PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

REPORT GENERATION FROM CHEST X-RAYS: ANALYSIS OF NLP METRICS AND CLINICALLY CORRECT TEMPLATE-BASED MODEL

PABLO PINO

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:
DENIS PARRA

Santiago de Chile, March 2022

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

REPORT GENERATION FROM CHEST X-RAYS: ANALYSIS OF NLP METRICS AND CLINICALLY CORRECT TEMPLATE-BASED MODEL

PABLO PINO

Members of the Committee:
DENIS PARRA
ALVARO SOTO
JOCELYN DUNSTAN
ANGEL ABUSLEME

Thesis submitted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science in Engineering

Santiago de Chile, March 2022

ACKNOWLEDGEMENTS

This work was partially funded by the National Agency for Research and Development ANID, Millennium Science Initiative Program, Code ICN17 002, as well as by ANID, Fondecyt grant 1191791, and by ANID / Scolarship Program / Beca Magíster Nacional / 2020 - 22201476.

First, I would like to thank my advisor Denis Parra. I have worked with him for many years, I have learned a lot, he has contributed to a large extent to the development of my career, and has been very supporting with me when I needed. Also many thanks to all collaborators that we have worked with throughout this project, especially to Cecilia and Claudio, from the Department of Radiology of the School of Medicine at Pontificia Universidad Católica de Chile. Their feedback has been essential to help us better understand the problem from the radiologists' point of view, in the hope that we can close the gap between artificial intelligence and medicine in the (near) future.

I would also like to acknowledge the assistance of Pablo Messina throughout this project, who is currently working on his PhD thesis also regarding medical report generation (though focusing on a VQA approach to tackle the problem), and helped me with many technical issues along the way. And to all the students from the Image Medicine AI group and from the HAIVis group, for everything I have learnt and I keep learning from them.

I am always grateful to my family, for their advice, patience and support in all these years through the university.

Lastly, I want to thank Carolina for her love and support through these years. She has encouraged me to keep moving forward when I most needed it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Every year radiologists face an increasing demand of image-based diagnosis from patients, and computer-aided diagnosis (CAD) systems seem like a promising way to alleviate their workload. In recent years, many authors have proposed deep learning models to generate reports from medical images, but they mainly focus on improving Natural Language Processing (NLP) metrics, such as BLEU and CIDEr, which may not be suitable to measure clinical correctness in the reports, as indicated by multiple authors. Additionally, most approaches are end-to-end black box models that are difficult or impossible to understand by a human, which would make it very hard to implement in a clinical scenario.

In this thesis, we contest the state-of-the-art models and evaluations in the report generation from chest X-rays task. We provide further evidence showing that traditional NLP metrics are not enough to evaluate this task, by showing their lack of robustness in multiple cases. For example, we show NLP metrics are not able to discriminate sentences with opposite clinical meaning, and we show that slightly altering report wording from a model can increase its NLP performance while maintaining high clinical performance. We also propose a template-based report generation model that detects a set of abnormalities and verbalizes them via fixed sentences into a structured report. We benchmark our model in the IU X-ray and MIMIC-CXR datasets against naive baselines, deep learning-based models, and literature models, by employing the CheXpert labeler and NLP metrics. The proposed model is much simpler and inherently interpretable than other state-of-the-art methods, and achieves better results in medical correctness metrics, though worse in NLP. We conclude there is a need to improve the assessment methods in this research area, by analyzing the available data in detail, performing more extensive evaluations and involving expert physicians.

Keywords: medical image report generation, deep learning, templates, diagnostic captioning, chest X-rays.

RESUMEN

Cada año aumenta la demanda por exámenes de imágenes de radiología, y sistemas para diagnóstico apoyados por computador (CAD, por su sigla en inglés) parecen una opción prometedora para aliviar esta carga de trabajo. En los últimos años, muchos autores han propuesto modelos de aprendizaje profundo para generar reportes a partir de imágenes, pero se enfocan principalmente en mejorar métricas de Procesamiento de Lenguaje Natural (NLP, por su sigla en inglés), como BLEU y CIDEr, que pueden no ser apropiadas para medir correctitud médica en los reportes, como han indicado varios autores. Además, la mayoría de las propuestas son modelos de caja negra que son difíciles o imposibles de entender por humanos, lo que dificultaría su implementación en un escenario clínico real.

En esta tesis, analizamos los modelos y evaluaciones usadas por el estado del arte en la tarea de generar reportes a partir de radiografías de tórax. Mostramos evidencia indicando que las métricas tradicionales de NLP no son robustas para esta tarea, por ejemplo, no discriminan bien oraciones que tienen significado contrario en términos médicos, y que se puede alterar levemente la escritura de los reportes de un modelo para subir su rendimiento en NLP, mientras se mantiene su alto rendimiento en términos clínicos. Además, proponemos un modelo basado en plantillas que detecta anormalidades y usa oraciones predefinidas para escribir un reporte estructurado. Evalúamos el modelo en los datasets IU X-ray y MIMIC-CXR, usando la herramienta *CheXpert labeler* y métricas de NLP. El modelo propuesto es más simple e interpretable que otros métodos del estado del arte, y obtiene mejores resultados en métricas de correctitud médica, aunque peores en NLP. Concluímos que se necesita mejorar los métodos de evaluación en esta área de investigación, haciendo evaluaciones más exhaustivas e involucrando a médicos expertos.

Palabras Claves: generación de reportes médicos a partir de imágenes, aprendizaje profundo, plantillas, radiografías de tórax.

1. INTRODUCTION

Writing a report from medical images is an important daily activity for radiologists, yet it is a time-consuming and error-prone task, even for experienced radiologists (Jones et al., 2021). Furthermore, Topol (2019) indicates that the need for diagnosis and reporting from image-based examinations far exceeds the current medical capacity of physicians in the US. Artificial Intelligence (AI) could alleviate this workload by providing computer-aided diagnosis (CAD) systems that can analyze an imaging study and generate a written report, which could be used as a starting point by a radiologist to iterate until producing a final report. For chest X-rays, typically, the radiologists examine one or more images from a patient, indicate if there are abnormalities, describe their visual characteristics, and provide a diagnostic or conclusion (Demner-Fushman et al., 2015; A. E. W. Johnson et al., 2019). Figure 1.1 shows a chest X-ray imaging study and report example from the MIMIC-CXR dataset (A. E. W. Johnson et al., 2019).

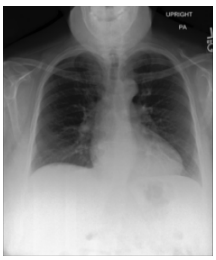
Many deep learning models are proposed in the literature to generate written reports from one or more images (Messina et al., 2020). Most works employ an encoder-decoder architecture, following ideas from the image captioning task in the general domain (Vinyals et al., 2015; Xu et al., 2015), using a CNN-based network as encoder to map the image into a latent space, and the decoder to generate the text. As decoder, there are mainly LSTM-based networks with attention mechanisms (e.g. Boag et al., 2020; Jing et al., 2018, 2019), and Transformer-based networks (e.g. Z. Chen et al., 2020; Lovelace & Mortazavi, 2020). Other approaches replace the decoder by a retrieval approach (e.g. Li et al., 2018; Biswal et al., 2020; Syeda-Mahmood et al., 2020; Kougia et al., 2021).

Despite the advances, it is hard to compare these approaches from a clinical perspective, since they are primarily evaluated by Natural Language Processing (NLP) metrics, such as BLEU (Papineni et al., 2002) or CIDEr-D (Vedantam et al., 2015), and these may not be suitable to measure correctness in the medical domain, as multiple authors suggest

Clinical information

Indication: history: ___ with cough and fever

Comparison: Chest radiograph ___ and chest CT ___



Output report

Findings:
The cardiac silhouette size is top normal. The aorta is mildly tortuous. Mediastinal and hilar contours are unremarkable. Pulmonary vasculature is not engorged. Minimal patchy opacities are demonstrated in the right lower lobe which may be infectious in etiology. Left lung is clear. No pneumothorax or pleural effusion is identified. No acute osseous abnormalities seen.

Impression:
Minimal patchy right lower lobe opacity which is concerning for infection in the correct clinical setting.

Figure 1.1. Report example from the MIMIC-CXR dataset (A. E. W. Johnson et al., 2019), with sections *indication*, *comparison*, *findings* and *impression*. The underscore tokens ("___") represent obfuscated information to de-identify the patients, such as dates and gender.

(e.g. Boag et al., 2020; G. Liu et al., 2019; Pino et al., 2020; Syeda-Mahmood et al., 2020; Pino et al., 2021). Furthermore, some of these metrics have been challenged in general domain tasks, such as machine translation or image captioning (Kilickaya et al., 2017; Reiter, 2018; Mathur et al., 2020; van Miltenburg et al., 2021). To overcome this problem, some authors have used metrics to evaluate the clinical correctness of the generated reports, including the CheXpert labeler (Irvin et al., 2019), MIRQI (Zhang et al., 2020), and other approaches (e.g. Xue et al., 2018; X. Huang, Yan, Xu, & Li, 2019; Alfarghaly, Khaled, Elkorany, Helal, & Fahmy, 2021), although these have not been tested with expert clinicians nor defined as a standard yet.

In addition to dealing with the correctness of the reports, there is the need to apply eXplainable AI (XAI) (Gunning et al., 2019) in a critical domain like medicine, since an application in a clinical setting would have a direct impact on patients (Reyes et al., 2020). The explainability aspect in the report generation task is still understudied (Messina et al., 2020), as most approaches use a post-hoc method to provide a local explanation, typically generating a saliency map indicating the pixels of most importance, using methods such

as Grad-CAM (Selvaraju et al., 2017) for CNN networks or visualizing attention maps for LSTM or Transformer networks. Nonetheless, some authors have argued against using isolated saliency maps as an explanation. For example, Rudin (2019) advocates for using *inherently interpretable models*, systems constrained by domain knowledge so they are transparent for humans to understand, instead of *black-box models with post-hoc explanations*. Ghassemi et al. (2021) argue current post-hoc methods are useful for troubleshoot and audit processes, but not suitable to explain decisions in a clinical setting. Furthermore, they suggest authors should favor testing rigorously their AI systems rather than proposing current post-hoc explanation methods for clinical practice.

In this thesis, we address the task of report generation from chest X-rays, and focus on the clinical correctness of the reports and the inherent interpretability of our model. Specifically, we state two hypotheses to contest:

- **Hyp 1:** traditional NLP metrics are not the most suitable for evaluating this task, compared to clinical correctness metrics.
- **Hyp 2:** a template-based model will be able to outperform state-of-the-art models measured by clinical correctness metrics, while being more understandable or transparent to a human, i.e. more inherently interpretable.

Our main contribution resides on approaching the problem by analyzing the data and evaluations available, and challenging the current state-of-the-art in these aspects. Sambasivan et al. (2021) indicated AI research usually undervalues data centered research, even though data quality is very important in high-stakes applications such as healthcare. Similarly, we believe there is too much emphasis in the models in the report generation task, and data and evaluations are being left out.

1.1. Problem definition

From an AI perspective, the following is the main task addressed by most articles in the literature:

Definition 1.1. *Given as input one or more chest X-rays of a patient, generate an output report as similar as possible to the one written by a radiologist.*

However, we argue that to generate the full report, it would be necessary to have additional context that is critical in clinical tasks (Cabitza, Rasoini, & Gensini, 2017). Thus, we narrow this definition, by focusing only in the content that (1) our model can generate and (2) the evaluations can measure, as will be detailed next.

The typical **report structure** is shown in Figure 1.1, that presents an example from the MIMIC-CXR dataset (A. E. W. Johnson et al., 2019), showing frontal and lateral X-rays, and the report with four different sections. The *indication* section describes the reason to perform the exam, *comparison* mentions previous patient exams, *findings* mainly describes the abnormalities present in the images, and *impression* summarizes the findings and may provide or suggest a diagnostic (Demner-Fushman et al., 2015; A. E. W. Johnson et al., 2019). Typically, the physician asking for the imaging exam is the primary care physician or a medical specialist. Then, to review the images, the radiologist receives the patient clinical information in the *indication* and *comparison* sections, and write their findings in *findings* and *impression*. Most authors in the report generation literature choose one or both of *findings* and *impression* to be generated automatically from the images, as indicated in the survey by Messina et al. (2020).

We **narrow the problem** considering four main aspects. First, the *comparison* and *indication* sections of the main datasets (Demner-Fushman et al., 2015; A. E. W. Johnson et al., 2019) are rather scarce, and most works do not take them into account (Messina et al., 2020). Hence, we rule out clinical information as additional input, even though it could

be very important in a real scenario, as mentioned by some authors in the medical domain (Oakden-Rayner, 2020; Summers, 2021). Additionally, radiologists may refer to clinical information in the *findings* and/or *impression* sections, for example, by directly mentioning past exams (e.g. *"comparison with previous exam"*), or by indirectly mentioning a past condition (e.g. *"size unchanged"*). Consequently with the inputs, we consider these mentions as information that is *out of reach* for the model, i.e. is considered impossible to be generated with the available information.

Second, radiologists may suggest a secondary exam to the primary doctors given the initial findings (e.g. *"recommend follow-up CT abdomen or CT torso for further evaluation"*, *"opacities concerning for pneumonia, follow-up in four weeks"*). We argue that to generate these suggestions the radiologists require clinical information, multiple views or exams if available, and domain knowledge that cannot necessarily be inferred from the images alone. Therefore, this information is also considered *out of reach*, even though complementing with more imaging exams and giving recommendations to the referring physician is an important step in a clinical scenario (Lukaszewicz, Uricchio, & Gerasymchuk, 2016).

Third, frontal views display much more important information regarding abnormalities than lateral views (Rodrigues & Qureshi, 2014), in particular, most abnormalities can still be detected if analyzing only frontal views, but not if analyzing only lateral views. Consequently, we define the frontal view as a mandatory input, either postero-anterior (PA) or antero-posterior (AP) projection, and the lateral view is optional.

Lastly, we argue that the *findings* section contains more descriptive information, in contrast to the *impression* section that summarizes, make inferences or conclusions, which may need additional patient information. Thus, we decide to generate the *findings* section of reports, since it is more likely to be generated with information from the input image only.

To sum up, for this thesis we **define the task of report generation from chest X-rays** as the following:

Definition 1.2. *Given one or more frontal, plus optionally one or more lateral images of a patient, automatically generate the findings section of a report, focusing on describing the abnormalities that are ascertainable on the images alone.*

The problem does not consider additional patient information as input, such as clinical history, previous images or symptoms, following most of the literature and data availability. The output report focus on describing the abnormalities that can be inferred from the image only, consequently with the model inputs. Other information in the reports, such as references to previous exams, suggestion of follow-up exams, prognosis or predictions of health conditions, are considered *out of reach*, i.e. impossible to generate with the available input.

1.2. Outline

The thesis is composed by seven chapters, including this one. Chapter 2 *(Background and Related Work)* defines background concepts regarding medical and machine learning topics, describes the datasets, metrics and models used in the related work. Chapter 3 *(Proposed Method)* states our proposed template-based method, Chapter 4 *(Materials)* describes the materials used in our experiments, and Chapter 5 *(Results)* show the experiment results and main analyses. Lastly, in Chapter 6 *(Discussion)* we discuss the results in terms of the proposed hypotheses, and also state limitations and possible avenues of future work, and Chapter 7 *(Conclusion)* closes with our main conclusions.

2. BACKGROUND AND RELATED WORK

There are several surveys reviewing the medical report generation task (Allaouzi et al., 2018; Pavlopoulos et al., 2019; Monshi et al., 2020; Messina et al., 2020; Ayesha et al., 2021; Kaur et al., 2021), which mainly describe the datasets, metrics and deep learning models employed, and discuss limitations and challenges in this area. We believe the survey by Messina et al. (2020) is the most comprehensive one regarding the topics of clinical correctness evaluation and explainability, which are both key aspects to this thesis. Hence, in this thesis we use this survey as the main reference for the related work, and we additionally consider papers addressing the report generation task that were published after the survey submission date.

The next sections present background definitions and the related work in the medical report generation task. The first (2.1) and second (2.2) sections present relevant medical and machine learning concepts, respectively, the third section (2.3) describes the available datasets, the fourth section (2.4) discusses the metrics used in this task, and the last section (2.5) provides an overview of the deep learning models used.

2.1. Medical background

Diagnosis task: to identify the medical condition of a patient, given background information such as imaging or laboratory exams, clinical information, symptoms, etc.

Prognosis task: to forecast the outcome of a medical condition in a patient in the future.

Imaging exam pipeline. The typical pipeline for a patient that needs an imaging exam is as follows. The *referring physician*, usually the primary doctor or medical specialist, requests an imaging exam for a patient, indicating the main reason, such as symptoms or other conditions. Then, the image is analyzed by a *radiologist* that writes a report

indicating the findings observed from the study. The written report is the main instrument for communication between the radiologist and referring physician (Lukaszewicz et al., 2016). The context of the patient is usually one of: *emergency*, *inpatient* (hospitalized) or *outpatient* (non-hospitalized); and accordingly, the reports in each of them may describe different findings and conditions.

Internal vs external validation. In healthcare related literature these terms are usually defined as follows. **Internal validation:** evaluation with data from the same original source; splitting a dataset in training-validation-test splits falls into this category. **External validation:** evaluation with data from a different source, typically a different hospital or clinical facility.

Chest X-ray technical factors when capturing and reviewing chest X-rays. The two main patient positions are frontal and lateral. For a frontal X-ray, the two possible projections are **Postero-Anterior (PA)** and **Antero-Posterior (AP)**, the former is the standard, provides a better image, and requires the patient to be standing facing the X-ray receptor (Rodrigues & Qureshi, 2014); the latter is typically used for hospitalized patients. Other technical factors include the rotation and inspiration of the patient, among others, (Rodrigues & Qureshi, 2014), which may affect the overall quality of the assessment. In general, the chest X-ray is a somewhat limited exam since the images are a 2-dimensional representation of the 3-dimensional body, implying there may be important attenuation of the image, making interpretation more difficult (Jones et al., 2021).

2.2. Machine Learning background

Common image-based tasks. We introduce some machine learning tasks that receive an image and make a prediction. **Binary classification:** predict one of two classes, usually identified as positive or negative. **Multi-label (binary) classification:** make a binary classification for N existing labels. **Regression:** predict a numeric value. **Segmentation:**

predict the presence of objects at pixel-level. **Image captioning:** generate a natural text description of the image.

Additionally, throughout this thesis we use common **classification metrics**, such as accuracy, precision, recall, F1-score and ROC-AUC (Witten, Frank, & Hall, 2011).

