A Pilot Study in Surveying Clinical Judgments to Evaluate Radiology Report Generation

William Boag
MIT
USA

Hassan Kané
WL Research
USA

Saumya Rawat
MIT
USA

Jesse Wei
Beth Israel Deaconess Medical Center, Department of Radiology
USA

Alexander Goehler
Beth Israel Deaconess Medical Center, Department of Radiology
USA

ABSTRACT

The recent release of many Chest X-Ray datasets has prompted a lot of interest in radiology report generation. To date, this has been framed as an image captioning task, where the machine takes an RGB image as input and generates a 2-3 sentence summary of findings as output. The quality of these reports has been canonically measured using metrics from the NLP community for language generation such as Machine Translation and Summarization. However, the evaluation metrics (e.g. BLEU, CIDEr) are inappropriate for the medical domain, where clinical correctness is critical. To address this, our team brought together machine learning experts with radiologists for a pilot study in co-designing a better metric for evaluating the quality of an algorithmically-generated radiology report. The interdisciplinary collaborative process involved multiple interviews, outreach, and preliminary annotation to design a larger scale study – which is now underway – to build a more meaningful evaluation tool.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; • **Human-centered computing** → *Collaborative and social computing design and evaluation methods.*

KEYWORDS

Radiology Report Generation, Participatory ML

ACM Reference Format:
William Boag, Hassan Kané, Saumya Rawat, Jesse Wei, and Alexander Goehler. 2021. A Pilot Study in Surveying Clinical Judgments to Evaluate Radiology Report Generation. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3442188.3445909

1 INTRODUCTION

Over the past few years, there has been considerable interest in the task of creating deep learning models to interpret medical images. Many of the use cases include tumor detection/localization [37], automatic segmentation of CT [18], and X-ray scans [19, 28], and image registration [13, 38]. All of these scenarios have structured outputs, such as pixel labels or image labels, which allows standard loss functions and evaluation metrics to lead to models which converge towards high performance.

The situation is different, however, for tasks which generate clinical text, such as document summarization, clinical question answering, and report generation. Evaluation metrics used in language generation tasks are not as reliable, both general and domain-specific settings, and are an area of active research.

In this work, we provide three main contributions:

- We develop and conduct an annotation task to collect clinical judgment on 400 candidate reports from 100 radiology images. This provides much-needed guidance on what makes a report good or bad.
- We examine what radiologists look at when evaluating a report. This is useful both to understand the current limitations of the task framing and also to help inform future development of a better evaluation metric for Chest X-Ray report generation.
- We demonstrate some of the outreach tools we used for initial contact with domain experts to help create discussions and eventual partnerships.

Our findings highlight the need for data scientists to work closely with clinical experts to build meaningful tasks and models. Without domain knowledge and wisdom, it is far too easy to fall into the trap of incorrect modelling assumptions, such as treating radiology report generation as a simple image captioning task with readily available labels in the reports that accompany frontal radiographs. And once the task is properly defined, interdisciplinary teams must work together to develop sound evaluation metrics which service as sound proxies for clinical usefulness.

The rest of the paper is as follows: Section 2 contextualizes this effort in relation to related work; Section 3 describes the methodological design of the pilot, report candidates, and evaluation metrics; Section 4 shows the results of the pilot and related analyses; Section 5 discusses the implications of these results; Section 7 offers a conclusion of our results; and Section 6 discusses the limitations of this study and how they inform future work.

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445909

2 BACKGROUND AND RELATED WORK

2.1 Generating and Evaluating Radiology Reports

In recent years, several Chest X-Ray datasets, including both images and clinical reports, have been made publicly available [4, 8, 16]. Some efforts in radiology AI have worked to model text and images jointly in order to: predict disease severity of illness [32], identify regions of interest over chest X-ray images [26], and allow information retrieval [14]. Since these datasets have come out, many works have generated radiology reports, including with template-informed approaches [11] as well as reinforcement learning and radiology-derived metrics in the objective function [21].

