Improved Metrics

In Table 3.1, we saw how current metrics agreed with the radiologists about 40-60% of the time. Researchers use evaluation metrics to improve machine learning algorithms and to create better predictions. Thus it’s important for these metrics to accurately depict what we’re looking for; in this case, the current metrics don’t perform as well as expected. This confirmed our suspicion that current metrics are not geared well for clinical text, specifically that of chest x-ray reports. Our discussion with radiologists informed us of how they used correctness as their most important ranking motivation. However, CheXpert, the metric which determines correctness of the content in the chest x-ray reports, performed poorly (best CheXpert score = 57.3%) when compared to the radiologists’ rankings. This presents more CheXpert discrepancy and may represent instances where a report presented CheXpert level correctness [11] but was incoherent to radiologists. It further emphasizes the need to explore multidimensional analysis as one factor is unable to define the full understanding of clinical text.

3.3.3 Study Improvements

We also acknowledge that this pilot study could be further improved to provide more meaningful results. Specifically, we could have more radiologists perform the task. A third radiologist would be beneficial towards providing a tie breaker and enable more comparisons to be evaluated. We could also include more advanced generation methods for our report captions. In this study we used more general methods such as random-report, nearest neighbor, and 3-gram; however, there are many more advanced image captioning algorithms available that could provide a better report and that are more commonly used in image captioning. In this pilot study, we used 100 images and 400 captions which resulted in a much smaller pool of pairwise comparison evaluation. We can use a more diverse set of images and more images to improve the reliability of our data and results. The next two experiments will incorporate these improvement suggestions.

3.4 Conclusion

In this initial radiologist ranking task study, we further understand how radiologist determine if a report is good or not and how their expert judgement compares to current state of the art evaluation metrics. Our study suggests that current metrics are inadequate as they only match radiologists' opinions around 40-60% of the time. Our analysis of the generated reports also suggests that additional research with more advanced techniques is required to capture what precisely radiologists are considering outside of correctness when determining which reports are better than others. In the next study, we will continue to further understand how radiologists evaluate reports across multiple dimensions.

Chapter 4

Radiologist Rating Task — Multi-Dimensional Evaluation

The previous pilot study, radiologist ranking task, concluded a need to better understand the factors radiologists consider when determining if a report is of good quality or not. Understanding the balance between these factors can help us better create metrics for clinical text that more accurately represent the opinions of actual clinicians. In this radiologist multi-dimensional evaluation study, we asked three expert radiologists to evaluate a diverse set of chest x-ray images and their respective four generated captions across four dimensions: factual correctness, comprehensiveness, style, and overall quality.

4.1 Methods

In this study, we asked three radiologists to rate 4 generated report captions for each of the 200 chest x-ray DICOM (Digital Imaging and communications in Medicine) [21] images presented across 4 dimensions: factual correctness of interpretation, comprehensiveness, presentation/style, and overall rating. We used a different set of images and models than what was used in the previous experiment to increase the image diversity and use of advanced generation models. We then analyzed how well current metrics performed compared to the radiologists' dimensional ratings to further

understand if metrics correlate with a specific dimension.

4.1.1 Report Generation Models

We have seen significant progress in research for computer vision models, specifically those adapted for chest x-ray report generation. When trained correctly, these models take in a chest x-ray image and output a correct, readable report caption. In this study, we use two generation model techniques: Show, Attend, and Tell; and TieNet. Show, Attend, and Tell [26] models first use convolutional neural networks to build a feature map encoding the image and then use a recurrent neural network module integrated with attention to generate word by word tokens.¹ TieNet (Text-Image Embedding network) models use end to end convolutional neural network - recurrent neural networks integrated with multi-level attention to extract meaningful image and text representations [25].²

4.1.2 Selecting Diverse Images

It's important to select a group of diverse images that would present quality captions. We chose our images based on the diagnoses presented in the reference report to create a wider set of diverse diagnoses presented in the images and corresponding report captions. We first generated the following captions for our total set of images (obtained from the CXR dataset): Show, Attend, and Tell with beam size 1 (Show, Attend, and Tell-1); Show, Attend, and Tell with beam size 5 (Show, Attend, and Tell-5); TieNet with beam size 1 (TieNet-1); TieNet with beam size 5 (TieNet-5). For each image, we determined the CheXpert labels for the reference caption and each of the 4 generated captions. To experiment with high quality report captions, we determined which images had the best set of generated captions by calculating the average accuracy between the reference (correct) CheXpert labels and each of the generated captions' CheXpert labels. We then chose the top 200 images with the

¹An implementation can be found at <https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning>.

²An implementation can be found at <https://github.com/MLforHealth/TieNetReproduction>.

highest agreement accuracy for our task.

4.1.3 Generated Report Captions

In this study, we used the advanced caption generation methods to create higher performing captions to more precisely understand how radiologists score reports. For each of the 200 chest x-ray images, radiologists were presented with 4 randomly sorted captions (mix of advanced generation, standard generation [4], and reference):

- Reference
- TieNet with beam size 5 (highest performing advanced generation method)
- KNN OR Random - each chosen with a 50% probability
- Show, Attend, and Tell with beam size 5 OR TieNet with beam size 1 — chosen based on which one had the highest CheXpert similarity score

The reason these captions were chosen in this manner was to present radiologists with mostly high quality generated captions while also adding diversity of captions. This focuses the radiologist’s attention on evaluating captions with more detail and highlights more factors of the evaluation process outside of factual correctness.

4.1.4 Multi-Dimensional Analysis

Following conversation with radiologists regarding our previous pilot ranking study, we focus this study on evaluating captions across multiple factors. Our discussions led us to identify 4 scoring dimensions: factual correctness of interpretation, comprehensiveness, presentation/style, and overall rating. Dimensions:

1. Factual Correctness of Interpretation: focuses on the correctness of the information presented in the report. This is not an overall correctness of the report but correctness of the information presented. Missing information is addressed in comprehensiveness.