Natural Language Processing (NLP) is an area that explores how to analyze, understand or manipulate natural text using computer systems. Natural Language Generation (NLG) is a sub-area that specializes in tasks that generate natural text from different types of data, such as *image captioning*, *machine translation* (translate natural text from one language to another), and more. There are automatic metrics used in this sub-area to evaluate the quality of the generated text, such as BLEU (Papineni et al., 2002) or CIDEr (Vedantam et al., 2015), which we refer interchangeably as **NLP or NLG metrics** throughout this thesis.

2.3. Datasets

A dataset for the image-based medical report generation task consists in a set of medical images taken from patients alongside a report written by a radiologist. Messina et al. (2020) identified at least 18 datasets covering multiple image modalities and body parts, though many of them are not publicly available, and most research efforts focus on two chest X-rays datasets in English: IU X-ray (Demner-Fushman et al., 2015) and MIMIC-CXR (A. E. W. Johnson et al., 2019). Consequently, in this thesis we mainly focus on the chest X-ray image modality, since there is more data and research available. There are also chest X-ray datasets that only provide abnormalities tagged for each image, i.e. *classification datasets*, which are commonly used for pre-training, auxiliary tasks and other strategies (Messina et al., 2020). Relevant chest X-ray datasets are listed in Table 2.1.

The next subsections describe the most relevant chest X-ray datasets, report (2.3.1) and classification (2.3.2) categories, and lastly mention datasets from other modalities or body parts (2.3.3).

Table 2.1. Chest X-rays datasets used in the literature. Labels were automatically annotated from the reports using automated tools, unless stated otherwise.

Dataset	**# Samples**	**Labels**	**Source**
IU X-ray, Demner-Fushman et al., 2015	Images: 7,470 Reports: 3,955 Patients: 3,955	MeSH and RadLex concepts (manual). MTI tags	Indiana University Hospital Network (US), outpatients only
MIMIC-CXR, A. Johnson et al., 2019	Images: 377,110 Reports: 227,827 Patients: 65,379	14 CheXpert labels	Beth Israel Deaconess Medical Center (US), patients from 2011-2016
PadChest, Bustos et al., 2019, in Spanish	Images: 160,868 Reports: 109,931 Patients: 67,625	297 UMLS concepts (findings, diagnoses, anatomy)	Hospital Universitario de San Juan (Spain), patients from 2009-2017
CheXpert, Irvin et al., 2019	Images: 224,316 Reports: 0 Patients: 65,240	14 CheXpert labels	Stanford Hospital (US), inpatients and outpatients from 2002-2017
ChestX-ray14, X. Wang et al., 2017	Images: 112,120 Reports: 0 Patients: 30,805	14 abnormalities	National Institutes of Health (US), patients from 1992-2015

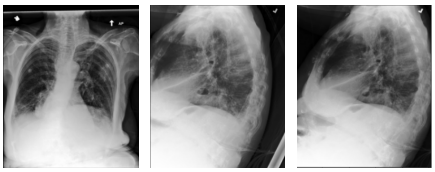
2.3.1. Chest X-ray report datasets

Literature (Messina et al., 2020) indicates the main two datasets used in this task are IU X-ray (Demner-Fushman et al., 2015) and MIMIC-CXR (A. E. W. Johnson et al., 2019); while PadChest (Bustos et al., 2019) has not been widely used yet. The reports usually have four sections: *indication*, *comparison*, *findings* and *impression* (Demner-Fushman et al., 2015; A. Johnson et al., 2019), the former two are provided by the referring physician, and the latter two are written by the radiologist analyzing the image. *Indication* states

Clinical information

Indication: ___ with fever, hypotension. Eval for acute process

Comparison: ___



Output report

Findings:
AP upright and lateral views of the chest provided. Suture material is noted projecting over the left upper lung as on prior compatible with prior resection. There is focal opacity in the right lower lobe and left mid lung, could represent pneumonia though follow-up to resolution advised. There is a retrocardiac opacity containing a fluid level most compatible with a hiatal hernia. No large effusion or pneumothorax. Cardiomediastinal silhouette is unchanged. Bony structures are intact.

Impression:
Right lower and left mid lung opacities, concerning for pneumonia, though follow-up to resolution to exclude underlying mass.

Figure 2.1. Report example from MIMIC-CXR (A. E. W. Johnson et al., 2019) with sections *indication*, *comparison*, *findings* and *impression*. Colors indicate the nature of each sentence, there are sentences describing abnormalities: healthy, uncertain and abnormal in green, orange and red, respectively. In blue there is *out of reach* information (i.e. the model would require more information to generate it), and in light blue referring to technical factors.

the main reason to perform the image study (e.g. symptoms), and *comparison* references previous exams from the patient. In *findings*, the radiologist mainly indicates the presence or absence of abnormalities and describes visual characteristics of the positive findings, such as location, severity, shape, size, among others. Additionally, the radiologist may mention technical factors in the image, may reference previous exams from the patient, and may even suggest additional follow-up exams for the patient. In the *impression* section, the radiologist summarizes the observations into a diagnostic or conclusion, usually in one or a few sentences. See an example in Figure 2.1.

In addition, report datasets have tags or labels annotated from the text, as listed in Table 2.1. For example, IU X-ray (Demner-Fushman et al., 2015) and PadChest (Bustos et al., 2019) are annotated with broad sets of concepts that include anatomical landmarks, abnormalities, diagnoses, and more; namely MeSH (Rogers, 1963), RadLex (Langlotz, 2006), MTI tags (Mork, Yepes, & Aronson, 2013) and UMLS concepts (Lindberg, Humphreys,

& McCray, 1993). MIMIC-CXR (A. E. W. Johnson et al., 2019) is annotated with the set of CheXpert labels (Irvin et al., 2019) that mainly includes specific chest abnormalities (refer to Table 2.2 in the next subsection for the labels). In all cases, authors typically use these labels to train their models with auxiliary tasks.

2.3.2. Chest X-ray classification datasets

The most used classification datasets are the ChestX-ray14 (X. Wang et al., 2017) and CheXpert (Irvin et al., 2019), typically used for pre-training or auxiliary tasks in the report generation literature (Messina et al., 2020). Both contain chest X-rays annotated with 14 labels (see Table 2.2 for the labels), which were automatically extracted from reports, though the reports are not publicly available. Alongside the CheXpert dataset, the authors proposed the CheXpert labeler (Irvin et al., 2019), a rule-based tool to label written reports with the 14 labels as *positive*, *negative* or *uncertain* (refer to section 2.4.2.1 for details on the labeler). Both datasets are imbalanced towards having more negative samples, as commonly occurs in medical related data.

Table 2.2. Abnormality labels in each dataset.

Anatomy	**CheXpert**, Irvin et al., 2019	**ChestX-ray14**, X. Wang et al., 2017
Heart and Mediastinum	*Cardiomegaly*, *Enlarged Cardiomediastinum*	*Cardiomegaly*
Lungs	*Atelectasis*, *Consolidation*, *Edema*, *Pneumonia*, *Lung Lesion*, *Lung Opacity*	*Atelectasis*, *Consolidation*, *Edema*, *Pneumonia*, *Emphysema*, *Fibrosis*, *Infiltration*, *Mass*, *Nodule*
Pleural space	*Pneumothorax*, *Pleural Effusion*, *Pleural Other*	*Pneumothorax*, *Effusion*, *Pleural Thickening*
Other	*No Finding*, *Fracture*, *Support Devices*	*Hernia*

2.3.3. Other modalities report datasets

Literature (Messina et al., 2020) shows there are multiple datasets covering different modalities and body parts beyond chest X-rays. For example, among the publicly available datasets there is INBreast with mammography X-rays (Moreira et al., 2012), PEIR Gross with gross lesions (Jing et al., 2018), STARE with retinal fundus images (Hoover, 1975), ROCO with multiple radiology modalities, such as CT scans, PET scans, Fluoroscopy images, etc.; (Pelka et al., 2018) and the ImageCLEF datasets from their caption challenges, containing varied biomedical images extracted from PubMed Central[1](Eickhoff et al., 2017; García Seco de Herrera et al., 2018; Pelka et al., 2021). Refer to the appendix A for the full list and details. From these, the most used datasets are from the ImageCLEF challenges, for example by recent works in their latest challenges (Schilling et al., 2021; Quintana et al., 2021; Castro et al., 2021; Schuit et al., 2021). Nonetheless, there is much less research addressing the report generation task in those sub-domains, compared to chest X-rays (Messina et al., 2020).

The nature of each of these datasets is very different to chest X-rays. In each subdomain there is a different set of target abnormalities and body parts observed; the images look different, some modalities produce 3D volumes of images (e.g. CT scans); the reports are usually structured differently, by having a different vocabulary, longer or shorter reports; the patient population may differ, and so on. As an example, consider Figure 2.2 showing two samples from the ImageCLEF caption 2021 dataset (Pelka et al., 2021).

2.4. Metrics

The survey by Messina et al. (2020) categorizes evaluation metrics in the report generation task in two main categories: *text quality* measures, which are traditional Natural Language Processing (NLP) or Natural Language Generation (NLG) metrics, and *clinical*

[1] `https://www.ncbi.nlm.nih.gov/pmc/`

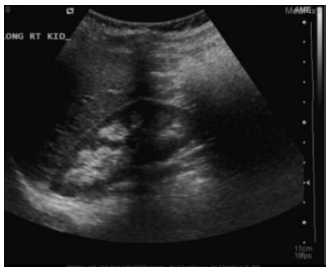
Image	Report
	*Contrast enhanced CT of the abdomen and pelvis demonstrates a large enhancing mass in the left kidney. The mass has multiple foci of fat density. Area of increased density within the mass is consistent with hemorrhage.*
	*Longitudinal ultrasound image of the right kidney demonstrates diffusely hyperechoic medullary pyramids.*

Figure 2.2. Examples from the ImageCLEF 2021 dataset (Pelka et al., 2021), including a slice of an abdominal CT (top) and a kidney Ultrasound (bottom).

correctness measures, which aim to assess the clinical facts stated in the reports. Most works evaluate the report generation performance using only the first category, *NLP metrics*, namely BLEU (Papineni et al., 2002), CIDEr-D, (Vedantam et al., 2015), ROUGE-L (Lin, 2004) and METEOR (Lavie & Agarwal, 2007), which measure n-gram matching between the ground truth and a generated text. These metrics are very popular in machine translation, image captioning and other NLP tasks; although there is growing evidence that they may not be suitable to measure correctness in clinical reports (Boag et al., 2020; G. Liu et al., 2019; Lovelace & Mortazavi, 2020; Pino et al., 2020; Babar et al., 2021; Pino et al., 2021) or even in other tasks (Kilickaya et al., 2017; Reiter, 2018; Mathur et al., 2020; van Miltenburg et al., 2021). To overcome this problem, some authors (e.g. G. Liu et al., 2019; Ni et al., 2020; Pino et al., 2021; Zhang et al., 2020; Biswal et al., 2020; X. Huang et al., 2019; Jing et al., 2019; Xue et al., 2018) also evaluate their models using the second category metrics, *clinical correctness*. The most common approach is

to use the CheXpert labeler (Irvin et al., 2019) or a variation of it, though there are other approaches proposed, and there is still not a defined standard. To the best of our knowledge, none of the clinical metrics have been validated with expert clinicians, but they aim at assessing medical accuracy, unlike NLP metrics.

The next subsections describe in further detail the NLP metrics (2.4.1) and clinical correctness metrics (2.4.2), and then both categories are briefly analyzed (2.4.3) to illustrate why NLP metrics isolated may not be suitable for this task, and clinical metrics appear to be more appropriate. In each metric subsection, the first paragraphs give an overview of the metric, while the next paragraphs detail the calculation and formulas; the latter can be skipped for easier reading.

2.4.1. NLP / NLG metrics

NLG metrics were designed for natural language related tasks in the general domain, such as machine translation, text summarization or image captioning. As such, the metrics were designed to give an score between a *ground truth* (or *reference*) text, and a *generated* (or *candidate*) text. Furthermore, in the general domain the metrics are designed to receive one or more references per sample, to account for multiple ways of re-phrasing a sentence while conveying the same meaning. The next subsections present the details of BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Denkowski & Lavie, 2014), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). The former four are calculated based on n-gram matching, while SPICE does a semantic parsing of the text. All metrics range from 0 (worst) to 1 (best), except for CIDEr-D (the robust variant of CIDEr), which ranges from 0 to 10. The most common library used in Python for calculating the metrics derives from the Microsoft COCO Captions Challenge (X. Chen et al., 2015): `coco-caption`[2].

[2] `https://github.com/tylin/coco-caption`

2.4.1.1. BLEU

Overview. Papineni et al. (2002) proposed Bilingual Evaluation Understudy (BLEU) for machine translation, which is a precision-based metric that evaluates n-gram overlaps from a target text with one or more ground truth texts. For a specific value of n, BLEU-n can be calculated, such as BLEU-1 using up to unigrams, BLEU-2 using up to bigrams, etc. In the report generation task, typically BLEU-n are calculated with values from 1 to 4 (Messina et al., 2020). The BLEU metric is oriented to precision and not recall, thus it measures how consistent is the generated report with the ground truth, but not how much information from the ground truth is being captured or being left out. To compensate for this fact, it includes a penalization for brief candidate sentences in its calculation.

Details. The authors propose calculating a *modified n-gram precision* p_n for each value of n, shown in equation 2.1. The counters i and j sum over all the samples in the corpus, C_i and C_j are candidate sentences, $Count_{C_j}(\text{m-gram})$ is the amount of times that m-gram appears in the candidate C_j, $Count_{C_i \text{ clip } GT_i}(\text{n-gram})$ is the amount of times n-gram appears in the candidate C_i and in the ground truth GT_i, clipped to disallow matching the same n-gram multiple times.

$$p_n = \frac{\sum_{i \in Samples} \sum_{\text{n-gram} \in C_i} Count_{C_i \text{ clip } GT_i}(\text{n-gram})}{\sum_{j \in Samples} \sum_{\text{m-gram} \in C_j} Count_{C_j}(\text{m-gram})} \tag{2.1}$$

To compensate for the precision only orientation, the calculation includes a penalization for short sentences, namely the *brevity penalty* (BP), shown in equation 2.2, where r is the length of the reference and c is the length of the candidate text.

$$BP = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases} \tag{2.2}$$

Lastly, BLEU-N is calculated as the geometric average of the modified precision values up to N, weighted by custom w_n factors. Typically, w_n are uniform (e.g. $w_n = 0.25$ for $N = 4$).

$$\text{BLEU-N} = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.3}$$

2.4.1.2. ROUGE-L

Overview. Lin (2004) presented Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which is a set of metrics to assess text similarity in the text summarization task, namely ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S. Most works in the medical report generation task use the ROUGE-L metric (Messina et al., 2020), which is based on measuring the longest common sub-sequence between the generated and ground truth texts, and it has an hyper-parameter to bias the metric towards precision, recall, or an average of both (F-score). In practice, the score is slightly biased toward recall in the `coco-caption` package.

Details. Let a generated text be a sequence of words $Gen = w_1\ w_2\ \ldots\ w_n$ and a ground truth text be a sequence $GT = r_1\ r_2\ \ldots\ r_m$. As a reminder, by definition a sequence $X = x_1 \ldots x_N$ is a subsequence of $Y = y_1 \ldots y_M$ if all of its elements x_i appear in Y in the same order, though there may be other elements y_j in between. Then, let $LCS(Gen, GT)$ be the length of the longest common subsequence between Gen and GT. Intuitively, if Gen is more similar to GT, the longer the longest common subsequence found will be. Hence, a notion of recall (R_{lcs}) and precision (P_{lcs}) can be computed:

$$R_{lcs} = \frac{LCS(Gen, GT)}{length(GT)} \tag{2.4}$$

$$P_{lcs} = \frac{LCS(Gen, GT)}{length(Gen)} \tag{2.5}$$

Thus, ROUGE-L is calculated as a harmonic average between the two measures (F-score), using a hyper-parameter β.

$$\text{ROUGE-L} = F_{lcs} = \frac{(1+\beta^2) \cdot R_{lcs} \cdot P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{2.6}$$

Notice if $\beta = 1$, the F_{lcs} is exactly the F-1 score; if $\beta = 0$ is the precision, and if $\beta \to \infty$ it approximates the recall. In practice, β is set to 1.2 in the `coco-caption` package.

2.4.1.3. METEOR

Overview. S. Banerjee and Lavie (2005) presented Metric for Evaluation of Translation (METEOR), which was later updated by the same authors (Lavie & Agarwal, 2007; Denkowski & Lavie, 2010, 2011, 2014). METEOR attempts to find a uni-gram matching between the generated and ground truth sentences, and then an F-score of the words matched is computed. The metric includes using synonyms, stemming and paraphrasing sentences to find word matches, and the calculation includes several hyper-parameters, which were tuned by the authors to optimize the correlation with human judgment on a specific machine translation setting and dataset (Denkowski & Lavie, 2014). This may be appropriate for the general domain in some of their cases, but we believe is not necessarily appropriate in the medical domain. In the report generation task, most of the times is not clear which implementation or set of hyper-parameters were used (Messina et al., 2020), making it difficult to compare results.

Details. METEOR creates a matching by linking a token from the generated text with a token from the ground truth, which indicates that they have the same meaning (example in Figure 2.3). The main steps for uni-gram linking are, in order: exact word match, stem match, WordNet synonym (Miller, 1995) match, and paraphrase match (from a list of known paraphrases). Linking the same token twice is not allowed. Notice in some cases multiple matches may be possible between a candidate and reference sentence, so

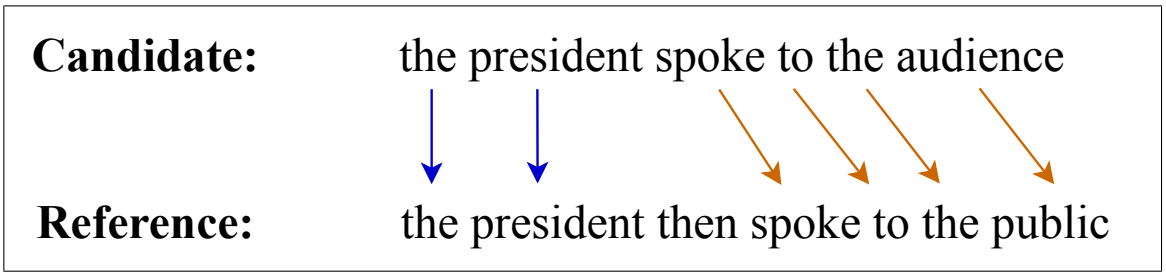


Figure 2.3. METEOR matching example, taken from the original paper (S. Banerjee & Lavie, 2005). Words are matched if they mean the same, and each word from the candidate may be matched to only one word from the reference. In the example there are two *chunks* matched, i.e. a contiguous group of words matched, shown with blue and orange arrows each.