Once text gets generated by a model, it needs to be evaluated both in training and at inference time to have an understanding of the model performance. The gold standard would be a bespoke human evaluation, but this has many challenges, including the bottlenecks of speed, cost, and scale. The machine learning community designed automatic metrics to solve some of these challenges. Existing metrics can be broadly categorized into:

- **n-gram matching**: count the number of n-grams that occur in the reference and candidate text (potentially with reweightings) [1, 20, 27, 36].
- **embedding matching**: compare the soft similarities of dense, fuzzy representations for words [7, 22, 23, 39] and sentences [9] .
- **learned metrics**: metrics are trained to optimize correlation with human judgments [24, 31, 33].
- **radiology-specific metrics**: rule-based parsing to identify the presence or absence of 14 specific diagnoses [15].

Evaluation metrics are meant to be proxies for bepoke human evaluations, therefore any new metric that is developed is measured to demonstrate its correlation with some form of human judgment. The original BLEU paper had human evaluators score candidate translations from 1 (very bad) to 5 (very good) and then demonstrated that BLEU ranks 5 systems in the same order that human annotators do [27]. CIDEr was created in 2014 for image captioning, and became very popular for benchamrking on the MS COCO dataset [36]. Annotators were asked to decide which of two candidate captions better agrees with a reference caption, and then CIDEr was shown to agree with the annotator rankings more strongly than previous metrics like BLEU and METEOR. Several papers have proposed moving away from correlation and more towards multidimensional evaluation where sentence corruptions are introduced as "unit tests" for how well the metric responds [2, 17].

Many of the most popular metrics have been widely criticised [29]. In the context of text simplification, BLEU has a very weak – and in some cases, negative – correlation with human judgment on grammaticality, meaning preservation, and simplicity [34]. Even in the context of machine translation (for which it was originally created), BLEU correlates poorly with human judgment on both adequacy (i.e. whether the hypothesis sentence fully captures the meaning of the reference sentence) and on fluency (i.e. the quality of language in the hypothesis sentence) [5].

Table 1: Linguistic characteristics of PASCAL-50S (CIDEr's annotations) and the MIMIC-CXR radiology reports.

Average Characteristic	PASCAL-50S	MIMIC-CXR
number of references	50.00 ± 0.0	1.00 ± 0.0
sentence count	1.00 ± 0.0	5.29 ± 1.9
word count	9.82 ± 3.2	55.25 ± 25.2
Dale–Chall readability score	5.23 ± 3.2	9.61 ± 1.0
Gunning-Fog index	11.20 ± 4.7	20.06 ± 2.8
Flesch readability score	96.08 ± 15.6	63.26 ± 12.5

2.2 Challenges in the Clinical Domain

Although these metrics have been used in evaluating radiology report generation [21], they are often designed for specific contexts with underlying assumptions about the number of reference sentences available as well as the complexity of the sentences to be analyzed. When CIDEr was introduced, its authors showed that "humans and CIDEr agree with a high correlation," but they did so when there were 20-50 reference captions per images as well as very low-complexity descriptions (e.g. "a cow is standing in a field"). It is not clear whether these findings would hold in the clinical domain, where there is: only one reference report, many more tokens, and a strong emphasis on factual correctness.

To better understand the differences between the simple general domain image descriptions and clinical text, we use standard readability scores to assess the complexity of a given piece of text. The Dale–Chall readability formula and Gunning-Fog index measures the years of formal education a person needs to understand the text on the first reading [6, 12]. In the case of Gunning-Fog, the score is meant to directly indicate the number of school years (e.g. 7 means 7th grade, 12 means 12th grade, etc) and Dale-Chall works similarly but on a 1-10 scale. The Flesh readability score rates documents on a 100-point scale based on the number of words and sentence and syllables per word [10]. Unlike the previous two indices,higher Flesch scores indicate easier-to-read documents.

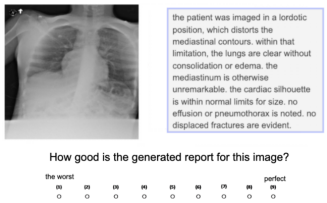
Table 1 demonstrates many of the differences between PASCAL-50S (a dataset introduced in the CIDEr paper) and MIMIC-CXR. We observe that PASCAL-50S indeed has 50 reference reports, each of which has 5 times fewer tokens than a MIMIC-CXR report. Additionally, the Dale-Chall and Gunning-Fox readability scores suggest that nearly twice as many yeears of formal education are required to understand radiology reports than simple image descriptions. Clinical text is demonstrably more complicated than the general domain text that previous metrics were developed for.