Category	Scale	Scoring Guidelines
Factual Correctness of Interpretation	1-5	1: Mentioned facts are totally incorrect 2: Mentioned facts are mostly incorrect 3: Mentioned facts are mostly correct but important details incorrect 4: Mentioned facts are mostly correct 5: Mentioned facts are entirely correct
Comprehensiveness	1-5	1: Misses most / all important findings 2: Misses 1-2 important findings 3: Misses no important findings but more than one minor finding 4: Misses no important finding but one minor finding 5: Covers all important and minor findings
Presentation / Style	1-5	1: The report legitimately does not make sense. Perhaps the author is a computer or otherwise struggles with English. 2: Individual phrases are sound but the overall report is not fully comprehensible 3: The report seems to be written by someone who can communicate in English but who has no radiology training (e.g. atypical phrases/word choices or deviation from radiology reporting standards) 4: Report partially follows radiology standards but has some flaws. 5: A medical expert wrote this report. It is stylistically good.
Overall Rating	1-7	If a radiology resident wrote this, what overall numerical grade would you give them? 1: very bad 2: bad 3: below average 4: average 5: above average 6: good 7: very good

Figure 4-1: Scoring Guidelines provided to radiologists.

2. Comprehensiveness: focuses on if the report mentions all the relevant and necessary information. A report that mentions all the relevant information but incorrectly represents some piece of information should still receive a high comprehensiveness (and a lower factual correctness).
3. Presentation / Style: focuses on the language of the report to gauge if it looks like something an actual radiologist would write. You should not take into account any incorrectness or comprehensiveness.
4. Overall Rating: focuses on the general: is this a good report. Correctness, comprehensiveness, and style are all considered together.

Figure 4-1 showcases the scoring guidelines we provided to radiologists to identify how to rate across these dimensions and reduce the variability amongst scores. We identified rating ranges of 1 to 5 and 1 to 7 based on discussions with radiologists and given feedback from the previous study. Ratings can offer more insight into how

much better or worse a caption is than another, which presents more insights than a ranking task. However, this means it's more important to emphasize the scoring guidelines.

4.1.5 Rating Task

Three highly-skilled radiologists were presented with 200 images and their respective 4 captions. Radiologists selected caption ratings from a dropdown which listed the score along with the score's explanation to continuously reinforce our scoring rubric. They were also presented with a DICOM viewer [1] link to better view the chest x-ray in their preferred format. Figure 4-2 displays an example of the completed task for one image.

4.1.6 Evaluation Techniques

Of the 200 images, all three radiologists completed the task for 167 images (one of the radiologists did not finish the task completely). Thus, we base our results and findings on that group of 167 images. The 167 images resulted in 501 caption pairs of the reference to another generative caption. With the collected radiologists' ratings we evaluated the following:

- We compared the radiologists' rating scores to common NLP metrics: symbolic (Meteor, Nubia, Rouge), rule-based (CheXpert Recall, CheXpert Accuracy, CheXpert Precision), and transformer architectures (Bleu, Bertscore, Bleurt).
- We analyzed how radiologists ranked generated captions across the four dimensions to understand how they weigh different factors when determining the quality of a report.
- We compared the radiologist's multi-dimensional rankings to the metric scores' rankings to identify what specific dimensions metrics address, if any.
- We also looked at the agreement level for the group of three radiologists to identify how similar the expert radiologists' evaluations are.

Rating Task 139:				
The four following reports are all trying to describe this image (some of them might be factually incorrect). Please rate them for each category. Category scoring guidelines are on the Scoring Guidelines sheet.				
Dicom Viewer Link: https://www.pacsbin.com/ [REDACTED]				
Caption	Factual Correctness (1 to 5)	Comprehensiveness (1 to 5)	Presentation / Style (1 to 5)	Overall rating (1 to 7)
the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.	5: Mentioned facts are entirely correct	4: Misses no important finding but one minor finding	5: A medical expert wrote this report. It is stylistically good.	6: good
the lungs are relatively hyperinflated but clear without consolidation, effusion, or edema. moderate cardiac enlargement is noted compatible with patient's history. no acute osseous abnormalities.	3: Mentioned facts are mostly correct but important details incorrect	4: Misses no important finding but one minor finding	5: A medical expert wrote this report. It is stylistically good.	6: good
heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. no pleural effusion or pneumothorax is seen. there are no acute osseous abnormalities.	3: Mentioned facts are mostly correct but important details incorrect	4: Misses no important finding but one minor finding	5: A medical expert wrote this report. It is stylistically good.	6: good
ap upright and lateral views of the chest provided. lungs are hyperinflated and lucent likely reflecting underlying NAME. prominent costal cartilage accounts for para nodularity projecting over the left lower lung. no large effusion or pneumothorax. cardiomeastinal silhouette is normal. bony structures are intact. no free air below the right hemidiaphragm.	4: Mentioned facts are mostly correct	2: Misses 1-2 important findings	1: The report legitimately does not make sense. Perhaps the author is a computer or otherwise struggles with English.	2: bad

Figure 4-2: Task presented to radiologists, where all captions are shown at the same time.