METEOR has specific rules to optimize the matching (refer to section 2 of the paper for details, Denkowski & Lavie, 2014).

Given a match between a generated and ground truth sentences, precision (P) and recall (R) scores can be computed by counting the amount of words matched in each text. The specific calculation of P and R includes hyper-parameters δ, w_{exact}, w_{stem}, w_{syn} and w_{par} regarding the different matching steps (refer to the paper for details). Then, an F-score is proposed with an α hyper-parameter to bias toward precision or recall, shown in equation 2.7. Additionally, a penalization is added to prefer matchings with longer phrases (equation 2.8): where a *chunk* is a contiguous group of uni-grams matched contiguously to the reference (in Figure 2.3, the number of *chunks* is 2), and γ and β are additional hyper-parameters to tune the penalization. Lastly, METEOR is calculated as equation 2.9.

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \tag{2.7}$$

$$Pen = \gamma \cdot \left(\frac{\#chunks}{\#unigrams\,matched} \right)^{\beta} \tag{2.8}$$

$$\text{METEOR} = (1 - Pen) \cdot F_{mean} \tag{2.9}$$

2.4.1.4. CIDEr

Overview. Vedantam et al. (2015) presented Consensus-based Image Description Evaluation (CIDEr) as a metric for the image captioning task. CIDEr represents each sentence with a term-frequency and inverse-document-frequency (TF-IDF) score over its n-grams, where the TF term gives more importance to the presence of each n-gram in the sentence, while the IDF term gives more importance to the more rare n-grams in the dataset, assuming those will provide more valuable information. Two sentences are then similar if their n-gram TF-IDF representations are similar, and the authors argue that this captures both precision and recall notions, and preserves grammatical and semantic aspects and by using multiple n values. They also presented the variant CIDEr-D (Vedantam et al., 2015), which is less susceptible to gaming effects. The original CIDEr ranges from 0 (worst) to 1 (best), and CIDEr-D ranges from 0 (worst) to 10 (best). In the report generation task, most of the authors do not specify which variant is used (Messina et al., 2020), but the implementation from `coco-caption` is CIDEr-D.

Details. To calculate the TF-IDF score for a given sentence s and an n-gram k, consider the following. Intuitively, the TF term will represent the amount of times the n-gram k appears in s, with respect to all the n-grams in s. On the other hand, the IDF term will measure the inverse of how much appears the n-gram k in the whole dataset. Thus, the TF-IDF score, named $g_k(s)$, is roughly calculated as:

$$g_k(s) = TF \cdot IDF \tag{2.10}$$

$$g_k(s) = \frac{\text{\# appearances k in s}}{\text{\# appearances any n-gram in s}} \cdot \log\left(\frac{\text{dataset size}}{\text{\# appearances k in dataset}}\right) \tag{2.11}$$

$$g_k(s) = \frac{h_k(s)}{\sum_{l \in \text{n-grams}} h_l(s)} \cdot \log\left(\frac{\#Images}{\sum_{i \in Images} \sum_{q=1}^{m} h_k(GT_{iq})}\right) \tag{2.12}$$

where $h_y(x)$ is the amount of times the n-gram y appears in the sentence x, and GT_{iq} for $q \in \{1, \ldots, m\}$ are the m ground truth sentences for the image i[3]. Then, given all existing n-grams $k_1, k_2, \ldots, k_M$, a vector $\vec{g}^n(s)$ is composed for each sentence s, in each position containing the TF-IDF score for each n-gram $k_1, k_2, \ldots, k_M$. Lastly, the likeness between two sentences is calculated as the cosine similarity between their two vectors, as equation 2.13 for a specific n, and equation 2.14 to average up until N-grams. N is typically set to 4, using uniform weights $w_i = 0.25$.

$$\text{CIDEr}_n(Gen_i, GT_i) = \frac{1}{m} \sum_{j=1}^{m} \frac{\vec{g}^n(Gen_i) \cdot \vec{g}^n(GT_{ij})}{||\vec{g}^n(Gen_i)|| \ ||\vec{g}^n(GT_{ij})||} \tag{2.13}$$

$$\text{CIDEr}(Gen_i, GT_i) = \sum_{n=1}^{N} w_n \ \text{CIDEr}_n(Gen_i, GT_i) \tag{2.14}$$

Lastly, the same authors presented the variant CIDEr-D that is more robust against gaming effects, by including a penalization for length differences between the sentences, and using a more robust counting mechanism that clips n-gram matches, i.e. disallows matching the same n-gram multiple times.

2.4.1.5. SPICE

Overview. Anderson et al. (2016) proposed Semantic Propositional Image Caption Evaluation (SPICE) as a metric for image captioning, which captures the underlying meaning of the sentences describing an image, by parsing each caption to a graph representing the objects in the scene, relations between them and their attributes. SPICE is then the F-1 score for the match between the generated graph and ground truth's graph, assuming two captions have a similar meaning if and only if their graphs are similar.

Although this metric is essentially different from the previously described n-gram based metrics, SPICE is still categorized as a text quality metric by Messina et al. (2020),

[3]In the medical report generation task, this is limited to one ground truth report, but CIDEr is designed to work with more than one ground truth.

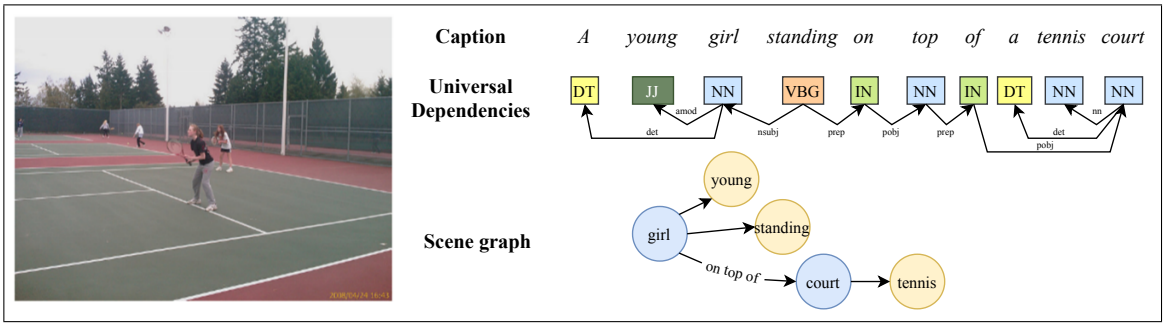


Figure 2.4. SPICE parsing example. The caption is parsed with universal dependencies and part-of-speech tagging, and then a scene graph is generated to represent the image. In the scene graph, blue nodes are objects, yellow nodes are attributes, and the edges may convey a specific meaning between nodes. Example inspired from the original paper (Anderson et al., 2016).

since the objects, relations and automated tools used in the parsing are not necessarily designed for the medical domain. As reported by Kilickaya et al. (2017), the quality of this metric highly depends on the quality of the parsing. From the chest X-rays report generation literature, B. Hou, Kaissis, Summers, and Kainz (2021) is the only work that uses SPICE as a metric, though they do not analyze or make particular comments on the performance measured by this metric.

Details. First, the text is parsed as universal dependencies with a Probabilistic Context-Free Grammar dependency parser (Klein & Manning, 2003), and then parsed to a scene graph with hand-crafted linguistic rules. Each node of the graph represents either an object or the attribute of an object, and each edge provides a meaning between objects or objects and attributes (e.g. *is in*, *looks like*, etc). See a scene graph example in Figure 2.4.

Second, the scene graph of a caption is represented as a set of tuples *(object)*, *(object, attribute)* and *(object1, relation, object2)*, containing all of its nodes and edges. Then, the ground truth and generated scene graphs are matched through their tuple representations in a tuple-wise fashion, and calculating precision, recall and F-1 score, where the SPICE metric is the final F1-score. Similar to METEOR, lemmatization and WordNet (Miller,

1995) synonymy are included to further improve the semantic matching between objects and attributes.

2.4.2. Clinical Correctness Metrics

Clinical correctness metrics are designed by authors addressing the report generation task in a medical sub-domain. Their objective is to capture the health condition of the patient written in the report, specifically, by detecting abnormalities, diseases, pathologies, or a diagnostic mentioned. Unlike NLP metrics, only one *ground truth* report is considered in this case, tailored for medical datasets. Notice a *positive* finding in an abnormality refers to a non-healthy or abnormal patient, while a *negative* finding refers to a healthy patient.

The next subsections describe the CheXpert labeler (Irvin et al., 2019) (2.4.2.1), which is the most commonly used metric in chest X-rays, MIRQI (Zhang et al., 2020) (2.4.2.2) and other approaches (2.4.2.3) found in the literature. From these approaches, there is still no defined standard, and to the best of our knowledge, they have not been evaluated against physicians judgment.

2.4.2.1. CheXpert Labeler

Overview. The CheXpert labeler[4] (Irvin et al., 2019) is a rule-based tool that classifies a set of abnormalities from a written report into *positive*, *negative*, *uncertain*, or *unmentioned*; by using manually curated patterns and a dependency parser to obtain semantic relations between the words and sentences. Then, the abnormality classification is simplified to a binary classification setting: the *unmentioned* class is merged with the *negative* class, as both indicate *the patient is healthy*; and *positive* with *uncertain*, since the latter usually indicates that *an abnormality should not be ruled out*. Lastly, the abnormality classification is used to compute binary classification metrics (precision, recall,

[4]Official implementation: `https://github.com/stanfordmlgroup/chexpert-labeler`.

F-1, ROC-AUC) between a generated and ground truth reports, also known as *CheXpert labeler metrics*.

Thirteen works in the chest X-ray report generation task used the CheXpert metric (G. Liu et al., 2019; Boag et al., 2020; Z. Chen et al., 2020; Nishino et al., 2020; Ni et al., 2020; Lovelace & Mortazavi, 2020; F. Liu, Yin, et al., 2021; D. Hou et al., 2021; Nguyen et al., 2021; B. Hou et al., 2021; Najdenkoska et al., 2021; Miura et al., 2021; Kougia et al., 2021), though to the best of our knowledge the tool has not been tested against expert judgment as a report generation metric. Nonetheless, the authors (Irvin et al., 2019) evaluated the labeling tool against a set of 1,000 reports manually labeled by two board-certified radiologists, achieving F1 scores macro averaged across labels of: 0.948 for *positive*, 0.899 for *negative*, and 0.777 for *uncertain*; demonstrating a good performance, and particularly showing better results than other labelers (e.g. NegBio, Peng et al., 2018).

Some authors have proposed improvements over the CheXpert labeler. McDermott et al. (2020) proposed CheXpert++, which is a BERT-based network trained to predict the 14 labels from the text. Similarly, Smit et al. (2020) proposed CheXbert, and they evaluated it further with a manually labeled test set of their own, showing it has slightly better results than CheXpert. Furthermore, Jain et al. (2021) proposed VisualCheXbert, which improves CheXbert by also using the chest X-rays during training, to address discrepancies they found between images and reports. In the report generation task, some authors used a variation instead of the original CheXpert labeler tool for the evaluation; for example, some used an LSTM model trained to predict the original labels (Lovelace & Mortazavi, 2020; Nguyen et al., 2021), and others used CheXbert (Miura et al., 2021).

Details. The process to label a report consists of three steps, namely extraction, classification and aggregation (Figure 2.5). First, textual mentions of the observations are extracted using pre-fixed patterns manually curated by radiologists. Second, the text is

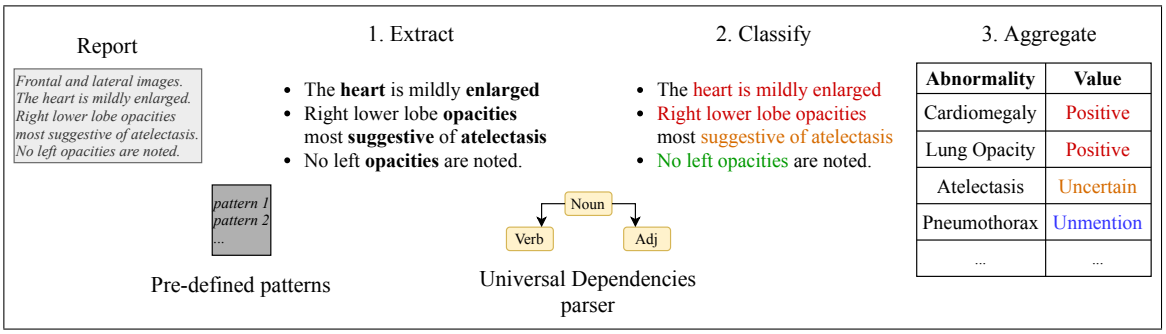


Figure 2.5. CheXpert labeler process comprising three main steps: (1) extracting abnormality mentions using pre-defined patterns, (2) parsing into Universal Dependencies to then classify the mentions, and lastly (3) aggregate the results.

parsed with the Bllip parser (Charniak & Johnson, 2005) trained with a biomedical model (McClosky, 2010), and then the universal dependency graph of the sentence is obtained using the Stanford CoreNLP tool (De Marneffe et al., 2014). With this information, each textual mention is classified into one of three classes, *negative*, *positive* or *uncertain*. Notice the latter captures both uncertainties about the diagnostic itself (e.g. *"diffuse opacity may represent pneumonia"*) or an ambiguous report (e.g. *"heart size is unchanged"*). Third, the classifications are aggregated into a final vector, where each observation can be one of the above classes, or else is considered *unmentioned*. The 14 labels or observations are: *No Finding*, *Enlarged Cardiomediastinum*, *Cardiomegaly*, *Lung Lesion*, *Lung Opacity*, *Edema*, *Consolidation*, *Pneumonia*, *Atelectasis*, *Pneumothorax*, *Pleural Effusion*, *Pleural Other*, *Fracture* and *Support Devices*. Notice all labels except for *No Finding* can be considered abnormalities.

To use as an evaluation metric, the labeler is applied over the generated and ground truth reports, and the outputs are then simplified to a binary setting. The *unmentioned* class is merged with the *negative* class, as both indicate the absence of an abnormality, and the *uncertain* class is merged with the *positive* class, since most of the uncertain cases indicate that *an abnormality should not be ruled out*, as shown in the examples before. Lastly, the findings are evaluated using classification metrics, such as precision, recall,

F-1 score or accuracy. The metrics can be shown disaggregated by each abnormality (e.g. *Lung Opacity* precision and recall), and/or aggregated into an average (e.g. recall macro averaged across 14 labels). Since most datasets are imbalanced towards having less abnormal cases, commonly accuracy is not very meaningful, and precision, recall and F-1 scores are more useful.

2.4.2.2. MIRQI

Overview. Zhang et al. (2020) proposed the metric Medical Image Report Quality Index (MIRQI)[5] to measure correctness in chest X-ray reports in terms of abnormalities and attributes mentioned. MIRQI works similar to the CheXpert labeler plus a few improvements: uses manually designed patterns and a dependency parser to classify a set of 20 abnormalities, and leverages the semantic dependencies to capture fine-grained attributes or *modifiers* related to the findings, such as body parts, abnormality severity, size, shape, etc. Then, both the abnormality classification (*positive*/*negative*/*uncertain*) and the modifiers are used to perform a matching between the generated and ground truth findings, and classification like metrics are calculated, with notions of precision (MIRQI-p), recall (MIRQI-R) and F1-score (MIRQI-F1). Figure 2.6 shows an extraction example; notice there are some general words mistakenly captured as modifiers, such as *present* and *is*.

Details. The process to extract and classify abnormalities is very similar to the CheXpert labeler, though expanding the rules to include 20 abnormalities (refer to the official code implementation for more details). However, the score calculations differ significantly. MIRQI is calculated with notions of precision (MIRQI-p), recall (MIRQI-r) and F-1 score (MIRQI-F1), with slight differences from the common classification metrics. First, the abnormalities and attributes from the ground-truth and generated reports are

[5]Official implementation: `https://github.com/xiaosongwang/MIRQI`.

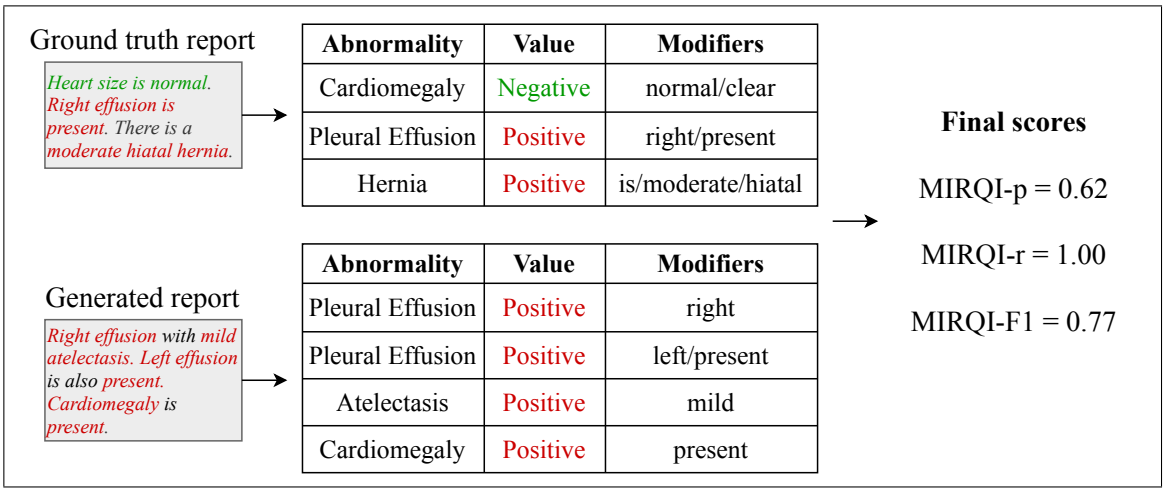


Figure 2.6. MIRQI extraction and matching example. Abnormalities with modifiers are extracted from each report, then matched, and MIRQI scores are calculated.

matched, and the score of *true positives (TP)* is calculated as shown in equation 2.15.

$$TP = (1 - w_{attr}) \cdot TP_{abn} + w_{attr} \cdot TP_{attributes} \tag{2.15}$$

Where TP_{abn} are the *true positives* in terms of *positive* or *uncertain* abnormalities matched (i.e. *positive* and *uncertain* are merged, same as the CheXpert labeler), and $TP_{attributes}$ are the *true positives* modifiers detected (accounting only the matched abnormalities cases). Thus, the true positive TP value accounts for detection of abnormalities and also of their modifiers. w_{attr} is a hyper-parameter set to 0.2 in the original paper (Zhang et al., 2020) (although is set to 0.3 in the original code implementation).