3 METHODOLOGY

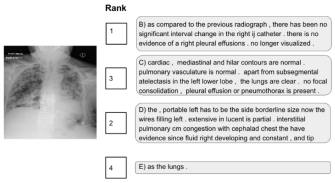
The long-term goal is to eventually design a better evaluation metric for determining whether an automatically-generated radiology report is good. However, such a task requires the close collaboration with clinicians on the design and improvement of the metric. As such, we begin by focusing on the process of this participatory design, and will then apply these lessons learned in a standalone work.

This collaborative effort was done in two parts: qualitative discussions to design the framing/task and a pilot study. Computer

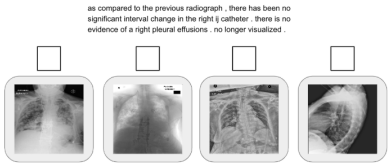
Figure 1: Three different annotation tasks we considered for the radiologists to perform.



(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.

scientists conducted interviews with radiologists from three hospitals, including from Boston, MA and Atlanta, GA. After these conversations, an interdisciplinary team of computer scientists and radiologists conducted a pilot study based on feedback during the initial outreach.

3.1 Designing the Annotation Task

Based on prior work in radiology report generation [3] and evaluation metric creation [27, 36], we had a rough idea of the collaboration and data collection we had in mind: radiologists need to read generated reports and decide whether it is good or not.

The simplest way to go about this could be to display an image + report and ask the radiologist to rank how good it is from 1-to-5. Unfortunately, this approach suffers from broader design issues: behavioral economics demonstrates that humans can be inconsistent. We can see an example of this from the Sentences Involving Composition Knowledge (SICK) dataset [25], where prior work observes that the same kind of sentence transformations can be scored inconsistently [2]:

(1) A man is holding a frog had a 2.1/5 similarity with There is no man holding a frog.
(2) A man is playing soccer had a 4.8/5 similarity with There is no man playing soccer.

Although some of these annotator calibration concerns can be solved with mean normalization, ensuring that a particularly harsh annotator doesn't distort the average, the larger problem is that annotators are not only inconsistent with one another, but they can also be inconsistent with themselves depending on their context and priming [35].

With this in mind, we explored a few potential ways to pose the annotation task for doctors. Figure 1 demonstrates a few ways to ask annotators to make judgments, such as a ranking-based approach (1b) or image selection (1c).

3.2 Interviews with Radiologists

In order to reach out to radiologists to discuss this project, we created a "1-pager" to send to them before our call, inspired by the *Collabsheets* list of "simple" questions for computer scientists and clinicians to discuss [30]. The 1-pager is shown in Figure 3, and its purpose of the document is to give a background on where we are coming from and focus the conversation on the kinds of questions that seemed important to us.

On many questions, there was a strong consensus among the radiologists. They all agreed that clinical correctness is the most important factor in determining whether a report is good. Additionally, each radiologist talked about their field's move towards more structured, templated reports, with some suggesting that perhaps an evaluation metric should try to favor regularity. Additionally, there was overall agreement that for an annotation task like this (where they were not being asked to write their own reports) it might be nice to have a DICOM image viewer that could allow them to zoom, adjust contrast, etc. but such functionality would not be necessary.

During the course of the interviews, there were a few other concepts raised by radiologists which we had not considered when designing the 1-pager in Figure 3, including:

- Many images are simply "normal heart, normal lungs, etc." We should purposefully select a diversity of diagnoses in the annotation set.
- When designing a metric eventually, it may be useful to look for words conveying levels of uncertainty (e.g. "consistent with" vs "suggests").

There was, however, some disagreement among the clinicians.[1] One radiologist questioned whether any of the proposed annotation tasks were the most meaningful thing to measure: they suggested perhaps we should create an interface where the generated report is a "first draft" for the annotator to modify until they are satisfied with the final product. This would, however, be a much more involved task for our annotators. Additionally, radiologists disagreed over whether it was worth including background information about the patient (e.g. "51 y/o female suffering from cough").

3.3 Pilot Study: Annotation Task

Based on the feedback from initial conversations, we conducted a data collection pilot study. Two radiologists annotated 100 images (400 captions) a piece. Figure 4 demonstrates one instance of this task: for a given image, radiologists needed to rank 4 possible

[1] Interestingly, the notion that different doctors could disagree on healthcare expert opinions was surprising for some computer scientists. A helpful analogy for how experts in a field have different opinions is to consider the strong opinions computer scientists have on Bayesian vs Frequentist statistics. No field is a monolith.