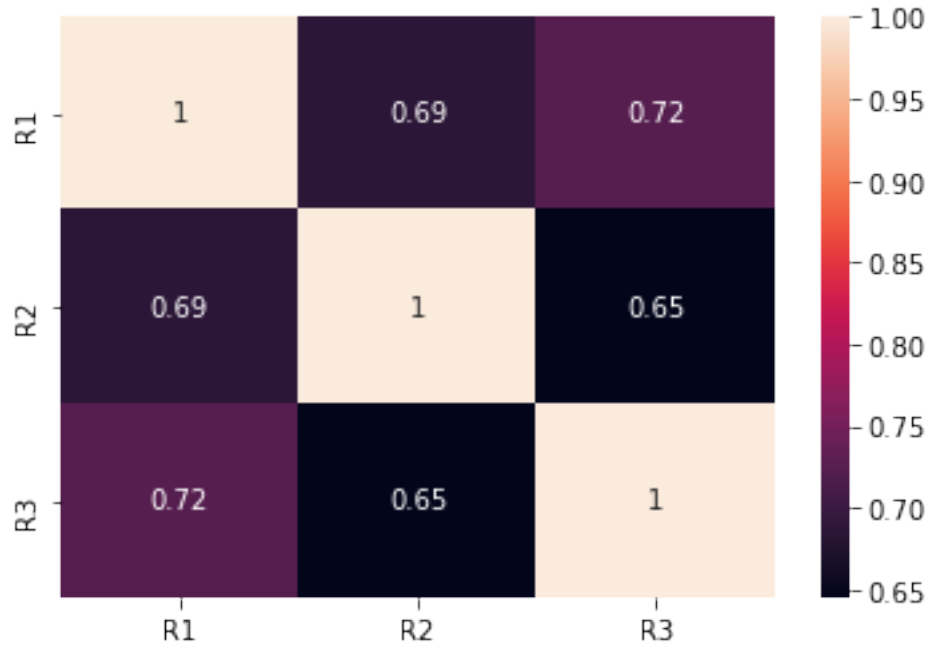


Figure 4-3: Overview of radiologist inter-annotator agreement

In the cases where we evaluate the radiologists’ scores with the metrics, we use the average score of the three radiologists’ rating scores for each of the rating dimensions.

4.2 Results

For the 167 image tasks that all three radiologists completed, we evaluated how well they agreed with each other and with the evaluation metrics, how their ratings and metrics compared across the four dimensions, and how well the generated reports performed against each other.

4.2.1 Radiologist Agreement

We had three radiologists complete a multi-dimensional rating task for 167 images. In order to assess the quality of the data collected, it is important to understand how well the radiologists agreed with each other. Figure 4-3 displays how well each radiologist agreed with the others for each of the 501 comparisons. These Pearson correlations lie between 0.65 and 0.72, indicating that radiologist mostly agree with each other. How-

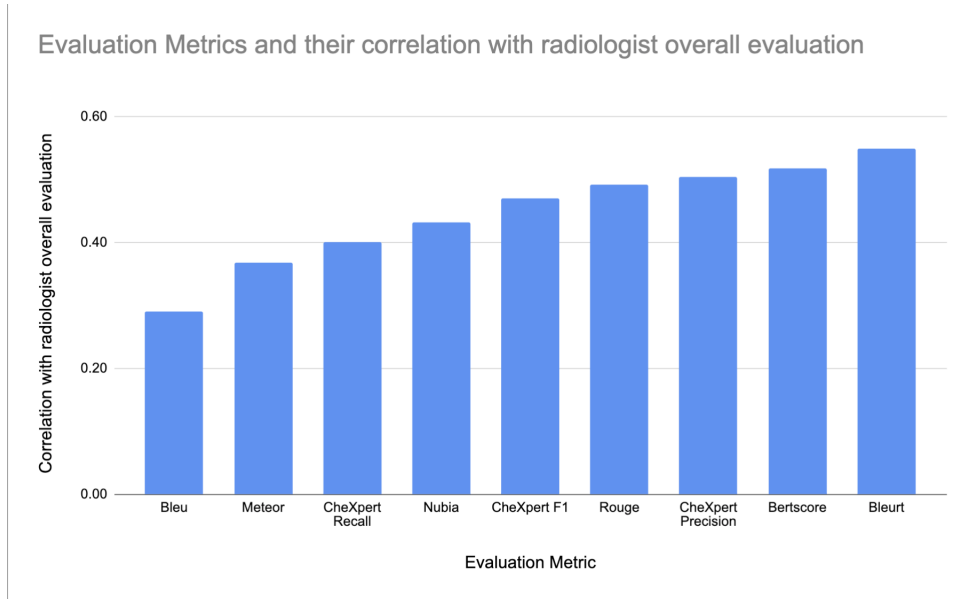


Figure 4-4: Overview of different evaluation metrics and their agreement with radiologist overall rating

ever, this agreement level varies more across the dimensions. The factual correctness dimension has agreement levels of about 0.8 while the presentation/style dimension have an agreement level around 0.5.

4.2.2 Evaluating Metrics

In this study, we analyzed how well three groups of metrics matched the radiologists' overall rating evaluation dimension. This allows us to understand how well evaluation metrics accurately reflect expert evaluation. In figure 4-4, we see that Bleurt and Bertscore are slightly more correlated with expert opinion. However, Bleurt (the most correlated metric) only has a 0.55 correlation.

4.2.3 Evaluating Generated Reports

We also evaluated which generation method ranked higher over others given our metrics and radiologist judgement. Based on the unanimity in rankings, we grouped the metrics into two groups:

- Group 1 Metrics: Bleu, Rouge, NUBIA, Bleurt. These metrics ranked gener-

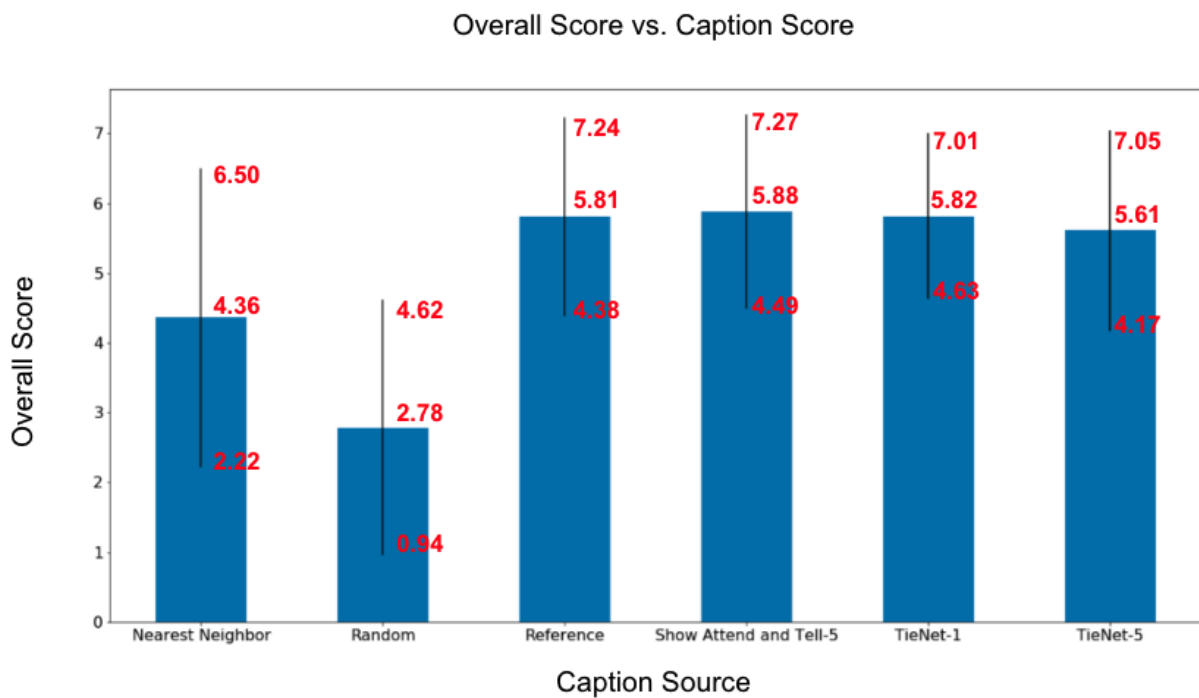


Figure 4-5: This graph shows the average overall rating (range of 1 to 7 (best)) radiologists scored for each caption generation method along with one standard deviation error bars.

ation methods best to worst as: Show Attend and Tell-5, TieNet-1, TieNet-5, Nearest Neighbor, and Random.

- Group 2 Metrics: Bertscore, Meteor. These metrics ranked generation methods best to worst as: Show Attend and Tell-5, TieNet-1, Nearest Neighbor, TieNet-5, and Random.

In figure 4-6, we see how for overall, comprehensiveness, and factual correctness, radiologists and metrics all chose Show Attend and Tell-5 as the best caption (not considering the reference). However, for style, the radiologists chose TieNet-1. Additionally, we see random performing the worst for all metrics and radiologist evaluation dimensions, which makes sense and reemphasizes that our other generated reports represented the image at least better than a random report selection. We also see the advanced generation methods (Show Attend and Tell and TieNet) performing better than nearest neighbor.

Figure 4-5 shows the average overall rating (range of 1 to 7 (best)) radiologists scored for each caption generation method. We see that Show Attend and Tell 5 was given on average the highest score. We also see that two of the advanced generation methods were perceived to be similar to the actual reference caption.

4.2.4 Multi-Dimensional Analysis

We wanted to understand how various factors (factual correctness, comprehensiveness, and style) contribute towards the overall rating of a generated report. Figure 4-7 displays how our four dimensions correlated with each other. We see that factual correctness and comprehensiveness are both highly correlated with the overall rating and style is only slightly less correlated.

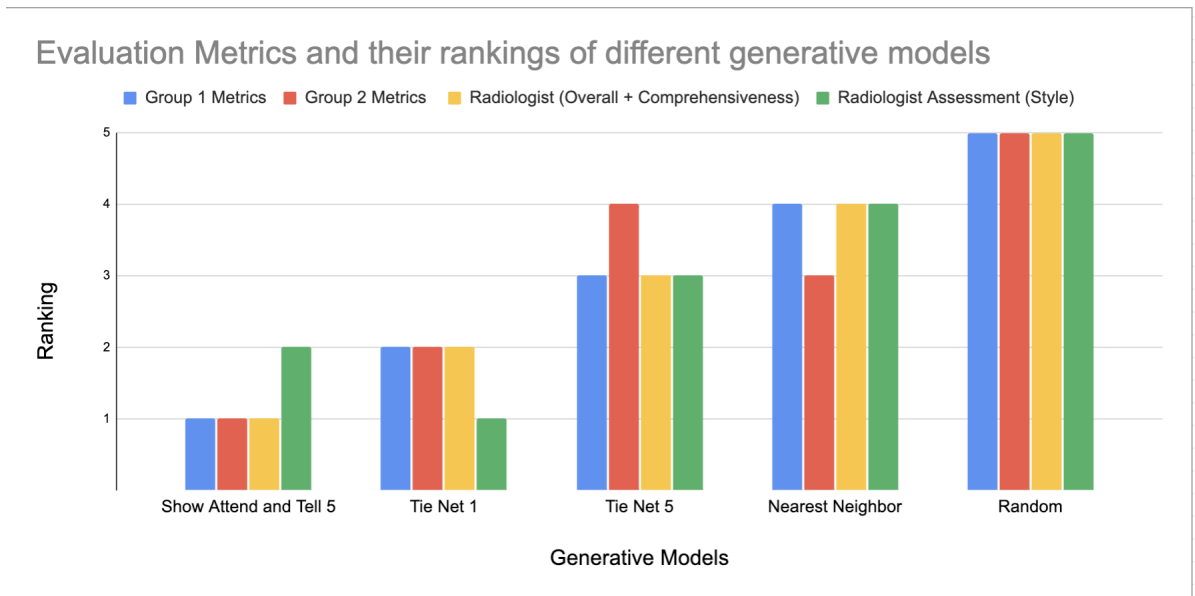


Figure 4-6: This graph shows how metrics and the averaged radiologists' dimensional ratings ranked the different types of generative models. Lower bars indicate a better ranking and better performing report.