Then, to calculate the final scores both *positive* (abnormal) and *negative* (healthy) predictions are considered, as MIRQI-r takes both recall and specificity into account (equation 2.17); and MIRQI-p takes both precision and the negative predictive value (NPV) into account (equation 2.19). The ture negatives (TN), false positives (FP) and false negatives (FN) are calculated as usual in binary classification. w_{pos} is a hyper-parameter to balance between favoring positives or negative predictions, which is set to 0.8 in the paper (Zhang

et al., 2020). Lastly, the MIRQI metric is defined as the F-1 score shown in equation 2.20.

$$\text{MIRQI-r} = w_{pos} \cdot \frac{TP}{TP + FN} + w_{neg} \cdot \frac{TN}{TN + FP} \tag{2.16}$$

$$= w_{pos} \cdot recall + w_{neg} \cdot specificity \tag{2.17}$$

$$\text{MIRQI-p} = w_{pos} \cdot \frac{TP}{TP + FP} + w_{neg} \cdot \frac{TN}{TN + FN} \tag{2.18}$$

$$= w_{pos} \cdot precision + w_{neg} \cdot NPV \tag{2.19}$$

$$\text{MIRQI} = \text{MIRQI-F1} = \frac{\text{MIRQI-r} \cdot \text{MIRQI-p}}{\text{MIRQI-r} + \text{MIRQI-p}} \tag{2.20}$$

2.4.2.3. Other clinical metrics

Overview. In several works the authors evaluated their model with other methods to assess the clinical performance of the generated reports. Most of the approaches detect abnormalities in the text and use classification metrics for the evaluation, similar to the CheXpert labeler, though each of them using a different procedure. Other simpler approaches use a set of keywords and compare the number of appearances in the ground truth and generated reports with classification metrics. Despite these methods may have been useful in their experiments, the authors do not provide enough details or an official implementation, and they did not perform a formal evaluation of the correlation with expert judgment or similar. Thus, no consensus or standard can be drawn from these proposals yet (Messina et al., 2020). The methods are cited and briefly mentioned next.

Details. Among the *abnormality-based* methods, which attempt to classify abnormalities from the report and then use classification-like metrics: Jing et al. (2019) used manually designed patterns to classify findings; Biswal et al. (2020) used a character-level CNN to classify CheXpert labels; and Moradi et al. (2016) used a proprietary software to extract semantic descriptors. Other approaches are *keyword-based*, which compute the ratio of a set of keywords found between the generated report and ground truth: X. Huang et al. (2019) used MeSH terms (Rogers, 1963); Xue et al. (2018) used 438 MTI terms

Table 2.3. NLP wrong scoring examples: ground truth (GT) and generated (Gen) reports examples with metrics BLEU (B, average BLEU 1-4), ROUGE-L (R-L), CIDEr-D (C-D, ranges from 0 to 10) and CheXpert F-1 (macro-averaged across abnormalities mentioned). Correct and incorrect sentences are in green and red, respectively, while *out of reach* information for the model is in blue.

	Report	**B**	**R-L**	**C-D**	**F-1**
GT 1	Heart size is mildly enlarged. Small right pneumothorax is seen.				
Gen 1	Heart size is normal. No pneumothorax is seen.	**0.493**	**0.715**	**2.34**	0
Gen 2	The cardiac silhouette is enlarged. No pneumothorax.	0.146	0.464	1.09	0.5
Gen 3	Mild cardiomegaly. Pneumothorax on right lung.	0.075	0.289	0.15	**1**
GT 2	Comparison to previous exam. Heart size is enlarged. Dr XXXX was contacted.				
Gen 1	Comparison to previous exam. Heart size is enlarged. Dr was contacted.	**0.655**	**0.846**	**5.03**	**1**
Gen 2	Heart size is enlarged.	0.135	0.458	0.70	**1**

(Mork et al., 2013); Li et al. (2018) calculated precision and false positive rate of the 10 most frequent abnormality-related terms in the dataset; and L. Wu et al. (2017) calculated accuracy, sensitivity and specificity for a set of keywords. Lastly, Alfarghaly et al. (2021) used a different approach by measuring the similarity between reports using cosine similarity between the average of the word embeddings, vector extrema (Forgues, Pineau, Larchevêque, & Tremblay, 2014) or greedy matching (Rus & Lintean, 2012).

2.4.3. NLP vs Clinical Correctness: design differences

NLP metrics have been widely used in general domain tasks such as machine translation and image captioning (Kilickaya et al., 2017; Mathur et al., 2020). However, some authors have already contested their efficacy in tasks in the general domain (Kilickaya et al., 2017; Reiter, 2018; Mathur et al., 2020; van Miltenburg et al., 2021), and in the medical report generation task (e.g. Boag et al., 2020; Pino et al., 2021). Regarding clinical

metrics there is still no standard defined or validated by expert judgment, but the CheXpert labeler stands out as the most used in the literature. The next paragraphs describe multiple disadvantages from NLP metrics, to argue that they are not the most suitable by design for the medical report generation task.

First, the NLP metrics reviewed are designed to work with more than one ground truth sentence (Papineni et al., 2002; Lin, 2004; S. Banerjee & Lavie, 2005; Vedantam et al., 2015), to account for multiple ways of rephrasing the same sentence. However, in the medical domain the datasets contain just one report written for each imaging study, thus the metrics are limited to one ground truth per sample. This implies that the n-gram matching approach may (1) overlook negation and uncertainties in the sentences, which are very important in clinical reports (Irvin et al., 2019); (2) be unaware of synonyms or different ways of mentioning the same findings; and (3) evaluate a generated report against noisy ground truths. Consider the examples from Table 2.3, showing ground truth and generated examples, and the performance they achieve using NLP and CheXpert F-1 metrics. In the first example (**GT 1**), the first generated report is clinically incorrect and achieves high NLP scores, while the third sample is correct but achieves low NLP scores, due to the specific wording and synonyms used in the reports. In the second example (**GT 2**), there is *out of reach* information, i.e. data that the model is not able to infer only from the input image (a *comparison* to a previous patient exam, and the *contact with a doctor*). However, those sentences are accounted for with the n-gram approach, giving a higher score to the output that replicates the ground truth more closely.

Second, detecting type of errors, inspecting error gradations or other error analysis would be essential in a clinical setting (Oakden-Rayner, Dunnmon, Carneiro, & Re, 2020), and to this end, NLP scores provide too coarse information in contrast to clinical metrics. For example, if a generated report achieves low BLEU, the score alone does not provide much information into what is wrong with the prediction. In contrast, a low score in

Table 2.4. Error gradation examples: ground truth (GT) and generated (Gen) reports with metrics BLEU (B, average BLEU 1-4), ROUGE-L (R-L), CIDEr-D (C-D, ranges from 0 to 10) and CheXpert F-1. Wrong words are marked in red.

	Report	**B**	**R-L**	**C-D**	**F-1**
GT 1	There is a large right sided effusion.				
Gen 1	There is a minimal right sided effusion.	0.711	**0.875**	4.88	**1**
Gen 2	There is a large left sided effusion.	0.711	**0.875**	5.06	**1**
Gen 3	There is a large right sided mass.	**0.779**	**0.875**	**5.90**	0
GT 2	Opacities in the lung bases may represent atelectasis.				
Gen 1	Opacities in the left lung may represent atelectasis.	0.675	**0.888**	4.57	**1**
Gen 2	Opacities in the lung bases may represent pneumonia.	**0.809**	**0.888**	**6.98**	0.5

CheXpert Cardiomegaly recall indicates that the model is predicting too many *false negatives* (i.e. not capturing all the Cardiomegaly cases from the dataset). In general, a metric carefully designed for the medical domain could be stronger than general domain NLP metrics in this aspect. Regarding error gradation in the general domain, van Miltenburg et al. (2020) performed error analysis in NLG tasks and showed examples where metrics were not necessarily able to differentiate multiple kind of errors. Inspired by this, we built Table 2.4 with different error gradation examples in chest X-ray reports. In such cases, clinicians may need to define which ones are *critical* and which ones are *less bad* in each situation, and evaluation methods should help instead of hindering the process.

Lastly, design decisions for NLP metrics are intended for the general domain. Hence, implementations may be not specific enough for clinical reports, for example (1) when considering synonyms using databases like WordNet (Miller, 1995) (e.g. METEOR, SPICE), (2) when detecting objects and its relations (e.g. SPICE), or (3) when tuning hyper-parameters (e.g. METEOR). Consider the CheXpert labeler, that works in a similar fashion to SPICE by parsing the text to universal dependencies and then classifying mentions of interest, it uses a parser trained with a biomedical model (McClosky, 2010), and

uses text patterns manually curated by radiologists to detect the abnormalities. Furthermore, we argue that for different medical sub-domains different implementations may be needed, such as chest X-rays, abdominal CTs, histopathology images, and others; since the specific target vocabulary (abnormalities, body parts, synonyms, and more) may differ significantly.

2.5. Models

The survey by Messina et al. (2020) shows most models follow the design pattern depicted in Figure 2.7. The input is one or more images, there is a Visual Component to handle the image, a Language Component for the report generation, the main output is the generated report, and optional outputs are an auxiliary classification and/or a heatmap over the image, for example using CNN saliency methods such as Grad-CAM (Selvaraju et al., 2017), or leveraging the attention mechanisms from the Language Component. For the Visual Component, most works use networks based on common Convolutional Neural Networks (CNNs), such as Densenet (G. Huang et al., 2017) or ResNet (He et al., 2016), that outputs features in a latent space, and optionally an auxiliary classification (e.g. using as target the set of CheXpert abnormalities). In contrast, there are more varied approaches as Language Component, where the most common approach derives from the general domain image captioning task with an LSTM-based or Transformed-based decoder, but there are also retrieval and hybrid retrieval-generation approaches proposed (Messina et al., 2020). Most authors that propose end-to-end models use the traditional teacher-forcing method to train recurrent networks (Williams & Zipser, 1989), though a few have tried Reinforcement learning.

The next subsections summarize the main models proposed in the literature for chest X-ray report generation, focusing on the differences in the Language component, categorizing in deep learning decoders (2.5.1) and retrieval-based approaches (2.5.2). Lastly, other learning approaches are briefly described (2.5.3).

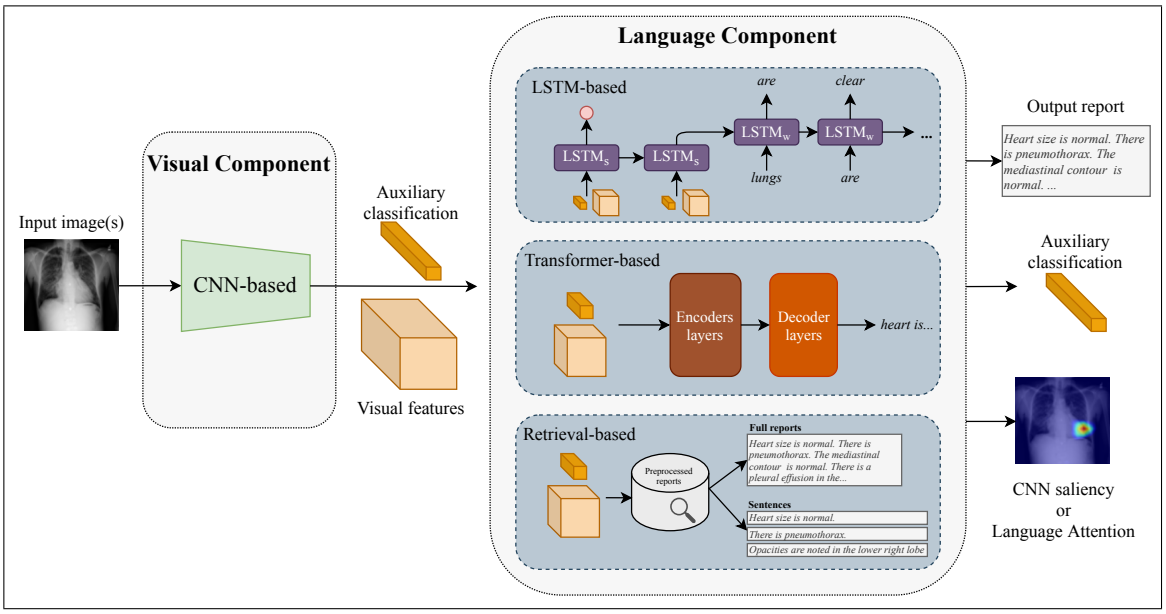


Figure 2.7. Overview of the models proposed in the literature. The Visual Component is a CNN-based network that outputs latent features and/or an optional auxiliary classification vector. The Language Component shows multiple variations, LSTM-based network (arranged in a hierarchical manner), Transformer-based networks, and Retrieval-based approaches. As additional output, some approaches output visual heatmap and the auxiliary classification.

2.5.1. Deep learning language components

The simplest approach uses a RNN to generate the output report word by word, such as an LSTM (Boag et al., 2020) or a GRU (Nishino et al., 2020), based on the image captioning model by Vinyals et al. (2015), and commonly incorporating an attention mechanism, based on Xu et al. (2015). Furthermore, some works (Jing et al., 2019, 2018; G. Liu et al., 2019; D. Hou et al., 2021; Zhang et al., 2020) have employed two LSTMs arranged hierarchically to generate sentences and words in each step. In most cases an attention mechanism is also included, either at every sentence step or every word step.

Some works (Z. Chen et al., 2020; Lovelace & Mortazavi, 2020; Xiong, Du, & Yan, 2019; Najdenkoska et al., 2021; You et al., 2021; B. Hou et al., 2021; Nguyen et al.,

2021; Miura et al., 2021) have used transformer-based networks (Vaswani et al., 2017) as a decoder to generate the report. The usual approach is to use one or more transformer encoder layers to receive the image features from the CNN and further process them, and then one or more transformer decoder layers to generate the text in a recurrent manner. Additionally, Z. Chen et al. (2020) included an external memory with the transformer in the decoding process.

2.5.2. Retrieval-based language components

Three papers have proposed purely retrieval methods. Syeda-Mahmood et al. (2020) manually curated a set of around 70 fine-grained labels (i.e. mentions of abnormalities and their characteristics), proposed a CNN-based model that detects the labels in the images, and then retrieves sentences with the found labels to generate the output report. Kougia et al. (2021) searches the most similar image in the training set, considering an image embedding and a set of predicted tags (both by a CNN), and returns its caption as output report. Lastly, Ni et al. (2020) trained cross-modal embeddings between images and sentences to relate abnormal image regions with abnormalities described in the reports; then, given an input image the embeddings are used to retrieve meaningful sentences and generate a report.

Four papers follow the hybrid retrieval-paraphrasing approach. Yang, Ye, You, and Ma (2021) proposed MedWriter, which uses image embeddings to retrieve relevant reports and sentences from a database, and then paraphrases them with a hierarchical LSTM model. Li et al. (2018) proposes HRGR, that consist of a hierarchical LSTM with a gate module to choose between retrieving from a database or generating with a word LSTM. CLARA (Biswal et al., 2020) was designed as an interactive tool: a human introduces *anchor words* and the prefix of a sentence, and the model uses the retrieval tool *Lucene* (Branko, Danijela, & Dušan, 2010) to retrieve sentence templates from a database. A sequence-to-sequence network then reads and paraphrases each sentence template to get the final

report. CLARA can also operate fully automatically by receiving an empty prefix and predicting the anchor words itself. Li, Liang, Hu, and Xing (2019) proposes KERP, a graph-based network to map the visual input into a sequence of templates from a curated database, and then paraphrase each sentence with another instance of their graph-based network.

2.5.3. Other learning approaches

Most works train the proposed models using the traditional teacher-forcing method (Williams & Zipser, 1989), i.e. predict each word conditioned on the previous words, therefore imitating the ground truth word by word. Two works use Reinforcement Learning (RL) to reward the model using a clinically relevant score: G. Liu et al. (2019) used the CheXpert labeler to define a medical reward, and Nishino et al. (2020) trained a BERT model to emulate the CheXpert labeler and used it as reward. Three works (Li et al., 2018; Xiong et al., 2019; Jing et al., 2019) also used RL, but with rewards based on NLP metrics (CIDEr, ROUGE and BLEU-4). Lastly, Lovelace and Mortazavi (2020) trained an LSTM to emulate the CheXpert labeler in a differentiable manner, and trained their model with a *clinically coherence loss* without requiring RL.

Most recently, F. Liu, Yin, et al. (2021) proposed a contrastive attention approach to train any encoder-decoder model, which helps discriminating better between normal and abnormal images. Similarly, F. Liu, Ge, and Wu (2021) proposed a framework to train any encoder-decoder model in a curriculum learning fashion, by ranking the image-report samples on different visual and textual difficulty measures.

3. PROPOSED METHOD

We propose a template-based model that detects a set of abnormalities in the image using a CNN as a binary classifier, and then relies on fixed sentences as templates for the text generation, as depicted in Figure 3.1. The reports generated are then *structured reports* covering all abnormalities in the set, containing one sentence per abnormality classified as present or absent. Ganeshan et al. (2018) has advocated for structured reports in radiology as a way to reduce diagnostic errors, facilitate the communication with the referring physician by reducing excessive language and styles, and ultimately reduce obstacles for optimal patient care. The details of our approach follow in the next subsections.

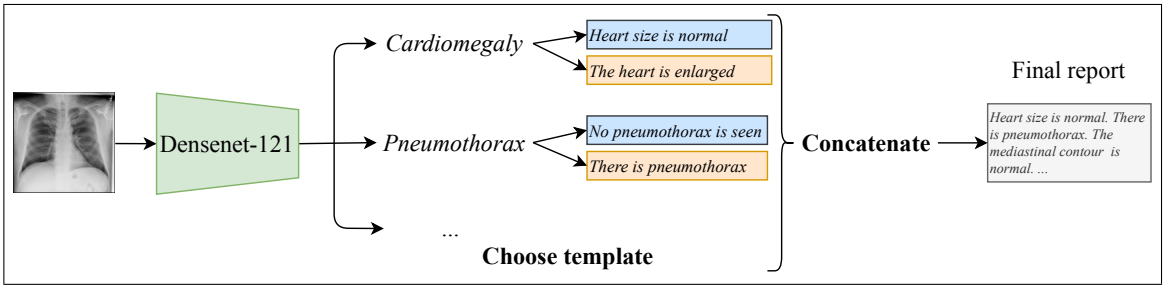


Figure 3.1. Template-based model for report-generation

3.1. Abnormality classification

To detect abnormalities, we implement a CNN that receives a chest X-ray and performs multi-label classification of the presence of the 13 abnormalities in the CheXpert set of labels[1](Irvin et al., 2019). We chose the CheXpert set of labels as target since is the most commonly used in the report generation literature (Messina et al., 2020). We use a Densenet-121 (G. Huang et al., 2017), which has shown good results in report generation

[1]All labels except for except for *No Finding*: namely *Cardiomegaly*, *Enlarged Cardiomediastinum*, *Atelectasis*, *Consolidation*, *Edema*, *Pneumonia*, *Pneumothorax*, *Lung Lesion*, *Lung Opacity*, *Pleural Effusion*, *Pleural Other*, *Fracture*, *Support Devices*.