Figure 3: This "1-pager" document was sent to radiologists during outreach when setting up initial conversations about this project's goals.

Evaluation Metric for Automatically-Generated Radiology Reports

Overview

With advances in deep learning and image captioning over the past few years, researchers have recently begun applying computer vision methods to radiology report generation. Typically, these generated reports have been evaluated using general domain natural language generation (NLG) metrics like CIDEr and BLEU. However, there is little work assessing how appropriate these metrics are for healthcare, where correctness is critically important.

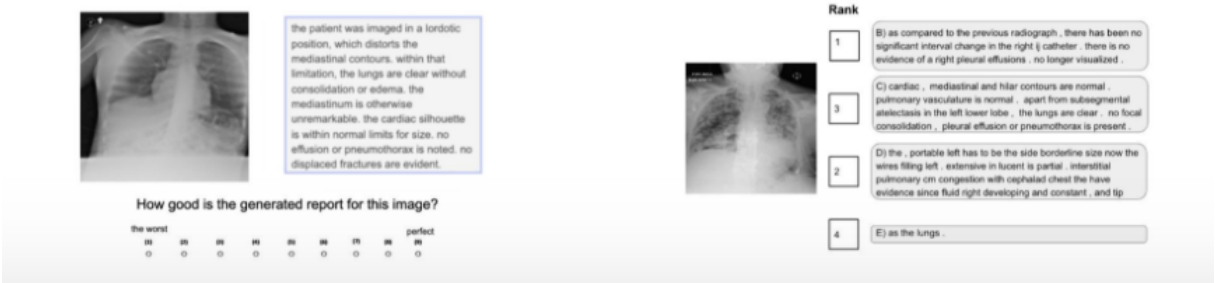
Many works have shown that language generation metrics can give incorrect or unintuitive results, suggesting the need for a more principled evaluation of these methods. In this work we are interested in benchmarking current image captioning metrics. We want to understand their failure modes in order to inspire further work on clinically accurate language generation metrics.

The key to this step of the project is for radiologists to assess the output of machine learning systems. See Figure 1 for a proof of concept for how we could ask for clinical judgment.

Open Questions:

- What is the most sensible way to get clinical judgment?
  - Option 1: {1 image, 4 reports} and rank candidate reports from best to worst
  - Option 2: {1 report, 4 images} and pick which image you think is being described
  - Option 3: {1 reference report, 3 candidate reports} pick candidates most similar to reference text
- Should we be trying to closely mimic the experience that doctors are used to?
  - e.g. Show image in interactive dicom viewer instead of jpeg?
- Should we display additional information as well?
  - Example 1: with every image, the context of "51 y/o female suffering from cough"
  - Example 2: run each report through CheXpert sentence labeler and present the alleged diagnoses alongside the images
- How clear should we be about a ranking task? What makes one report better than another?

Figure 1: Different ways to pose the annotation task to radiologists to determine whether a caption is good or bad. On the left, we ask the simplest version "Here is an image and a proposed caption, please rate how good the caption is from 1-to-10." On the right, we use a ranking-based task, where the doctor will decide which captions are better than other captions.