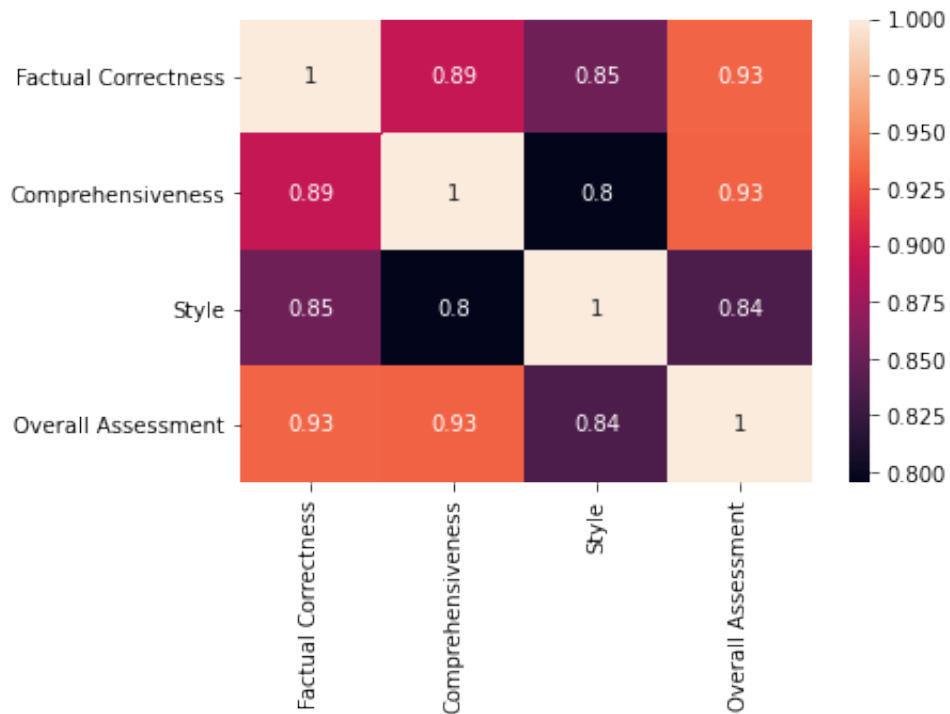


Figure 4-7: This heat map shows the Pearson correlation for four dimensions based on the average of the three radiologists' ratings for each dimension.

4.3 Discussion

4.3.1 Generation Methods

In the task, we incorporated a mix of advanced generation techniques (Show Attend and Tell, TieNet) alongside more simple techniques (random-report, nearest neighbor). In figure 4-5, it was interesting to see these advanced generation techniques outperform the actual reference report caption. This shows how generated clinical text reports are a viable option as they can provide a factually correct report in a more standardized format.

4.3.2 Need for Improved Metrics

When looking at metric agreement with our expert radiologists, we saw transformer based architectures outperform domain specific evaluation metrics. CheXpert is specifically designed for chest x-ray reports, and it is interesting to see learning based metrics perform better as they are not trained on the chest x-ray report context yet have a better understanding of what reports are better. Additionally, the overall range of metrics' agreements with expert judgement of 0.55 to 0.3 suggests the need to design better evaluation metrics.

We also see how various dimensions of the radiologists' judgements correlate better with different groups of metrics and how those dimensions correlate with the overall rating.

4.3.3 Improved Study

Our results indicate potential areas of improvement in our study. The overall inter-annotator agreement rate (Pearson correlation) of 0.65 to 0.72 and 0.5 agreement rate for presentation/style highlights the subjective nature of this task. Despite clear scoring guidelines, radiologist ratings were variable. The high agreement between factual correctness and comprehensiveness dimensions could imply how these dimensions are not very distinct or were interpreted differently than what was intended.

Future studies may involve a training pilot program to help calibrate the radiologists scores and provide better data precision.

4.4 Conclusion

In this radiologist rating task and multi-dimensional evaluation study, we further understood how radiologists decide if a report is good and how metrics correspond to those multi-dimensional considerations. We saw that machine learning generated report captions perform competitively to reference captions, transformer based metrics performed better than domain specific metrics, and correlation existed between the rated dimensions. This study leads way to further investigation on how we can build a new metric based on the segmented evaluations of each dimension. The next experiment will focus on addressing the readability (presentation/style) dimension metric.

Chapter 5

Improved Readability Metric

Our previous two studies indicated a need to explore improvements for metrics that are applied to clinical text. In order to generate strong chest x-ray radiology reports from machine learning models, we need to develop new metrics that address the various factors radiologists consider and weigh them accordingly. These factors can be defined broadly as factual correctness, comprehensiveness, and readability. We can break down the future clinical metric into each of these components, develop those components individually, and combine them together with correct weights.

This study focuses on developing and understanding the readability metric factor further. In particular, we focus on adapting the Dale-Chall Readability Formula for clinical contexts. Of the common general readability metrics (Dale-Chall, Flesch-Kincaid, Gunning Fog, and more), Dale-Chall incorporates an element based on the difficulty of words. However, in clinical text, many words are considered “difficult” but are necessary for correctly explaining findings. It does not make sense to score necessary words that are common in a clinical setting lower; this lower scoring steers models away from then using those words. Thus, there’s an opportunity to improve this Dale-Chall Readability Formula by better defining what difficult words mean. I also investigated improving the Flesch-Kincaid formula; however, it’s a convoluted task to address clinical text in terms of sentence length and average syllables per word.