(Messina et al., 2020) and in multi-label classification in chest X-rays (Rajpurkar et al., 2017).

For training, we use a weighted binary cross-entropy loss, similar to X. Wang et al. (2017), to compensate for the unbalanced classes. We initialized the network with the pre-trained weights from ImageNet (Deng et al., 2009), then pre-trained on the CheXpert dataset (Irvin et al., 2019) for the same medical classification task, then fine-tuned in the MIMIC-CXR dataset (weights θ_M), and lastly fine-tuned in the IU X-ray dataset (weights θ_I). We used the θ_M weights for the template-based model in the MIMIC-CXR dataset, and the θ_I weights for IU X-ray dataset. During training, we augment the images by random crop, translation, rotation, shear, changing brightness, contrast and adding gaussian noise. We optimize the saved model by the PR-AUC metric on the validation test set. At inference time, the CNN outputs a continuous value, thus, to obtain a binary classification we apply a threshold for each abnormality. The threshold value for each abnormality is calculated by finding a value that optimizes the F1-score obtained in the validation set.

3.2. Text generation

We manually curated a set of two sentences per abnormality indicating presence and absence, totaling 26 sentences. We built the templates by examining the reports and picking existing sentences or creating new ones. To generate the full report, the image is fed to the CNN to compute the binary classification, then the corresponding absence or presence template is chosen for each abnormality, and the sentences are concatenated into the final report.

We tested the model using three template sets: *single*, *abnormal-only* and *grouped*. All of them provide the same meaning clinically (in terms of the presence of the 13 abnormalities), but are written differently.

Single. Concise sentences that indicate the presence or absence directly, for example: *"No pleural effusion"* and *"Pleural effusion is seen"*. The presence templates do not provide detailed visual characteristics, such as location, severity, size, or more, since the classification model does not predict this information. All sentences are shown in Table 3.1.

Abnormal-only. Contains only the *positive* sentences (i.e. indicating abnormalities) from the *single* set, and replaces the *negative* sentences by empty sentences. If the prediction for an image is *negative* for all abnormalities, the output report is filled with the sentence *"No active disease"*, to avoid outputting an empty report.

Grouped. To resemble more the reports from each dataset, we grouped multiple abnormalities into common sentences from the training set. For example, in IU X-ray, if all the lung-related abnormalities are classified as absent, the template chosen is *"The lungs are clear"*, instead of using their individual absence templates. If at least one of the abnormalities does not match the group, the model falls back to a set of individual sentences. For the IU X-ray dataset we use the *single* set as the individual fallback, and for MIMIC-CXR, we modified some of the sentences from the *single* set to further resemble the reports from the dataset. Full set of sentences are detailed on Appendix B.

3.3. Implementation details

Our implementation is available online[2]. We used the pytorch implementation[3] of the Densenet-121 (G. Huang et al., 2017) architecture. Specifically, given an input image, we (1) use the `features` layer to extract a feature vector of size $1024 \times H \times W$, (2) apply global average pooling to obtain a vector of size 1024, (3) apply a dropout layer with $p = 0.5$, (4) pass through a fully connected layer to obtain a vector of size 13 with predicted values, and (5) apply a threshold to obtain a binary classification for each abnormality.

[2] `https://pdpino.github.io/clinically-correct`
[3] `https://pytorch.org/vision/stable/models.html`

Table 3.1. Sentences in the *single* template set.

Abnormality	**Absence template**	**Presence template**
Cardiomegaly	*Heart size is normal*	*The heart is enlarged*
Enlarged Cardiomed.	*The mediastinal contour is normal*	*The cardiomediastinal silhouette is enlarged*
Consolidation	*No focal consolidation*	*There is focal consolidation*
Lung Opacity	*The lungs are free of focal airspace disease*	*One or more airspace opacities are seen*
Atelectasis	*No atelectasis*	*Appearance suggest atelectasis*
Pleural Effusion	*No pleural effusion*	*Pleural effusion is seen*
Pleural Other	*No fibrosis*	*Pleural thickening is present*
Pneumonia	*No pneumonia*	*There is evidence of pneumonia*
Pneumothorax	*No pneumothorax is seen*	*There is pneumothorax*
Edema	*No pulmonary edema*	*Pulmonary edema is seen*
Lung Lesion	*No pulmonary nodules or mass lesions identified*	*There are pulmonary nodules or mass identified*
Fracture	*No fracture is seen*	*A fracture is identified*
Support Devices	–	*A device is seen*

The specific threshold value for each label is calculated by finding a value that optimizes the F1-score obtained in the validation set.

When training, the weights from the convolutional layers were initialized with ImageNet pre-trained weights from pytorch, and the full model (convolutional and fully connected layer) were pre-trained in the CheXpert dataset. When pre-training in CheXpert, we used a batch size of 54, trained with the Adam optimizer for 15 epochs with learning rate 0.0001 and weight decay (L2-norm) 0.00001. When training in MIMIC-CXR or IU X-ray, we used a batch size of 110, trained with the Adam optimizer for 30 epochs with learning rate 0.00003 and weight decay 0.002. In both cases, we resized the input images to 256×256, and saved the model with the best PR-AUC evaluated in the validation set. We used a GPU Nvidia RTX 3090 and a GPU Nvidia RTX 2080 for the experiments.

4. MATERIALS

4.1. Datasets

We perform the experiments with two publicly available datasets: IU X-ray[1] (Demner-Fushman et al., 2015) and MIMIC-CXR[2] (A. Johnson et al., 2019; A. E. W. Johnson et al., 2019). For both datasets we used the *findings* section of the reports and kept only frontal X-rays (AP and PA projections). We used the official train-validation-test split for MIMIC-CXR, and we split the IU X-ray dataset in 8:1:1 proportions. To pre-process the reports we applied tokenization, manually corrected some typos, and kept all the tokens in the vocabulary. The amount of final images and dataset statistics is detailed in Table 4.1, and the distributions of number of words and number of sentences per report are shown in Figure 4.1. MIMIC-CXR presents somewhat longer and potentially more complex reports, which is consistent with the sources and sizes of the datasets.

Table 4.1. Datasets statistics. Abnormal images are those with one or more abnormalities present, as labeled by the CheXpert labeler.

Amount	**IU X-ray**	**MIMIC-CXR**
Total images	7,470	377,110
Total reports	3,955	227,827
Frontal images	3,311	243,326
Train images	2,638 (80%)	237,964 (97.8%)
Validation images	336 (10%)	1,959 (0.8%)
Test images	337 (10%)	3,403 (1.4%)
Healthy images	1,297 (39.2%)	40,165 (16.5%)
Abnormal images	2,014 (60.8%)	203,161 (83.5%)
Unique sentences	6,435	361,440
Unique tokens	1,578	11,598
Average number of words	37	56.5
Average number of sentences	4.6	5.2

[1] `https://openi.nlm.nih.gov/faq`
[2] `https://physionet.org/content/mimic-cxr-jpg/2.0.0/`

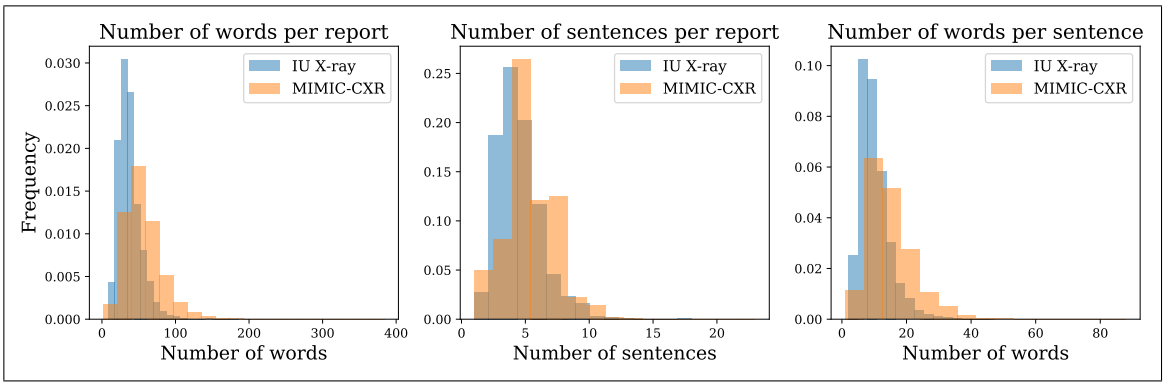


Figure 4.1. Word and sentence distributions.

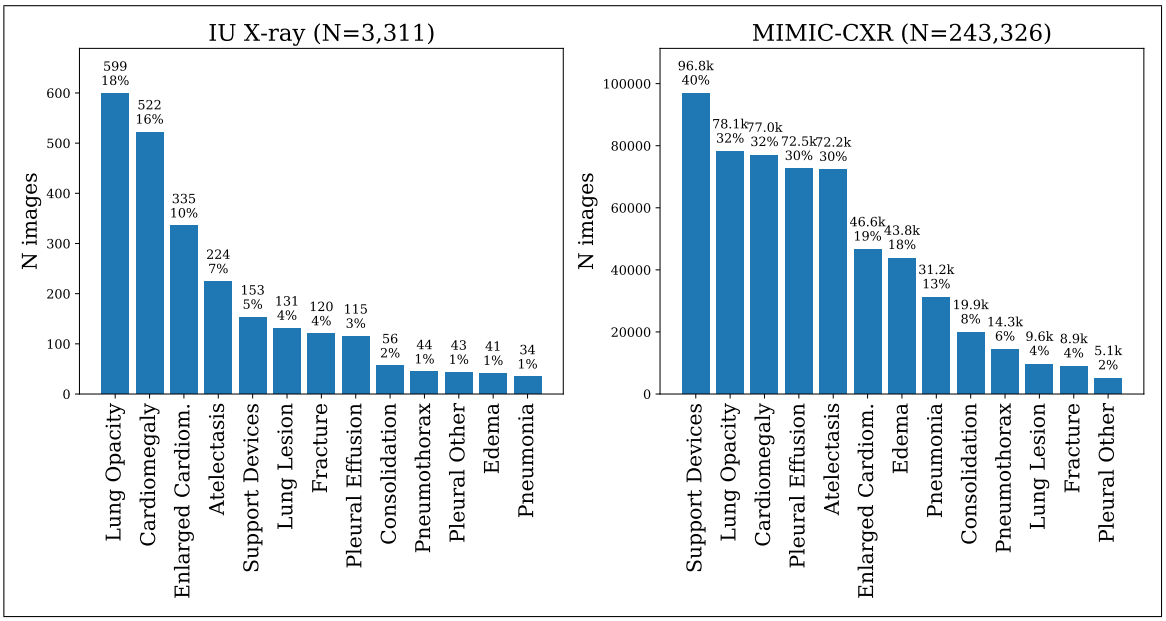


Figure 4.2. CheXpert label distributions.

We applied the CheXpert labeler (Irvin et al., 2019) to the processed reports to obtain the ground truth labels for the classification training task. Figure 4.2 shows the distributions of CheXpert abnormalities in each dataset, exhibiting the unbalance toward more negative samples in each label, as commonly seen in the medical domain. For example, in the MIMIC-CXR dataset, the five least common labels have less than 10% positive cases each; and in IU X-ray only four abnormalities have more than 200 positive cases.

4.2. Metrics

We used three NLP metrics: BLEU-N (B-N) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), and CIDEr-D (C-D) (Vedantam et al., 2015), implemented in a publicly available python library[3]. CIDEr-D ranges from 0 (worst) to 10 (best), while the others from 0 (worst) to 1 (best).

As clinical correctness metrics, we used the CheXpert labeler[4] (Irvin et al., 2019) and MIRQI[5] (Zhang et al., 2020), which were detailed in section 2.4.2. In both cases, we provide F1-score (F-1), precision (P), and recall (R). The CheXpert average values are the macro average across the 14 labels.

4.3. Baselines

4.3.1. Naive Models

We implement five simple baselines that are not clinically useful, but provide a reference value for the metrics. The first four models are unconditioned upon the query image, and all of them except by *Top-words-N* are grammatically correct.

- *Constant*: returns the same report for all the images, manually curated using sentences describing a healthy subject. We created four constant report versions for the different experiments, two using common sentences from each dataset (*version 1* and *version 2*), a *short* and a *long* variant (Table 4.2).
- *Random retrieval*: returns a random report from the training set.
- *Top-words-N*: generates a report randomly using the N most common words from the training set. First, it defines the number of words for the output report

[3] https://github.com/salaniz/pycocoevalcap
[4] https://github.com/stanfordmlgroup/chexpert-labeler
[5] https://github.com/xiaosongwang/MIRQI

by choosing randomly a length K from the training set, and then samples K words from the N most common words and returns them in a random order. We tested with $N \in \{10, 50, 100, 500\}$ and kept the best results in terms of clinical metrics.

- *Top-sentences-N*: same procedure as *Top-words-N*, using N most common sentences instead of words.
- *1-NN (Nearest-Neighbor)*: given a test image, search for the most similar image in the training set and return its report. It uses the embedded space by the last layer of features from a CNN to search for the closest image, using cosine distance.

Table 4.2. Reports used in the Constant models. The first two versions are based on common sentences from the two datasets.

Version	**Constant report**
Version 1 (IU X-ray)	*The heart is normal in size. The mediastinum is unremarkable. The lungs are clear. There is no pneumothorax or pleural effusion. No focal airspace disease. No pleural effusion or pneumothorax.*
Version 2 (MIMIC-CXR)	*In comparison with the study of xxxx, there is little change and no evidence of acute cardiopulmonary disease. The heart is normal in size. The mediastinum is unremarkable. No pneumonia, vascular congestion, or pleural effusion.*
Short	*No acute findings.*
Long	*Heart size is normal. The mediastinal contour is normal. No pulmonary nodules or mass lesions identified. The lungs are free of focal airspace disease. No pulmonary edema. No focal consolidation. No pneumonia. No atelectasis. No pneumothorax is seen. No pleural effusion. No fibrosis. No fracture is seen.*

4.3.2. CNN-LSTM based Models

We re-implemented two encoder-decoder models from the image captioning task, Show and Tell (ST) (Vinyals et al., 2015) and Show, Attend and Tell (SAT) (Xu et al.,

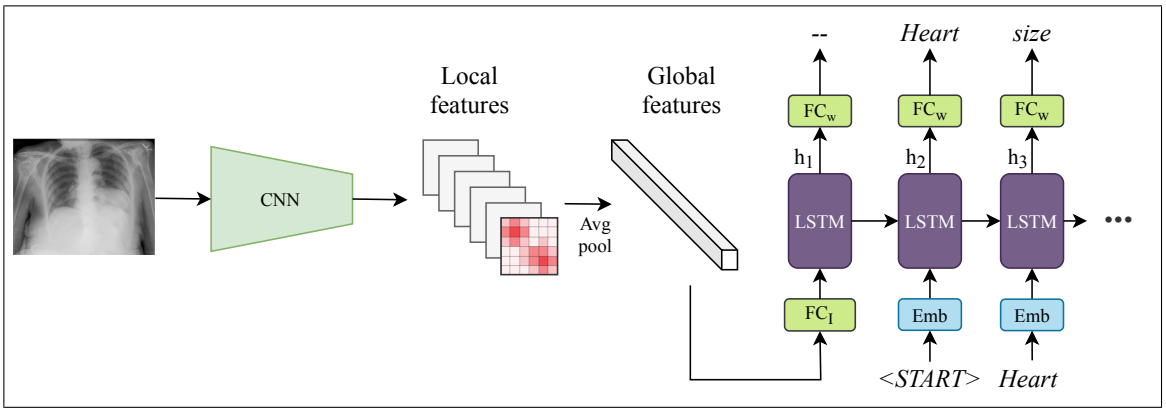


Figure 4.3. Show and Tell model (Vinyals et al., 2015). FC stands for fully-connected layer, where FC_I encodes the image features into the word embedding space, and FC_w receives the LSTM state and generates the report word by word.

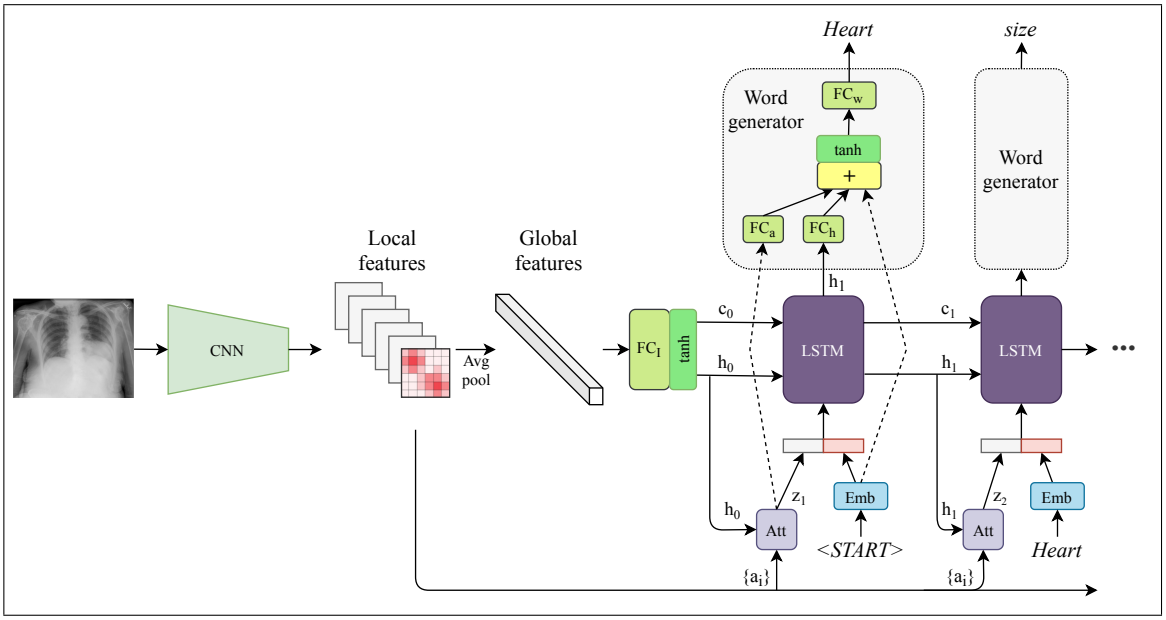


Figure 4.4. Show Attend and Tell model (Xu et al., 2015). FC stands for fully-connected layer, where FC_I encodes the global image features to initialize the LSTM hidden state, and FC_a, FC_h and FC_w are used to generate each word in each step.