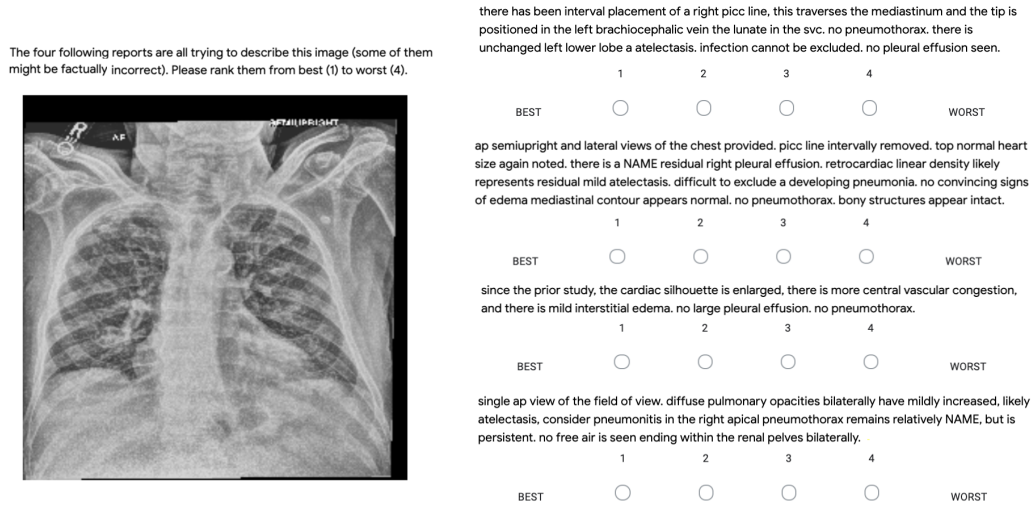


reports based on how well they describe the findings of the image. Each radiologist viewed the same images in the same order.

For each image, we presented the annotator with the following statements:

(1) "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)."
(2) "Briefly describe how you arrived at this ordering (a few simple bullet points is fine)"

Figure 4: An example HIT of the chosen annotation task.



(3) "Confidence that another radiologist would arrive at the same choice for best report (1=Not confident at all, 5=Very confident)"

For each image, we generate four different reports using the following methods: reference, 3-gram, nearest neighbor (1-NN), and random-report. The "reference" report refers to the actual report written by the radiologist, logged in the EHR. The nearest neighbor (1-NN) report is produced by returning the report of the closest image (in the DenseNet-induced feature space) training set. Similarly, the "random-report" is the associated report of a *randomly* selected image from the training set. Finally, the "3-gram" is produced by retrieving the 100 closest images and fitting a tri-gram model from their associated reports.

3.4 Performance of Evaluation Metrics

Ranking the four reports for a single image produces 6 comparisons (i.e. best > the other three, the second best > the other two, the third best > the worst), though 3 of those comparisons involve the reference sentence. To determine how strongly a given metric agrees with radiologist judgment, we compute the specific metric score for each of the 3 non-reference candidates[2] and determine the number of pairwise comparisons where the metric agrees with the experts.

For the metrics, we evaluate many evaluation metrics, including: baselines (random-score, length), readability scores (Dale-Chall), n-gram (BLEU, CIDEr), embedding (BERTScore), and CheXpert accuracy.

4 RESULTS

For 100 images, ranking 4 reports results in 600 binary comparisons. Of those 600 comparisons, the annotators agreed with each other on 459 (i.e. 76.5% of the time).[3] Of the 300 rankings which did not include the reference report, radiologists agreed on 199 rankings (i.e. 66% of the time).

Table 2: 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
ROUGE-1	56.3%	1%
CIDEr	58.8%	0
BERTScore	**61.3%**	0
chexpert-accuracy	43.7%	24.6%
chexpert-accuracy + .001*CIDEr	57.3%	0.5%

In line with prior work [3], the clinical correctness[4] of the 3-gram model (0.353) is higher than the random-report model (0.319) but the nearest neighbor achieves the highest level (0.437).

Table 2 shows how often each metric agreed with consensus radiologist rankings. We include the "random-score" metric (not to be confused with the "random-report" method) as a sanity check: if the metric were randomly assigning numbers, it would get the ranking correct 50% of the time. The "Percent Ties" column denotes how often a given metric was not able to pick either report (e.g.

[2]We do not compute the metric on the reference because, by definition, it would score 100% as you would be comparing the reference against itself.

[3]There were 5 data entry errors where two reports were given the same ranking (e.g. ranking of 1,2,3,3 or 1,2,4,4) even though the task did not allow for ties. For those five entries, the authors used the given explanations to infer what the annotator meant and break the ties.

[4]defined as the macro-average recall of CheXpert labels

Table 3: Top n-grams from the explanations provided by annotators for decision-making. Phrases containing stop words were removed.

unigram	Count
"factually"	16
"all"	17
"wrong"	18
"not"	21
"correct"	24

bigram	Count
"even though"	6
"most correct"	7
"hard to"	9
"factually wrong"	10
"all but"	13

trigram	Count
"are factually wrong"	3
"one and two"	3
"not sure if"	4
"all but one"	5
"is hard to"	6

Table 4: Readability scores of the 100 ngram-generated reports vs the 100 random-report candidates.