In this experiment, I replicated the Dale-Chall Readability formula, identified com-

mon terminology in chest x-ray reports, and improved the formula by altering how the common medical terminology was scored. I then evaluated the improved metrics’ performance against the current Dale-Chall implementation for the generated radiology reports presented in the multidimensional analysis rating task. Additionally, I evaluated how the improved metric aligned with radiologists’ multidimensional ratings; specifically addressing how correlated it is to the four dimensions (factual correctness, comprehensiveness, presentation/style, and overall).

5.1 Methods

5.1.1 Dataset

We evaluate our new improved readability metric on the data captured in the previous multi-dimensional radiologist rating task study. Specifically, we focus on the rating task of one of the radiologists who fully completed the task for all 200 images as described in the previous Radiologist Rating Task — Multi-Dimensional Evaluation study.

From that rating task, we have a resulting set of 800 generated report captions (4 captions for each of the 200 images). This group of 800 generated report captions includes 200 reference, 200 TieNet-5, 154 TieNet-1, 106 nearest neighbor, 94 random-report, and 46 Show Attend and Tell-5.

For each of these 800 report captions, we also use the radiologist’s ratings for the four dimensions: factual correctness, comprehensiveness, presentation/style, and overall rating.

5.1.2 Designing the Improved Dale-Chall Readability Metric

The first part in designing an improved readability metric was to re-implement the standard Dale-Chall Readability metric. Our previous studies used the Dale-Chall implementation found in the py-readability-metrics package.¹ In this implementation,

¹An implementation of the py-readability-metrics package can be found at <https://github.com/cdimascio/py-readability-metrics>.

the Dale-Chall score is calculated as:

$$\begin{aligned} \textit{percentage_of_difficult_words} &= \\ & \textit{number_of_complex_words}/\textit{number_of_words} * 100 \\ \\ \textit{raw_score} &= 0.1579 * \textit{percentage_of_difficult_words} \\ & \quad + 0.0496 * \textit{average_number_of_words_per_sentence} \\ \\ \textit{adjusted_score} &= \begin{cases} \textit{raw_score} + 3.6365, & \textit{percentage_of_difficult_words} \geq 5 \\ \textit{raw_score}, & \textit{otherwise} \end{cases} \end{aligned}$$

In these formulas, *number_of_complex_words* represents the number of words that are not present in a list of 2950 common words. Additionally, we use the Porter stemmer to check if the stem of a given word is present in the common word list to determine if it is classified as a complex or common word.

In our clinical text context, there are many words that can be considered common and meaningful for chest x-ray reports. However, these words (effusion, pneumothorax, pleural, lungs, clear, mediastinal, etc.) are considered complex in the standard Dale-Chall Readability formula and are thus result in a higher readability score. To address this, I adjusted the Dale-Chall common word list by adding common words that appeared in generated report captions. Specifically, I identified the words that appeared more than five times across the generated report caption, and I added this set of 235 words to the list of common words, resulting in an updated set of 3185 common words.

This initial design of an improved Dale-Chall readability metric incorporates the standard Dale-Chall Readability formula while using the updated set of context-specific common words. Readability scores correspond to reading grade level, so higher scores mean the text is at a higher reading level and more difficult to read.²

²An implementation of the improved readability metrics can be found at <https://github.com/23saumyar/Improved-Readability-Metrics>.

5.1.3 Evaluation Techniques

For the 800 generated report captions (4 captions per 200 images), I calculated the standard Dale-Chall readability score and the improved Dale-Chall readability score. To understand if there is a meaningful improvement, I analyzed how the improved formula compared to the standard formula for each type of generated caption and how the formulas correlated with the radiologist’s rated dimensions.

For each type of generated caption (reference, random-report, nearest neighbor, Show Attend and Tell-5, TieNet-1, and TieNet-5), I calculated the average standard Dale-Chall and improved Dale-Chall score to understand if certain generation methods produced more readable scores.

In previous studies, we discovered how radiologists evaluate the quality of a report through multi-dimensional factors. We hypothesize readability represents one of these dimensions. To evaluate this hypothesis, I evaluated how well the standard Dale-Chall and improved Dale-Chall score correlated with the radiologist’s rated dimensions (factual correctness, comprehensiveness, presentation/style, and overall).

5.2 Results

5.2.1 Evaluating Metrics

We calculated the standard and improved Dale-Chall readability score for each of the 800 generated report captions. Figure 5-1 shows the average readability score for each type of generation method. We see that the standard Dale-Chall readability scores the reports around 11 which represents a higher reading difficulty than the improved Dale-Chall report scores of around 8, as expected since we classified previously complex words as common words. Interestingly, the metric score rankings of the caption generation methods is different for the two metrics. For the standard Dale-Chall metric, the caption generation methods are ranked easiest to hardest as: TieNet 5, Show Attend and Tell 5, random-report, TieNet 1, nearest neighbor, and reference. In this case, the actual caption or reference caption is rated on average as the most