2015). Both use the CNN from the template-based model as encoder and a LSTM network as decoder to generate the full report word by word, as depicted in Figures 4.3

and 4.4. Show, Attend and Tell uses a visual attention mechanism that receives the local features from the image and the hidden state from the previous step, which is detailed below. The word embeddings are of size 100 and we initialized them with the RadGlove (Zhang, Ding, Qian, Manning, & Langlotz, 2018), which was pre-trained with radiology reports. The models are trained with a cross-entropy loss to generate the report word by word, using teacher forcing (Williams & Zipser, 1989). During training, in each epoch the CheXpert F-1 score was measured in the validation set, and the model with the best performance was kept.

Show, Attend and Tell attention mechanism. In each time step t, the attention is calculated as: $z_t = Attention(\{a_i\}, h_{t-1})$, where h_{t-1} is the LSTM hidden state in the previous step, $\{a_i\}$ are the local features in each image region i, and the attention output is the context vector z_t. Attention weights α_i are computed with a two-layer perceptron f_{att} (equations 4.1 and 4.2), where W_1, W_2, W_3 and b are learnable weights and bias. A β_t scalar is used as a gating mechanism to let the model decide whether to emphasize language or context in each step, and is computed with the fully-connected layer f_β followed by a sigmoid activation (equation 4.3). Lastly, the context vector z_t is computed as the weighted sum of weights α_i and features a_i (soft attention), multiplied by the gating β_t factor (equation 4.4).

$$\alpha_{ti} = \text{softmax}\left(f_{att}(a_i, h_{t-1})\right) \tag{4.1}$$

$$= \text{softmax}\left(W_1 \tanh\left(W_2 a_i + W_3 h_{t-1} + b\right)\right) \tag{4.2}$$

$$\beta_t = \sigma(f_\beta(h_{t-1})) = \sigma(W_4 h_{t-1}) \tag{4.3}$$

$$z_t = \beta_t \sum_i \alpha_{ti}\, a_i \tag{4.4}$$

4.3.3. Chest X-ray report-generation literature

We compare our approach with the results from nineteen models from the literature: seven LSTM-based (Jing et al., 2018; G. Liu et al., 2019; Boag et al., 2020; Nishino et al., 2020; Zhang et al., 2020; D. Hou et al., 2021; F. Liu, Yin, et al., 2021), six Transformer-based (Z. Chen et al., 2020; Lovelace & Mortazavi, 2020; Nguyen et al., 2021; Najdenkoska et al., 2021; Miura et al., 2021; B. Hou et al., 2021) three fully-retrieval (Syeda-Mahmood et al., 2020; Ni et al., 2020; Kougia et al., 2021), and three hybrid retrieval-generation based (Li et al., 2018, 2019; Biswal et al., 2020). We re-implemented the CoAtt model (Jing et al., 2018), and for the rest we show the results from their papers.

5. RESULTS

5.1. Report generation benchmark

We benchmark the naive baselines, baseline encoder-decoder models, state-of-the-art methods and our template-based model in the report generation task, using CheXpert and NLP metrics. We leave out MIRQI since it does not pass a minimal criterion detailed in the Stress tests results subsection. Tables 5.1 and 5.2 show the benchmark in the IU X-ray and MIMIC-CXR datasets, CheXpert metrics are macro averaged across 14 labels, and Table 5.3 show the benchmark in each CheXpert label in the MIMIC-CXR dataset. There are several remarks from these results, discussed next.

Template sets. The clinical performance for all the sets of the template-based model is the same, as expected, but their NLP performance are highly varied. The *abnormal-only* has the lowest NLP performance, probably due to its much shorter reports. The *grouped* set is able to achieve much higher NLP performance than the *single* set, particularly in the IU X-ray dataset, only by using sentences that appear more commonly in the dataset. Thus, we show that we can improve NLP metrics by changing the report wording while preserving the clinical correctness in terms of the seen abnormalities, suggesting NLP metrics are not robust enough to measure performance in this task.

NLP vs Clinical Correctness. From the whole benchmark, the medical correctness metrics seem to be better than NLP metrics at differentiating naive baselines versus deep learning models, considering the naive models are not clinically useful by design. The baselines that are unconditioned upon the query image, namely *Constant*, *Top-words-N*, *Top-sentences-N* and *Random retrieval*, achieve comparable performance to our template-based model, and even to some literature models, especially in the IU X-ray dataset. On the contrary, these naive models mostly achieve lower performance on CheXpert, whereas encoder-decoder, literature and template-based models show higher values. One exception

Table 5.1. IU X-ray report generation benchmark. CheXpert metrics are macro-averaged across 14 labels. [f+i] indicates they generated both *findings* and *impression* sections concatenated, while the rest generated *findings* only; * indicates we re-implemented the code; super script letters R, T and L indicate Retrieval, Transformer and LSTM-based approaches; super scripts RL and CA indicates the use of Reinforcement Learning and Contrastive Attention strategies.

	CheXpert			NLP			
Model	**F-1**	**P**	**R**	**B-1**	**B-4**	**R-L**	**C-D**
Constant-v1	0.038	0.026	0.071	0.462	0.173	0.366	0.307
Top-words-50	0.040	0.052	0.102	0.401	0.000	0.228	0.127
Top-sentences-100	0.051	0.043	0.072	0.387	0.104	0.296	0.160
Random retrieval	0.066	0.065	0.068	0.382	0.084	0.284	0.145
1-NN	0.153	0.143	0.170	0.407	0.106	0.303	0.207
ST*, Vinyals et al.[L]	0.065	0.103	0.089	0.246	0.072	0.292	0.192
SAT*, Xu et al.[L]	0.119	0.153	0.131	0.374	0.102	0.328	0.213
CoAtt*, Jing et al.[L]	0.144	0.162	0.147	0.403	0.114	0.316	0.221
CoAtt, Jing et al.[L,f+i]	-	-	-	0.517	0.247	0.447	0.327
Zhang et al.[L,f+i]	-	-	-	0.441	0.147	0.367	0.304
ARL, D. Hou et al.[L,RL]	-	-	-	-	0.125	0.262	0.366
F. Liu, Yin, et al.[L,CA]	-	-	-	0.492	0.169	0.381	-
Nguyen et al.[T]	0.152	0.142	0.173	0.515	0.235	0.362	-
Najdenkoska et al.[T,f+i]	-	-	-	0.493	0.154	0.375	-
Miura et al.[T,RL]	0.269	0.285	0.294	-	0.120	-	**0.996**
CLARA, Biswal et al.[R]	-	-	-	0.471	0.199	-	0.359
HRGR, Li et al.[R]	-	-	-	0.438	0.151	0.369	0.343
KERP, Li et al.[R]	-	-	-	0.482	0.162	0.339	0.280
RTEX, Kougia et al.[R]	-	0.193	0.222	-	0.055	0.202	-
Syeda-Mahmood et al.[R,f+i]	-	-	-	**0.560**	**0.490**	**0.580**	-
Templ. abnormal-only	**0.304**	**0.311**	**0.355**	0.018	0.002	0.125	0.006
Templ. single	**0.304**	**0.311**	**0.355**	0.299	0.073	0.282	0.033
Templ. grouped	**0.304**	**0.311**	**0.355**	0.450	0.145	0.352	0.252

to this trend is worth noticing, the *Top-words-100* model achieves high CheXpert scores in the MIMIC-CXR dataset, even though its sentences are not necessarily grammatically correct. This could indicate a weak point of the CheXpert labeler, since its parser is not

Table 5.2. MIMIC-CXR report generation benchmark. CheXpert metrics are macro-averaged across 14 labels. $^{f+i}$ indicates they generated both *findings* and *impression* sections concatenated, while the rest generated *findings* only; * indicates we re-implemented the code; ^{Ab} indicates they used a subset of the data only with reports that have one or more abnormal findings; super script letters R, T and L indicate Retrieval, Transformer and LSTM-based approaches; super scripts RL and CA indicates the use of Reinforcement Learning and Contrastive Attention strategies.

	CheXpert			**NLP**			
Model	**F-1**	**P**	**R**	**B-1**	**B-4**	**R-L**	**C-D**
Constant-v2	0.009	0.005	0.071	0.245	0.036	0.226	0.060
Top-words-100	0.212	0.234	0.237	0.326	0.002	0.189	0.052
Top-sentences-100	0.043	0.093	0.075	0.190	0.040	0.204	0.029
Random retrieval	0.203	0.236	0.186	0.277	0.040	0.198	0.035
1-NN	0.361	0.371	0.355	0.331	0.061	0.222	0.069
ST*, Vinyals et al.^{L}	0.189	0.331	0.180	0.273	0.067	0.235	0.054
SAT*, Xu et al.^{L}	0.275	0.400	0.256	0.263	0.067	0.248	0.083
CoAtt*, Jing et al.^{L}	0.185	0.381	0.186	0.256	0.071	0.260	0.076
Boag et al.^{L}	0.186	0.304	-	0.305	0.092	-	0.850
ARL, D. Hou et al.^{L,RL}	0.156	0.218	0.135	-	0.148	0.329	0.402
G. Liu et al.^{L,RL}	-	0.309	0.134	0.313	0.103	0.306	**1.046**
Nishino et al.^{L,RL}	0.217	-	-	0.217	0.048	-	-
F. Liu, Yin, et al.^{L,CA}	0.303	0.352	0.298	0.350	0.109	0.283	-
Z. Chen et al.^{T}	0.276	0.333	0.273	0.353	0.103	0.277	-
Lovelace and Mortazavi^{T}	0.228	0.333	0.217	0.415	0.146	0.318	0.316
Nguyen et al.^{T}	0.412	0.432	0.418	**0.495**	**0.224**	**0.390**	-
RATCHET, B. Hou et al.^{T}	0.276	0.332	0.271	0.232	-	0.240	0.493
Najdenkoska et al.$^{T,f+i}$	0.210	0.350	0.151	0.418	0.109	0.302	-
Miura et al.^{T,RL}	0.310	0.364	0.360	-	0.111	-	0.492
CVSE, Ni et al.^{R,Ab}	0.253	0.317	0.224	0.192	0.036	0.153	-
RTEX, Kougia et al.^{R}	-	0.229	0.284	-	0.059	0.205	-
Templ. abnormal-only	**0.462**	**0.442**	**0.547**	0.084	0.008	0.171	0.008
Templ. single	**0.462**	**0.442**	**0.547**	0.282	0.044	0.209	0.046
Templ. grouped	**0.462**	**0.442**	**0.547**	0.324	0.076	0.225	0.078

being able to detect the lack of grammatical structure while parsing the reports for this model.

Template-based clinical correctness. Regarding clinical metrics, our template-based models outperforms all other models in terms of CheXpert F-1, precision and recall macro-averaged across 14 labels in both datasets (Tables 5.1 and 5.2), and individual F-1 scores in 12 out of 13 abnormalities in MIMIC-CXR (Table 5.3). Specifically, the template-based model achieves much better performance than (1) the naive methods, showing it surpasses the first lower standard; and (2) the deep learning models, proving the approach to be more effective while simpler. The 1-NN baseline is also very strong, as it achieves CheXpert average scores higher than the encoder-decoder baselines, comparable to some literature models, and reached the highest F-1 score for the *Pleural Other* abnormality.

NLP metrics. Regarding NLP metrics, the naive and encoder-decoder baselines we trained perform closer to the literature models in IU X-ray than MIMIC-CXR, indicating that better results might be easier to achieve in the former dataset than the latter. This is consistent with the sizes and sources of the data, since MIMIC-CXR is much larger and varied in terms of patients and their conditions. Additionally, from the results in MIMIC-CXR (Table 5.2), CIDEr-D appears to be the most difficult to increase up to literature levels, from all the NLP metrics used.

5.2. Stress test on clinical metrics

We design and implement a simple stress test for the clinical correctness metrics to assess their validity, inspired by the behavioral tests for NLP models proposed in the CheckList methodology (Ribeiro, Wu, Guestrin, & Singh, 2020). The test consists of producing a set of generated reports with a baseline method, then manipulating the wording of the reports without changing their clinical meaning, and evaluate them both (i) baseline and (ii) manipulated methods with a target metric. Then, a clinical correctness metric is expected to stay constant between both methods, since the reports mention the same abnormal findings, just with different words. Otherwise, the metric may have specific design failures

Table 5.3. MIMIC-CXR report generation benchmark by CheXpert label.

F1-scores	**1-NN**	**SAT*, Xu et al.^{L}**	**Boag et al.^{L}**	**Lovelace and Mortazavi^{T}**	**RATCHET, B. Hou et al.^{T}**	**Miura et al.^{T,RL}**	**CVSE, Ni et al.^{R,Ab}**	**Templates single**
No Finding	0.237	0.266	0.407	**0.541**	0.451	0.454	0.300	0.354
Enlarged Cardiom.	0.294	0.272	0.134	0.059	0.015	0.077	0.061	**0.449**
Cardiomegaly	0.595	0.661	0.390	0.433	0.446	0.482	0.555	**0.708**
Lung Lesion	0.162	0.083	0.001	0.014	0.069	0.038	0.148	**0.277**
Lung Opacity	0.512	0.326	0.077	0.171	0.344	0.174	0.345	**0.644**
Edema	0.496	0.518	0.271	0.298	0.407	**0.622**	0.273	**0.622**
Consolidation	0.182	0.085	0.014	0.073	0.041	0.089	0.151	**0.274**
Pneumonia	0.209	0.160	0.030	0.039	0.234	0.267	0.270	**0.376**
Atelectasis	0.485	0.412	0.146	0.322	0.411	0.516	0.398	**0.610**
Pneumothorax	0.152	0.150	0.043	0.098	0.110	0.190	0.060	**0.237**
Pleural Effusion	0.655	0.679	0.473	0.480	0.633	0.730	0.539	**0.739**
Pleural Other	**0.152**	0.000	0.001	0.009	0.000	0.000	0.058	0.073
Fracture	0.143	0.084	0.001	0.000	0.000	0.071	0.056	**0.260**
Support Devices	0.782	0.787	0.613	0.660	0.697	0.635	0.334	**0.846**
Macro average	0.361	0.320	0.186	0.228	0.276	0.310	0.253	**0.462**

that make it unreliable or non-robust in some cases. We describe the test design, test cases, and present the results next.

Test design: unmention vs negative. In the context of clinical reports, when radiologists mentions an abnormality *negatively* they mean the same as *not mentioning* the abnormality. Consider the following reports: *"the heart is normal in size, no pneumothorax or pleural effusion observed"*; and *"no acute findings"*. The former mentions explicitly the abnormalities *Cardiomegaly*, *Pneumothorax* and *Pleural Effusion* as *negative*, though in the latter report they are *unmentioned* and the patient is assumed to be healthy in those

aspects. This is very common in real reports, as different radiologists adopt different styles when writing reports (Ganeshan et al., 2018). Hence, the test consists of manipulating the reports by replacing *negative* sentences by *unmentions*, or vice versa.

Test cases. As a first example, we used the proposed Template-based method with the *single* set of templates as baseline, and the *abnormal-only* set as the manipulated method: both contain abnormality predictions from the same CNN, but write different reports. The *single* set produces a maximum of 12 negative sentences[1], while the *abnormal-only* does not produce any negative sentence. Refer to the Proposed Method Chapter for details on the template sets. As a second example, we used the Constant model with the *version 1* as baseline, and the *long* and *short* versions as manipulated: the three of them describe a healthy subject, but with more or less sentences. The *version 1* has 6 negative sentences, *long* has 13 negative sentences, and *short* has no negative sentences. Refer to Table 4.2 in the Materials Chapter for the full reports.

Table 5.4. Stress tests results for CheXpert and MIRQI metrics, in IU X-ray and MIMIC-CXR datasets. Marks –, ↑ and ↓ indicate that the values stayed the same, increased and decreased from the baseline, respectively.

		CheXpert			MIRQI		
	Model	**F-1**	**P**	**R**	**F-1**	**P**	**R**
IU X-ray	Templ. single	0.239	0.225	0.357	0.529	0.534	0.540
	Templ. abn-only	0.239 –	0.225 –	0.357 –	0.457 ↓	0.447 ↓	0.478 ↓
	Constant-v1	0.038	0.026	0.071	0.469	0.462	0.481
	Constant-short	0.038 –	0.026 –	0.071 –	0.356 ↓	0.356 ↓	0.356 ↓
	Constant-long	0.038 –	0.026 –	0.071 –	0.462 ↓	0.452 ↓	0.481 –
MIMIC-CXR	Templ. single	0.462	0.442	0.547	0.739	0.801	0.720
	Templ. abn-only	0.462 –	0.442 –	0.547 –	0.833 ↑	0.801 –	0.897 ↑
	Constant-v1	0.009	0.005	0.071	0.134	0.118	0.167
	Constant-short	0.009 –	0.005 –	0.071 –	0.038 ↓	0.038 ↓	0.038 ↓
	Constant-long	0.009 –	0.005 –	0.071 –	0.121 ↓	0.102 ↓	0.176 ↑

[1]All abnormalities except of *Support Devices*, which is an empty sentence.

Results (Table 5.4). The CheXpert metrics pass both tests, as it does not change its value with any manipulation, while MIRQI does not pass the tests, since the scores are modified in almost all cases. To better understand this result, we studied the metrics design by inspecting their code in detail. By design, the CheXpert labeler merges the *unmention* and *negative* classes into one, as discussed in the Background subsection. However, MIRQI only considers *negative*, *uncertain* and *positive* mentions in its matching process, and *unmentioned* abnormalities are ignored. Thus, explaining why the MIRQI scores can be altered by manipulating reports with *negative* and *unmention* sentences.