Average Characteristic	random-report	3-gram
Dale–Chall readability score	9.61 ± 0.9	9.10 ± 1.5
Gunning-Fog index	20.03 ± 2.6	19.2 ± 3.7
Flesch readability score	63.66 ± 13.0	65.90 ± 17.3

if two reports were each correct on 9/14 findings, then chexpert-accuracy would be tied at 64% a piece). Because this is so common for chexpert-accuracy metrics, we also report how well it would perform when CIDEr is used to break the ties (i.e. *+ .001*CIDEr*) BERTScore attains the top performance of 61.3%; CheXpert only achieves 57.3%.

5 DISCUSSION

In this section, we explore unexpected or interesting results in an attempt to better understand how they arose.

5.1 Self-Reported Annotator Rationale

We were curious to test what would happen when radiologists needed to rank reports based on criteria which included style, factual correctness, grammar, and potentially other factors.

One of the annotators qualitatively described their process for completing the ranking task in an interview. They made it clear that the top criteria is factual correctness: "It doesn't matter how nice or brief a report is. If it's factually wrong, then it's bad." To rank the candidates, they would do an initial pass to group reports into two buckets: "plausible" and "wrong." From there, they would look at each bucket and identify which errors were more egregious (e.g. the report with a rare type of error was ranked worse than a report with a common type of error). Whenever two reports were both plausible without any disqualifyingly bad mistakes, they would look to see which one was more complete, especially since some omissions (e.g. failure to mention a lung lesion) would be more glaring than others.

Based on the quantitative results, the other annotator seemed to agree about the importance of correctness. After each image's ranking, annotators were asked to "Briefly Describe How You Arrived at This Ordering (a few simple bullet points is fine)." Table 3 depicts the top-5 most frequent unigrams, bigrams, and trigrams of the rationales experts gave in response. We can see through uses of phrases like "correct", "wrong", "factually", and "are factually wrong" that they are explaining their decisions as decisions of factual correctness. Additionally, they convey the challenges of comparing two non-perfect candidates through phrases like "most correct" and "all but one."

In both the qualitative and quantitative analyses, there was little to no discussion of readability or grammaticality.

5.2 Why Do Radiologists and CheXpert Disagree on 3-grams?

One surprising part about these "factual correctness"-based explanations is that Table 2 shows the Dale-Chall Readability Index agreed with the radiologists (58.3%) more often than any of the CheXpert-based metrics (57.3%). This finding continues the discrepancy initially discovered in the CXR-Baselines paper where 3-gram outperformed the random-report model on CheXpert's clinical correctness but underperformed on standard evaluation metrics [3].

We can see in Table 4 that ngram-based reports tend to have higher variance of readability, suggesting there may be especially difficult-to-read reports which are still coherent enough for CheXpert's rules to parse correctly. It may be the case that especially ungrammatical reports came across as "non-sensical" to annotaors, which *could* be considered part of "correctness."

6 LIMITATIONS AND FUTURE WORK

Many of the reports in the dataset were essentially incomplete, because they did not capture everything a radiologist would have when really performing their work. Over half of all reports in the annotation pilot referenced previous radiographs, which meant annotators needed to make their best guesses about unseen data. Additionally, although we only show the frontal chest x-ray, reports are usually written using multiple views, such as frontal and lateral.

On top of that, because there are so many diagnostic labels, it can be hard to know the focus of the report; in the true data generation process, the radiologist knows the reason for the exam (e.g. "check ETT tube placement"). This could lend itself more naturally to a question-answering task, where the reason is the "question," the radiograph is the "corpus," and the report is the "answer."

Another limitation encountered in this work is that the caption generation methods were very simple. When radiologists were completing the survey, at least one of them performed an initial "plausible" or "wrong" screening to filter out obviously inappropriate reports. In the followup study we hope to run, we will also use more advanced report generation techniques. Though we will continue to use random-report to serve as a strong baseline for

whether doctors can identify when a report looks good but is actually irrelevant to the context.