Average Readability Metric Scores for Generated Report Captions

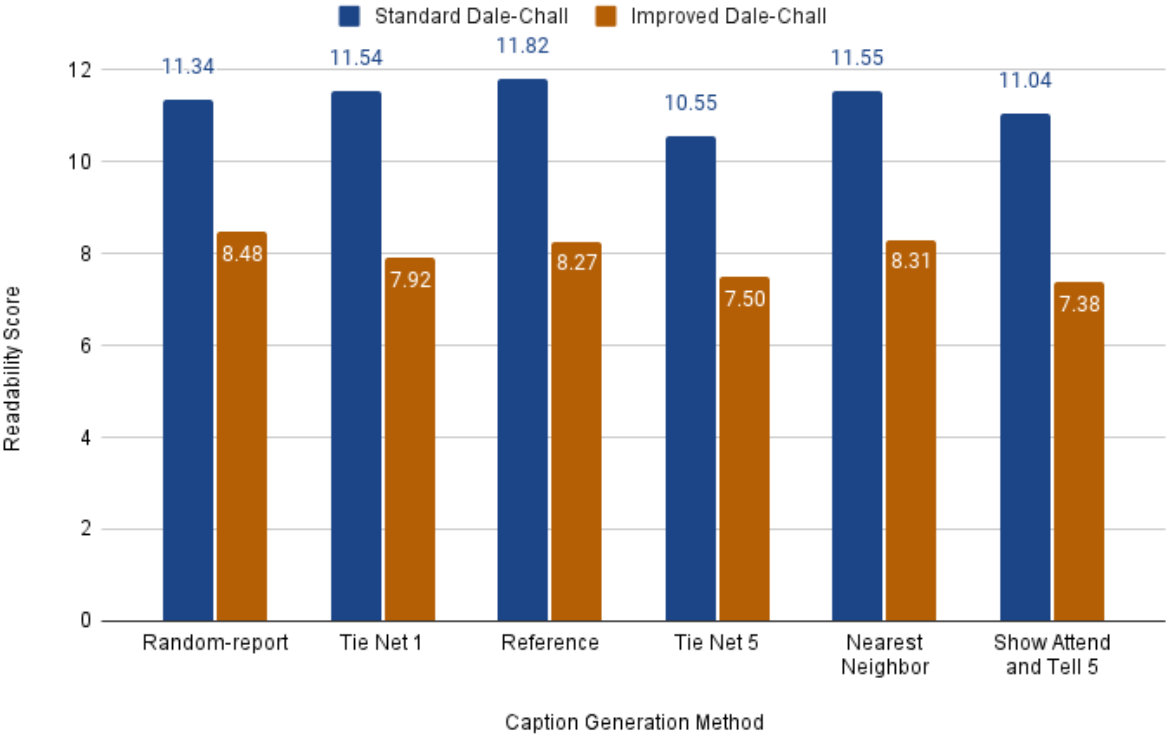


Figure 5-1: Average Dale-Chall readability scores for each report caption generation method where lower values represent the metric scoring the report as more readable.

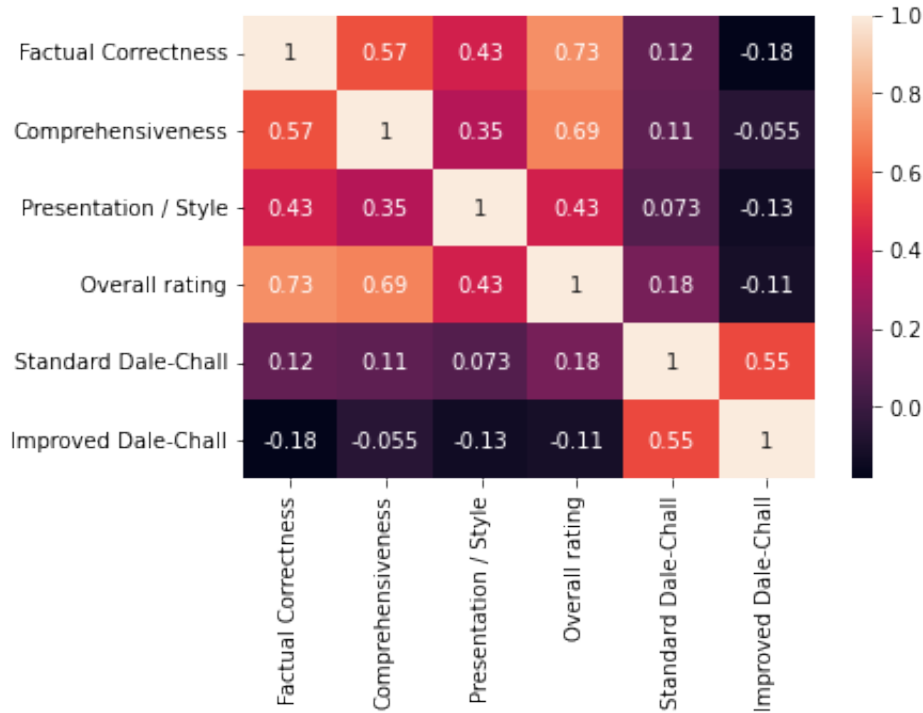


Figure 5-2: This heat map shows the Pearson correlation between the radiologist’s rated dimensions, standard Dale-Chall readability metric, and the Improved Dale-Chall readability metric. Note that for the four rated dimensions (Factual Correctness, Comprehensiveness, Presentation/Style, and Overall Rating) a higher score means a better report. However, for the Standard Dale-Chall and Improved Dale-Chall a higher score means a harder to read or worse report.

difficult to understand. On the other hand, for the improved Dale-Chall metric, the caption generation methods are ranked easiest to hardest as: Show Attend and Tell 5, TieNet 5, TieNet 1, reference, nearest neighbor, random-report. While the scores are overall lower, we still see the captions written by an actual radiologist (reference, nearest neighbor, random-report) having the highest (more difficult) scores.

5.2.2 Multi-dimensional Analysis

From our previous studies, we learned how radiologists evaluate multiple factors to determine if a report is good or not. We wanted to evaluate if the standard and improved Dale-Chall Readability metrics correlated with one of the factors the radiologist rated (Factual Correctness, Comprehensiveness, Presentation/Style, and Overall Rating).

Figure 5-2 shows a heat map of the Pearson correlation between the rated dimensions and the Dale-Chall scores. For the rated dimensions, a higher score means a better report. However, for the Dale-Chall scores, a higher score means a more difficult to read report or a worse report. Thus, a negative correlation between Dale-Chall and a rated dimension implies they both are correlated in choosing a better report. In the heat map, we see positive correlations between the Standard Dale-Chall Readability score and all four dimensions; this means when the rated dimensions choose a better report, Standard Dale-Chall chooses a more difficult to read report. However, we see negative correlations between the Improved Dale-Chall Readability score and all four rated dimensions. Specifically, the improved Dale-Chall Readability score was most negatively correlated with Factual Correctness followed by Presentation/Style, Overall Rating, and Comprehensiveness. These slight negative correlations hint that there's more to evaluating reports, even across specific related dimensions, than just the readability of them, further confirming what we saw in the previous studies. In particular, it was surprising to see the improved readability metric still have higher correlation with Factual Correctness than Presentation/Style; this hints that content is still somewhat considered when evaluating the Presentation/Style dimension.