Furthermore, notice that MIRQI changes differently in each case and in each dataset. In the Constant case it mostly decreases its score, except by MIRQI-r; and in the Template-based case increases in MIMIC-CXR and decreases in IU X-ray. We further inspect the Template-based test case by comparing the CheXpert label distribution in both datasets, grouped by *unmention*, *negative*, and *positive+uncertain*, and added across abnormalities (Figure 5.1). Notice that if we discard the *unmentions* (blue bar), as MIRQI does, both datasets present very different distributions between the other two classes; in particular, in IU X-ray there are more *negative* mentions, while in MIMIC-CXR there are more *positive+uncertain* mentions. Thus, a model biased to generate more abnormal than normal sentences could have more chances at a better performance in MIMIC and worse in IU; and vice versa with a model biased towards normal sentences. This could be a possible explanation to the observed discrepancy, but we believe it requires further analysis to understand to what extent we can game MIRQI.

To sum up, the CheXpert metric is able to pass our minimal stress test, and even though it has not been tested thoroughly in the literature, we believe is good enough to use as clinical correctness metric in this thesis. However, MIRQI does not pass the test, and we found at least one severe problem that makes it unreliable and even vulnerable to gaming effects. For example, we could leverage the biases in the datasets to create a model that achieves better performance, without changing the clinical meaning or actual usefulness

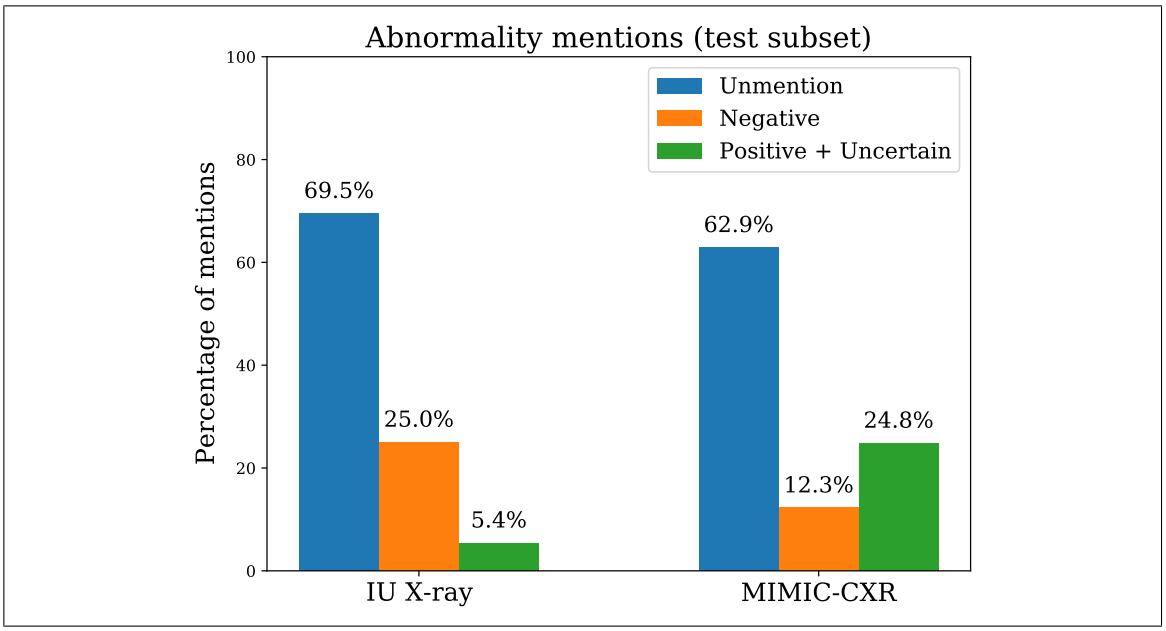


Figure 5.1. CheXpert distribution of labels in both datasets, considering test sets only.

of the generated reports. Hence, we decided not to use MIRQI as a clinical metric in the other experiments. It is worth mentioning MIRQI has some interesting ideas, as it also attempts to capture attributes and modifiers of the abnormalities, such as location, severity, shape, and more, which can also be significant in reports (Datta & Roberts, 2020).

5.3. NLP metrics in corpora with controlled clinical meaning

We test the NLP metrics in corpora with manually controlled clinical meaning to (1) observe the metrics behavior in different scenarios, and to (2) measure how well they differentiate clinically correct and incorrect text samples. Specifically, we refer to a corpus as a set of ground truth reports paired with generated texts, and we build a corpus by choosing the reports by their output from the CheXpert labeler. Assuming the labeler is correct *most of the times*[2], we argue we are building a *corpus with controlled clinical meaning*. We

[2]As detailed in the (Background subsection), the authors (Irvin et al., 2019) tested the labeler against samples manually assessed by radiologists, achieving near 0.8-0.9 F-1 score classification results.

also remark that CheXpert only accounts for the presence, absence or uncertainty of abnormalities, but ignores additional attributes mentioned, such as severity, location, shape, and more; hence, the controlled meaning will also ignore these details. The full analysis is divided in five main steps, which are detailed next.

Step 1: Label and group sentences. First, we list all unique sentences in a dataset and label them using the CheXpert labeler. Then, for each abnormality, we group the sentences according to one of the four outputs given by the labeler: *positive (pos)*, *uncertainty (unc)*, *negative (neg)* or *unmention (none)*. See example sentences for *Lung Opacity* and *Cardiomegaly* in Table 5.5. The amount of unique sentences per dataset is $6,435$ in IU X-ray and $361,440$ in MIMIC-CXR, and the amount of sentences per CheXpert group and per abnormality for each dataset is detailed in the Appendix C.1 (Table C.1).

Table 5.5. Example sentences grouped by their CheXpert output, regarding *Lung Opacity* and *Cardiomegaly* abnormalities. Words marked in red indicate a positive finding, in orange an uncertainty, and in green a healthy finding.

	Unmention	**Negative**	**Uncertain**	**Positive**
Lung Opacity	*No acute findings.*	*No focal opacity.*	*Definite infiltrate is not excluded.*	*Left basilar retrocardiac opacity.*
	Heart size is upper normal.	*No infiltrates or masses in the lungs.*	*Question left basilar atelectasis versus infiltrate.*	*Bibasilar and perihilar interstitial opacities.*
Cardiomegaly	*Overall clear lungs.*	*Heart size and mediastinal contours are normal.*	*Borderline heart size.*	*Slight cardiomegaly.*
	No edema or pneumonia.	*The heart is not enlarged.*	*Heart size is stable*	*The heart size is mildly enlarged.*

Step 2: Build controlled corpus. For a given abnormality (e.g. *Cardiomegaly*), we select two of its sentences groups (e.g. *negative* and *positive*), and we build a *corpus with a controlled clinical meaning* by using sentences from the first group as ground truth reports and sentences from the second group as generated reports (e.g. all ground truth are *negative* and all generated are *positive*). Notice we control the meaning of all the pairs in the corpus, with the aforementioned assumption that the labeler is correct always or most of the times. We use the notation *"Abnormality gt-gen"*, where *gt* represents the group chosen for the ground truth, *gen* the group chosen for the generated, and the abnormality may be omitted if it can be inferred from the context (the example would be *Cardiomegaly neg-pos*, and all samples would be *false positives* detections). Table 5.6 shows two corpora examples with some report samples.

The most exhaustive way to create a corpus with two groups is to take the Cartesian product between all sentences in the groups, i.e. pair all sentences in the first group with all the sentences in the second, in which we end up with a quadratic amount of samples. Considering the amount of sentences in each dataset, specially in MIMIC-CXR, working with a corpus this size is not feasible, so we apply a random sampling strategy to select pairs, detailed in the appendix C.2.

Step 3: Compute metrics. Given a corpus with clinically controlled meaning, we can compute NLP metrics for the full corpus. Notice that for each abnormality we can build a corpus between all pairs of CheXpert groups, totalling $4 \times 4 = 16$ corpora. Thus, we can compose a 4×4 matrix summarizing NLP scores for each abnormality and each NLP metric, see left of Figure 5.2 with an example for *Lung Opacity* and ROUGE-L. The cell in the first row and first column of the matrix shows ROUGE-L calculated with the corpus *none-none*, in the second row and first column for the corpus *neg-none*, and so on. Matrices for all abnormalities and NLP metrics are shown in the Appendix C.3.

Table 5.6. Examples of reports in clinically controlled corpora. Words marked in red indicate a positive finding, in orange an uncertainty, and in green a healthy finding.

Ground truth report	**Generated report**
Corpus 1: *Pleural Effusion pos-neg*	
Small left pleural effusion is noted, similar to previous examination. *Bilateral pleural effusions, small on the right, and trace on the left.* ...	*There is no evidence of right pleural effusion.* *No pleural effusions are currently seen.* ...
Corpus 2: *Pneumonia neg-unc*	
There is no evidence of lower pneumonia. *The right lung is clear and there is no evidence of acute pneumonia.* ...	*If clinical suspicion for pneumonia persists, followup radiograph may be helpful.* *In the correct clinical setting, superimposed infection is not excluded.* ...

In Figure 5.2, the diagonal of the matrix holds the corpora where the groups share the same CheXpert output (e.g. *neg-neg*, *unc-unc*, etc), and thus are correct in a clinical sense. On the other hand, the non-diagonal scores are corpora with incorrect samples from all sorts, *neg-pos*, *unc-none*, *none-pos*, etc. Hence, if ROUGE-L (or any NLP metric being analyzed) was correlated with clinical meaning in these corpora, we would expect diagonals with higher values than non-diagonal values. In the Figure 5.2 example, the higher values are in the corpora *neg-neg*, *unc-unc* and *pos-pos*; though the differences with non-diagonal values are not so large in some cases, such as *pos-pos* with its neighbors *unc-pos* and *pos-unc*, which indicates that *positive* sentences may be easily confused with *uncertain* sentences.

The right side of Figure 5.2 shows two plots with score distributions across the samples in some selected corpora, e.g. in the top plot, the blue histogram shows the distribution from the *neg-neg* corpus and the orange from the *neg-pos* corpus. By comparing distributions from two corpora, we can analyze the discrimination capabilities of the NLP metric.

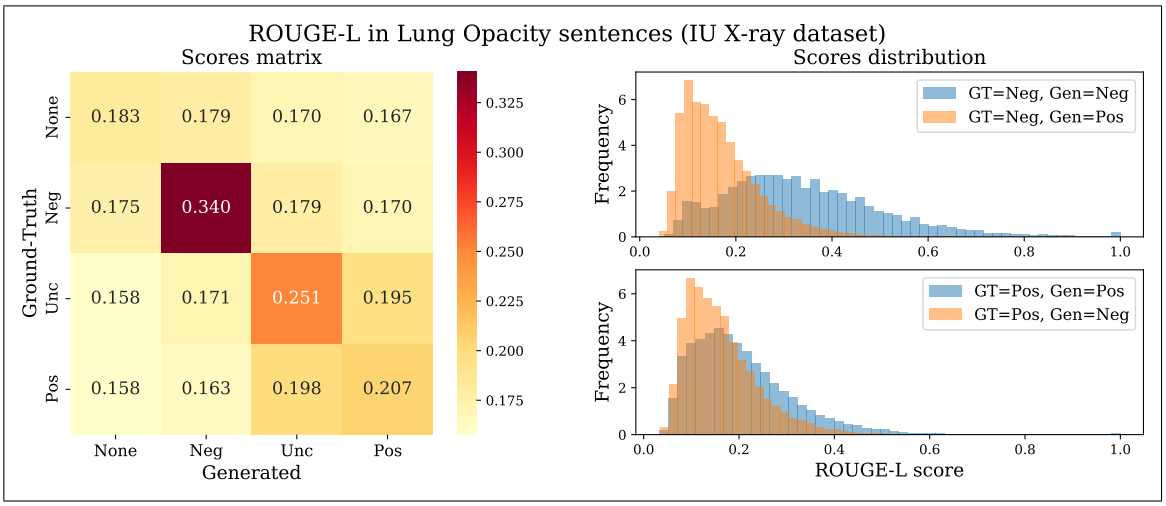


Figure 5.2. ROUGE-L scores computed in clinically controlled corpora from *Lung Opacity*, in the IU X-ray dataset. The matrix summarizes scores for each corpus built, and the histograms show the distribution of four selected corpora. The blue histograms are clinically correct corpora, while the orange histograms are clinically incorrect.

For example, the top plot represents the following case: *given a negative ground truth regarding Lung Opacity, discriminate between positive and negative generated sentences*. If ROUGE-L is able to discriminate such sentences, then we would expect that the distributions be highly separated. The top plot seems somewhat separated, while the bottom plot seems much less separated. We show more distribution plots for other abnormalities, NLP metrics and pair of corpora in the Appendix C.3, and we delve deeper into this separation problem in the next steps.

Step 4: Reduce CheXpert to binary. When using the CheXpert labeler as a classification metric, the outputs are usually reduced from four to two: an abnormality labeled as *unmention* is considered a *negative* case, and typically a label *uncertain* is considered a *positive* case, as discussed in the Background section. Hence, we repeat the previous steps merging the aforementioned outputs into a binary classification setting. Thus, we build 2×2 matrices for each abnormality and each NLP metric, see Figure 5.3 with an example

with BLEU-1 and *Cardiomegaly*, Figure 5.4 with examples for multiple abnormalities and NLP metrics, and see Appendix C.3 for more matrices and histograms.

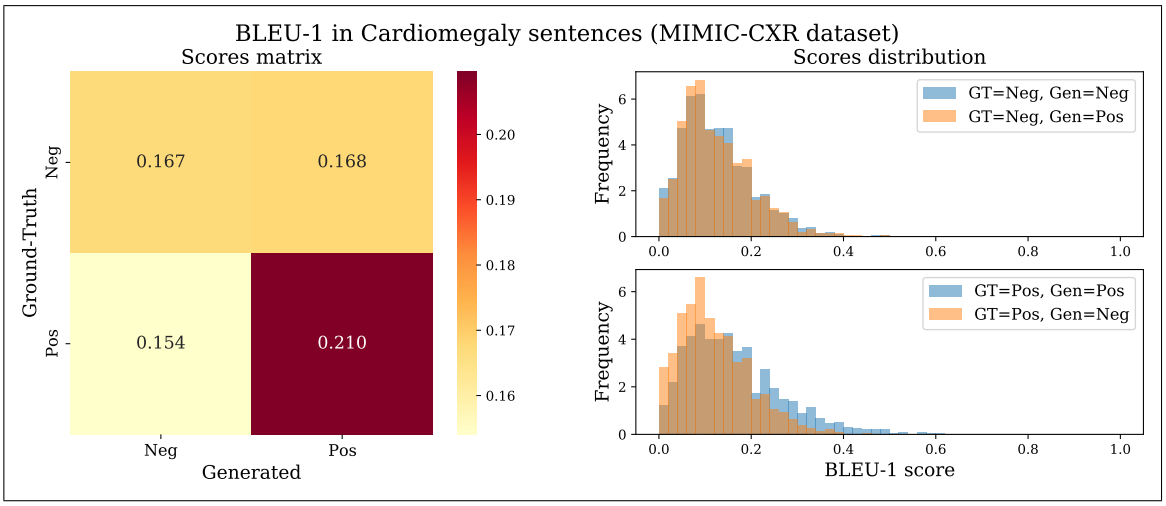


Figure 5.3. BLEU-1 scores computed in corpora controlled with *Cardiomegaly* in a binary setting. The matrix summarizes scores for each corpus built, and the histograms show the distribution of the four corpora. The blue histograms are clinically correct, while the orange histograms are clinically incorrect.

The matrix in the left of Figure 5.3 shows the 2×2 matrix, where the diagonal holds the *neg-neg*[3] corpus (true negative samples) and *pos-pos* corpus (true positives) corpus, while the non-diagonal cells hold the *neg-pos* (false positives) and *pos-neg* (false negatives) corpora. Same as before, we would expect diagonal values to be higher than non-diagonal values, if the metric is correlated with clinical meaning. From the matrix only, the corpus that stands out with a higher value is the *pos-pos*; on the other hand, *neg-neg* has similar scores to the clinically incorrect corpora.

The right of the Figure 5.3 plots the score distributions across samples in the four corpora. If the BLEU-1 metric was good at discriminating correct from incorrect sentences regarding *Cardiomegaly*, we would expect the histograms to be highly separated, but it

[3]We abuse notation by, from hereinafter, referring to the *pos* and *neg* groups as the merged classes (i.e. *positive* plus *uncertain* and *negative* plus *unmentioned*, respectively).

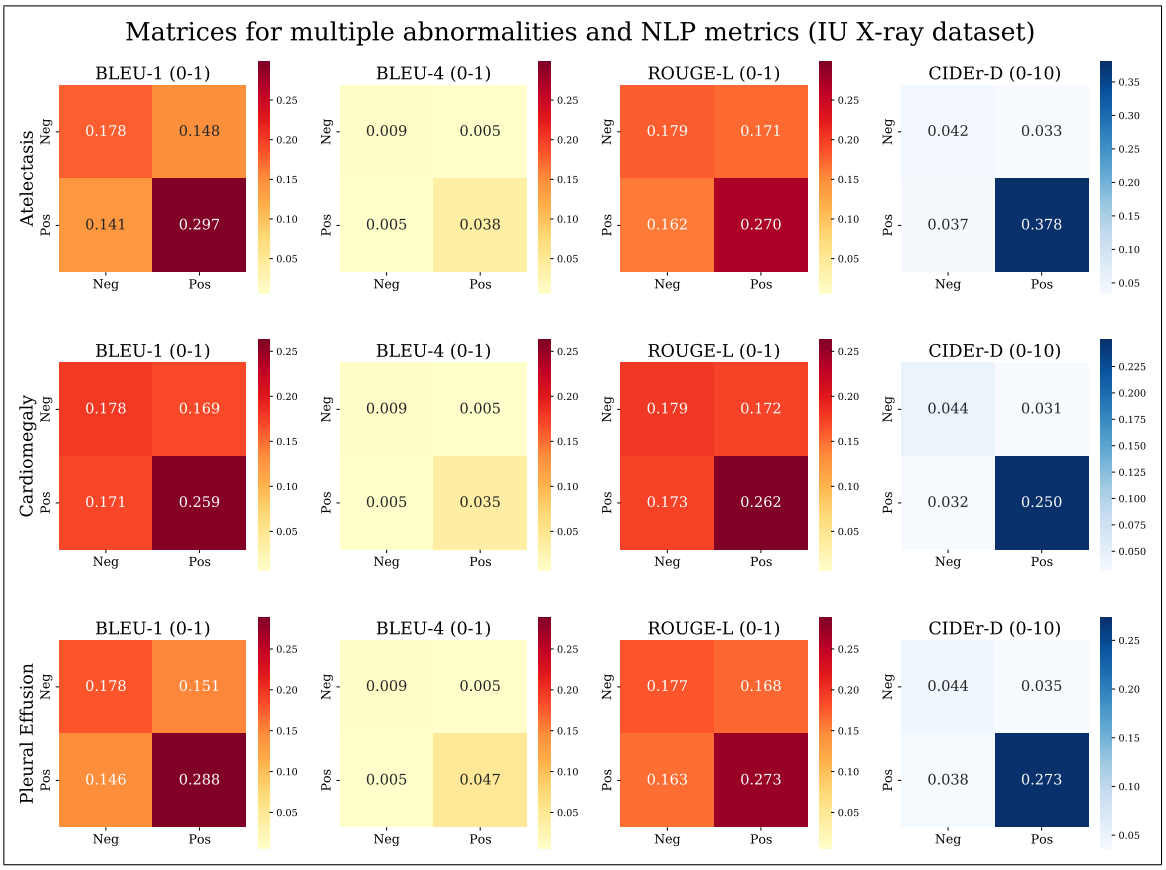


Figure 5.4. Matrices showing NLP metrics behavior with respect to CheXpert outputs in three abnormalities in the IU X-ray dataset. All matrices show a similar pattern, the most different score is achieved by the *pos-pos* corpus in the each case (bottom right cell in each matrix), while the rest of the cells show more similar scores among them.

does not seem to be the case. Specifically, given a positive *Cardiomegaly* ground truth (bottom histogram), BLEU-1 does not discriminate very well between a negative or positive generated texts; and given a negative *Cardiomegaly* ground truth (top histogram), the separation is even worse.