7 CONCLUSION

In this work, we performed a pilot study to assess clinician judgment for what makes one radiology report better than another one. The gold standard of a report's "good-ness" would be how well it improves outcomes for the patient or hospital: perhaps it catches more illnesses than a bad report would, or perhaps it saves time/money for hospital operations. Of course, one cannot run a randomized control trial because a bad model would result in significant harm to patients. That is why the clinical domain needs an appropriate metric to serve as a proxy when the true outcome itself cannot be obtained. The difficult challenge is in determining whether a proxy is appropriate; BLEU is a proxy, but the field has spent over a decade pointing out many flaws it has.

In addition to the clinical content of this work, the second main contribution of this study is in the collaborative process itself. Many domains (e.g. healthcare, criminal justice, social science, etc) are getting a lot of attention from computer scientists, but meaningful progress can only be made through meaningful engagement with the domain experts and (when possible) the stakeholders. In this work, we emphasize the tools we used for outreach and the conversations for co-designing a refined version of this task based on lessons learned from the pilot study.

ACKNOWLEDGMENTS

The authors would like to thank Geeticka Chauhan, Ramya Nagarajan, Dr. Catherine D'Ignazio, and Michael Noseworthy for their input and suggestions. The authors also thank Dr. Judy Wawira and Dr. Joanne Shepard for their insight and suggestions. This research was funded in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374.

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[2] William Boag, Renan Campos, Kate Saenko, and Anna Rumshisky. 2016. MUTT: Metric Unit TesTing for Language Generation Tasks. In *ACL* (Berlin, Germany).

[3] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alsentzer, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Baselines for Chest X-Ray Report Generation. Machine Learning for Health workshop at NeurIPS.

[4] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2019. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441* (2019).

[5] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

[6] Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: the new Dale-Chall readability formula*.

[7] Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2748–2760.

[8] 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 (2016), 304–310.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Rudolph Flesch. 1948. A new readability yardstick. , 221–233 pages. https://doi.org/10.1037/h0057532

[11] William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. 2018. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340* (2018).

[12] Robert Gunning. 1968. *The Technique of Clear Writing*.

[13] Grant Haskins, Uwe Kruger, and Pingkun Yan. 2020. Deep learning in medical image registration: a survey. *Machine Vision and Applications* 31, 1 (2020), 8.

[14] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. 2018. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615* (2018).

[15] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 590–597.

[16] Alistair EW Johnson, Tom J Pollard, Seth 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. *arXiv preprint arXiv:1901.07042* 1, 2 (2019).

[17] Hassan Kané, Yusuf Kocyigit, Pelkins Ajanoh, Ali Abdalla, and Mohamed Coulibali. 2019. Towards Neural Similarity Evaluator. (2019).

[18] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. 2017. Deep learning applications in medical image analysis. *Ieee Access* 6 (2017), 9375–9389.

[19] Matthew Lai. 2015. Deep learning for medical image segmentation. *arXiv preprint arXiv:1505.02000* (2015).

[20] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[21] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically Accurate Chest X-Ray Report Generation. In *Proceedings of the 4th Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 106)*. PMLR, Ann Arbor, Michigan, 249–269. http://proceedings.mlr.press/v106/liu19a.html

[22] Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the second conference on machine translation*. 589–597.

[23] Chi-kiu Lo. 2019. YiSi-A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. 507–513.

[24] Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined mt metric based on direct assessment—casict-dcu submission to wmt17 metrics task. In *Proceedings of the second conference on machine translation*. 598–603.

[25] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*. 216–223.

[26] Mehdi Moradi, Ali Madani, Yaniv Gur, Yufan Guo, and Tanveer Syeda-Mahmood. 2018. Bimodal network architectures for automatic generation of image annotation from text. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 449–456.

[27] 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* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[28] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

[29] Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics* 44, 3 (2018), 393–401.

[30] Shems Saleh, William Boag, Lauren Erdman, and Tristan Naumann. 2020. Clinical Collabsheets: 53 Questions to Guide a Clinical Collaboration. In *Proceedings of the 5th Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 126)*. PMLR.

[31] Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. 751–758.

[32] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2497–2506.

[33] Miloš Stanojević and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. 414–419.

[34] Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995* (2018).

[35] Richard H. Thaler and howpublished = Penguin Books year = 2009 isbn = 9780143115267 Cass R. Sunstein, title = Nudge: Improving Decisions About Health, Wealth, and Happiness Revised Expanded Edition. [n.d.]. .

[36] R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

[37] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. 2019. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 292, 1 (2019), 60–66.

[38] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

[39] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).