5.3 Discussion

5.3.1 Improved Metric Performance

Our results show how the Improved Dale-Chall Readability metric resulted in overall lower readability scores and had a negative correlation with radiologist's rated dimensions. It makes sense to have lower readability scores as we defined frequent clinical words as common words. Additionally, we also see the Improved metric rating the reports that were written by actual radiologists (random-report, reference, nearest neighbor) as more difficult to read. This makes sense since these reports have more variability in text while the machine learning generated reports generally have a more limited and standardized vocabulary. The negative correlation implies that for higher

radiologist rated dimensions' scores, there is a lower (easier) readability score. Thus, we see that the Improved Dale-Chall metric is slightly more accurately capturing the readability quality of the reports while the Standard Dale-Chall metric was rating better reports as more difficult to read.

5.3.2 Readability Metric Improvements

Overall, the slight improvement in the Dale-Chall Readability metric resulted in a metric that better reflected if a report was better or worse. However, there is still ample room for improvement and refinement. One potential way to improve this metric is to investigate which frequency of word occurrence (this study uses 5) moves it into the common words list. Additionally, the weights in the Dale-Chall formula can be tuned if the formula is trained given a large set of captions and their rated dimension.

5.4 Conclusion

The Improved Dale-Chall Readability metric presented in this study serves as the first step in designing more meaningful metrics for radiology report evaluation. Its higher correlation with radiologist's ratings shows how slightly adapting metrics for a given context can provide a much more meaningful evaluations core. The high correlation with presentation represents how we can tie certain metrics to a particular dimension and work towards combining these metrics and dimensions to build evaluation metrics that evaluate and score like an actual radiologist would. Further work in investigating and training the weights of the Dale-Chall Readability score formula can provide further precision for evaluating the readability levels of radiology reports. However, it's important to note the tension between readability and precision in a radiology report. Attempting to make a report more understandable can potentially wash out detail which can impact comprehensiveness and accuracy. We want to address this by continuing to understand how to best write and standardized report findings to make them easier to quickly understand while also recognizing the specificity of unique

situations that should not be removed.

Chapter 6

Conclusion

As we continue to explore machine learning generated clinical text, specifically in the context of chest x-rays, it is important to build and evaluate the models correctly. We need to understand what makes the clinical text good and how we can teach our models to look for those aspects through evaluation metrics. My thesis focuses on understanding how radiologists determine if a report is good or not, the current generation and evaluation metrics, and addressing the disparity between current metrics and expert opinion.

6.1 Main Contributions

Specifically, my main contributions through this work are:

- Designing ranking and rating tasks to understand how radiologists evaluate chest x-ray reports.
- Analyzing common NLP evaluation metrics against radiologists' evaluation data.
- Generating an improved readability metric as a first step towards creating better metrics for clinical text.

6.2 Future Work

From this work, we have a better understanding of what components radiologists evaluate when determining if the clinical text report is good or not. Further radiologist annotation tasks with a larger, diverse group of chest x-rays and more radiologists will shed more light on the dimensional weights of the factors (factual correctness, comprehensiveness, presentation/style) we identified through qualitative discussion with radiologists. Given a clearer understanding of these dimensional components, further work can be done to build a new metric that more accurately represents a radiologist's judgement. I envision this new metric to be an adaptation and combination of existing metrics that can be further adapted and improved for other clinical text where correctness is also incredibly important.

Bibliography

- [1] Pacsbin, your personal, anonymized dicom library.
- [2] Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Karen Dunn Lopez, Richard Cameron, and Gail M Keenan. Generating summaries of hospitalizations: A new metric to assess the complexity of medical terms and their definitions. In *Proceedings of the 9th International Natural Language Generation conference*, pages 26–30, 2016.
- [3] William Boag, Renan Campos, Kate Saenko, and Anna Rumshisky. Mutt: Metric unit testing for language generation tasks. In *ACL*, August 2016.
- [4] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alsentzer, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Baselines for chest x-ray report generation. In *Machine Learning for Health workshop at NeurIPS*. Machine Learning for Health workshop at NeurIPS, 2019.
- [5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*, 2019.
- [6] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.
- [7] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
- [8] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340*, 2018.
- [9] Robert Gunning. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13, 1969.
- [10] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpan-skaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels

- and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.
- [11] Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 1(2), 2019.
- [12] Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland), December 2020. Association for Computational Linguistics.
- [13] Hassan Kané, Yusuf Kocyigit, Pelkins Ajanoh, Ali Abdalla, and Mohamed Coulibali. Towards neural similarity evaluator. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.
- [14] David Kauchak and Gondy Leroy. Moving beyond readability metrics for health-related text simplification. *IT professional*, 18(3):45–51, 2016.
- [15] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.
- [16] George R Klare. A table for rapid determination of dale-chall readability scores. *Educational Research Bulletin*, pages 43–47, 1952.
- [17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [18] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*, 2019.
- [19] Matthew B.A. McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 913–927, Virtual, 07–08 Aug 2020. PMLR.
- [20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, USA, 2002. Association for Computational Linguistics.

- [21] Charles Parisot. The dicom standard. *The International Journal of Cardiac Imaging*, 11(3):171–177, 1995.
- [22] Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018.
- [23] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.
- [24] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020.
- [25] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.
- [26] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.
- [27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.