Step 5: Measuring separation with ROC-AUC. It is clear from the plots that the histograms are not to be easily separated, but we have not yet analyzed this quantitatively. To this end, we pose the discrimination problem as a binary classification problem, split

in two analogous tasks: given a fixed abnormality, an NLP metric outputs a score to discriminate the correct classification, i.e.,

(i) Given a ***positive*** **ground truth**, discriminate between a generated sentence as *negative* (incorrect) or *positive* (correct). As a probability, this is expressed as: $P(Gen = pos|GT = pos)$. Example in the bottom right plot in Figure 5.3.

(ii) Given a ***negative*** **ground truth**, discriminate between a generated sentence as *negative* (correct) or *positive* (incorrect). As a probability, this is expressed as: $P(Gen = neg|GT = neg)$. Example in the top right plot in Figure 5.3

We then compute ROC-AUC to measure the goodness of the classification in each task, without having to set a specific threshold for the score. If the ROC-AUC score achieved is close to 1, the NLP metric is able to separate sentences closely to the CheXpert labeler; on the other hand, if the ROC-AUC is close to 0.5, the discrimination is almost as bad as random.

Tables 5.7 and 5.8 show ROC-AUC scores computed for all metrics and all abnormalities in the IU X-ray and MIMIC-CXR datasets. The ROC-AUC scores are much higher for the first task than the second, in most metrics, in most abnormalities and in both datasets, indicating that the NLP metrics are much better at discriminating true positive than true negative mentions. Nonetheless, the ROC-AUC scores are still relatively low in the first task, and very low in the second task. In the former, the higher scores barely reach 0.7-0.8 scores, and in the latter they are around and even below 0.5 (a random classifier achieves 0.5). CIDEr-D seems to have the best discriminating power in the first task among all metrics, which could be due to the IDF scores used in its calculation; but still has very low performance in the second task.

Table 5.7. NLP metrics capabilities to separate clinically correct from incorrect samples, measured in ROC-AUC in binary classification tasks in the IU X-ray dataset. Task 1: given a *pos* ground truth, separate *pos* from *neg* generated reports; and Task 2: given a *neg* ground truth, separate *neg* from *pos* generated reports. ROC-AUC equal to 0.5 indicates the metric separates as good as a random classifier; ROC-AUC near 1 indicates perfect separation.

	Task 1 $P(Gen = pos \vert GT = pos)$				**Task 2** $P(Gen = neg \vert GT = neg)$			
Abnormality	**B-1**	**B-4**	**R-L**	**C-D**	**B-1**	**B-4**	**R-L**	**C-D**
Atelectasis	0.796	0.695	0.808	**0.899**	0.505	**0.588**	0.521	0.455
Cardiomegaly	0.645	0.633	0.690	**0.694**	0.518	0.500	**0.521**	0.517
Consolidation	0.717	0.623	0.727	**0.778**	0.484	**0.538**	0.497	0.455
Edema	0.729	0.703	0.784	**0.813**	0.567	0.566	**0.582**	0.541
Enlarged Cardiom.	0.650	0.577	0.672	**0.715**	0.511	**0.547**	0.524	0.495
Fracture	0.693	0.595	0.720	**0.790**	0.524	**0.550**	0.538	0.526
Lung Lesion	0.711	0.606	0.653	**0.734**	0.547	**0.650**	0.562	0.462
Lung Opacity	0.659	0.570	0.632	**0.700**	0.532	**0.615**	0.543	0.458
Pleural Effusion	0.754	0.660	**0.782**	0.769	0.514	**0.566**	0.525	0.468
Pleural Other	0.709	0.669	0.727	**0.760**	0.497	**0.556**	0.519	0.479
Pneumonia	0.691	0.624	0.652	**0.756**	0.565	**0.637**	0.584	0.493
Pneumothorax	0.726	0.594	0.764	**0.767**	0.508	**0.549**	0.529	0.481
Support Devices	0.622	0.540	0.614	**0.638**	0.521	**0.576**	0.524	0.490

Table 5.8. NLP metrics capabilities to separate clinically correct from incorrect samples, measured in ROC-AUC in binary classification tasks in the MIMIC-CXR dataset. Task 1: given a *pos* ground truth, separate *pos* from *neg* generated reports; and Task 2: given a *neg* ground truth, separate *neg* from *pos* generated reports. ROC-AUC equal to 0.5 indicates the metric separates as good as a random classifier; ROC-AUC near 1 indicates perfect separation.

	Task 1 $P(Gen = pos \vert GT = pos)$				**Task 2** $P(Gen = neg \vert GT = neg)$			
Abnormality	**B-1**	**B-4**	**R-L**	**C-D**	**B-1**	**B-4**	**R-L**	**C-D**
Atelectasis	0.704	0.614	0.703	**0.736**	0.497	0.493	**0.508**	0.450
Cardiomegaly	0.621	0.594	0.659	**0.680**	0.503	**0.525**	0.511	0.502
Consolidation	0.686	0.590	0.686	**0.706**	0.493	**0.530**	0.494	0.437
Edema	0.720	0.699	0.773	**0.774**	0.533	0.541	**0.575**	0.504
Enlarged Cardiom.	0.622	0.589	0.645	**0.675**	0.492	**0.526**	0.494	0.514
Fracture	0.707	0.663	0.760	**0.814**	0.551	0.537	**0.564**	0.545
Lung Lesion	**0.613**	0.577	0.593	0.601	0.520	**0.564**	0.534	0.507
Lung Opacity	0.635	0.579	0.611	**0.664**	0.460	0.509	**0.510**	0.441
Pleural Effusion	0.766	0.721	**0.779**	0.764	0.498	**0.553**	0.481	0.501
Pleural Other	0.671	0.683	0.694	**0.704**	0.501	**0.555**	0.531	0.464
Pneumonia	0.653	0.612	0.680	**0.744**	0.532	**0.572**	0.514	0.496
Pneumothorax	0.720	0.646	**0.764**	0.745	0.493	**0.524**	0.496	0.489
Support Devices	0.650	0.585	**0.677**	0.673	0.508	**0.512**	0.506	0.477

6. DISCUSSION

6.1. Hyp 1: NLP metrics are not the most suitable for the report generation task

Following the current message on NLP metrics, we provided further evidence showing that NLP metrics are not the most appropriate for the chest X-ray report generation task. We showed their lack of robustness by proposing naive baselines that achieve NLP performance comparable to the state-of-the-art in IU X-ray, and by improving the NLP performance of our template model without affecting its clinical performance. Additionally, we showed that NLP metrics poorly discriminate sentences with opposite meaning regarding a given abnormality, while using the CheXpert labeler (Irvin et al., 2019) as the gold standard. Given our results, we argue NLP metrics should not be used isolated to evaluate the generated reports in this task.

Despite the general lack of robustness of NLP metrics, we also observed they are better at detecting true positives than true negative sentences. This aspect could be leveraged to use them as complementary metrics, or be further studied to design other evaluation metrics. Furthermore, CIDEr-D seems to be the least fragile metric from our experiments, as it is less vulnerable to the manipulations tried, and it is the best at discriminating true positive sentences. It may be worth exploring if its TF-IDF design provides a robustness that the others do not have.

Regarding clinical correctness metrics, we proposed a simple behavioral stress test to contest them, inspired by the CheckList methodology (Ribeiro et al., 2020), by interchanging *negative* and *unmentions* of abnormalities in chest X-ray reports. We showed the CheXpert labeler (Irvin et al., 2019) passes this minimal criterion, but MIRQI (Zhang et al., 2020) does not, proving it to be vulnerable to gaming effects or adversarial inputs. Furthermore, we believe more tests can be created by leveraging the examples shown

in the Background subsection *NLP vs Clinical Correctness*, regarding detection of synonyms, negations, gradation of errors and more. Thus, in the future, any metric can be tested to pass minimal criteria (e.g. robustness in *negative vs unmention*, robustness with synonyms, etc.), and to accomplish different goals (e.g. detect type I/II errors, etc.).

Lastly, we showed that specialized clinical metrics have more advantages than NLP metrics by design. As discussed in the Background subsection *NLP vs Clinical Correctness* and in the *Clinically Controlled Corpora* experiments, a metric like CheXpert gives fine-grained results, as it can identify abnormalities in both ground truth and generated sentences. On the contrary, an NLP metric only outputs a score indicating the similarity of the ground truth and generated reports, without further context on abnormalities, error type, or anything else. Hence, a carefully designed metric can be much more insightful into what kind of errors are found.

In sum, regarding our first hypothesis, we conclude **NLP metrics are not enough to be used isolated in the report generation task, clinical metrics are more appropriated to this end**, and overall, **both could be used in tandem to compensate for their disadvantages.** Additionally, CheXpert is currently the most appropriate clinical correctness metric available for this task, since MIRQI proved to be vulnerable in certain cases. In general, we believe our findings imply that making progress in the report generation task in one medical sub-domain, like chest X-rays, does not guarantee progress for other modalities or body parts, such as abdominal CTs, brain MRIs, etc., and authors should pay special attention to their evaluations in each situation.

6.2. Hyp 2: Template-based model outperforms state-of-the-art measured by CheXpert

We proposed a template-based model that is simpler and achieves higher clinical performance than SOTA in the report generation task from chest X-rays. First, the model

is **much simpler** than the state-of-the-art methods, as is a smaller deep learning model than the common end-to-end LSTM-based or Transformer-based approaches, and it uses fewer templates and a more straightforward retrieval process than other Retrieval-based approaches. Second, we test the model thoroughly in the IU X-ray and MIMIC-CXR datasets, using CheXpert as the clinical correctness metric, which we argue is currently the most appropriate to this end. The template-based model is able to **achieve higher clinical performance** than SOTA models measured by CheXpert, even though its NLP performance is lower.

We believe authors should be more cautious when measuring improvements by proposed deep learning models, given our findings while contesting the metrics in this topic, and since traditional NLP metrics are even challenged in the general domain (Kilickaya et al., 2017; Reiter, 2018; Mathur et al., 2020; van Miltenburg et al., 2021). In practice, **we recommend comparing the model with simple baselines**, such as our template-based model, a 1-NN, which showed very strong performance, and even naive models if possible.

An advantage of our template-based model is that the language generation process **is more inherently interpretable than other models**, as defined by Rudin (2019). Rudin (2019) describes *inherently interpretable models* as systems structured or constrained by domain knowledge so they are useful and transparent for humans to understand; contrary to *explainable models*, which are typically too complex for a human to comprehend, and need post-hoc methods to be explained, methods that have proven to be unreliable (Adebayo et al., 2018; Ghassemi et al., 2021). Consider Figure 6.1, contrasting an end-to-end model (top) where the full process is opaque, to our template-based model (bottom) that performs abnormality detection inside a black box, but the text generation process is fully transparent and interpretable. This aspect allows, for example, modifying the templates easily, either by radiologists or developers.

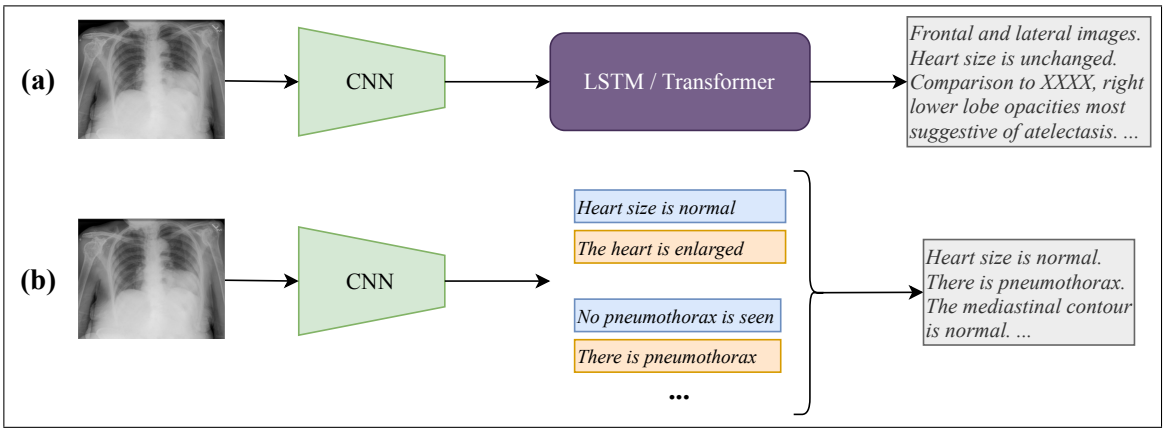


Figure 6.1. Interpretability comparison between: (a) end-to-end model as many from the literature, and (b) our template-based model. The former is completely opaque, while the latter performs text generation as a fully transparent process.

6.3. Are current evaluations sufficient for clinical deployment?

Beyond our hypotheses and metrics analyzed in this thesis, we argue that **most evaluations applied in the report generation task are still insufficient to assess the readiness of AI models in a clinical workflow**, as it even occurs in other image-based tasks applied in the medical domain, such as classification, regression or segmentation. For example, Roberts et al. (2021) reviewed over 400 pre-printed and published works in 2020 addressing diagnosis or prognosis of COVID-19 from radiology images using machine learning, and concluded that none of the models was potentially usable for clinical practice; one of the main reasons was insufficient or non-existent external evaluation of the models. Other authors have also observed poor generalization in deep learning models using chest X-rays, mainly due to domain shift issues, for example in abnormality classification (Pooch, Ballester, & Barros, 2020), and in COVID-19 lesion segmentation (Gonzalez et al., 2021). Additionally, multiple authors have found potential bias/fairness issues in chest X-ray AI applications, which could deepen racial disparities in medical practices. Seyyed-Kalantari, Zhang, McDermott, Chen, and Ghassemi (2021) showed that deep learning models achieved less performance, particularly less recall (sensitivity), on subgroups by

sex, age, race and more. I. Banerjee et al. (2021) showed deep learning models trained for clinical tasks were able to detect race of the patient from radiology images with high performance (over 0.9 AUC).

Even so, we also remark some works addressing the abnormality classification from chest X-rays task that perform more extensive evaluations. Seah et al. (2021) tackled the task of classifying 127 clinical findings, assessed the performance of radiologists with and without assistance of their automated DL-based system, and used data from three continents, including inpatient, outpatient and emergency settings. G. Wang et al. (2021) proposed a pipeline to classify viral, non-viral and COVID-19 pneumonia. They proved their model generalized well across multiple centers and multiple countries, and compared its performance against senior and junior radiologists.

Overall, automated systems proposed in the report generation task may **have a high potential impact in the clinical domain, but we believe that the evaluation methods used are still insufficient**. We argue that authors in this research field should aim to (1) perform more rigorous internal and external evaluations, (2) evaluate with expert physicians, (3) carry out further error analyses of the models, such as robustness and sensitivity, (4) address problems regarding fairness or demographic bias issues, and more. We believe our contribution of analyzing existing metrics and proposing a simple model is a first step towards this goal in the chest X-ray report generation task.

6.4. Limitations

We identify six main limitations in our work. First, we mostly report the results presented in the original articles, thus the benchmark could be improved. The evaluation protocols from each work may vary, specifically, the reports pre-processing steps, the train-test splits used, and the metric implementations. Also, many works do not provide

clinical correctness metrics, specially in the IU X-ray dataset, and most do not provide results disaggregated by CheXpert labels.

Second, our problem definition narrows the scope of the reports considerably, by removing clinical information as additional input, and denoting information in the output reports as *out of reach* (e.g. references to past exams). Thus, our experiments could be complemented by a more comprehensive analysis of the report inputs and outputs.

Third, our template model and the CheXpert labeler are limited to detect presence or absence of a specific set of abnormalities, disregarding (a) other abnormalities and (b) their visual characteristics, such as location, severity, or more. Furthermore, since the template-based model is especially tailored for the CheXpert labels, it has an inherent advantage over other models when measuring with the CheXpert metric.

Fourth, our templates are specific to our chest X-ray datasets and are not directly generalizable to other datasets. Hence, in order to use the template-based method with other image modalities or body parts, we would have to manually curate a set of templates covering relevant abnormalities, gather relevant labels, and train a classifier model.

Fifth, the ground truth reports and labels might be somewhat noisy, which could reduce the overall quality of fully-supervised models and of evaluations based in this data. There are two main causes for this noise. On the one hand, we use the CheXpert labeler as extraction tool, since is the most commonly used in the chest X-ray report generation literature (Messina et al., 2020), but it is yet to be proven to be a gold standard. On the other hand, we have ignored the intrinsic ambiguity observed in the medical domain, which exists because physicians may be uncertain of their diagnostics or disagree in some cases (Cabitza et al., 2017). Both these facts introduce noise in the ground truth reports and labels; for example, Oakden-Rayner (2020) found severe discrepancies between the ground truth labels and the abnormalities seen in the images in the ChestX-ray14 dataset (X. Wang et al., 2017), due to these problems.

Sixth, we argue the set of abnormalities used in the CheXpert (Irvin et al., 2019) or ChestX-ray14 (X. Wang et al., 2017) datasets may not be the most appropriate, even though they are the most used in the report generation literature (Messina et al., 2020). On the one hand, some abnormalities are not fully ascertainable on the images, since they cannot be distinguished without additional clinical information, and should at most be suggested by the radiologist (Lukaszewicz et al., 2016). For instance, in chest X-rays, *Consolidation*, *Pneumonia* and *Infiltration* are very hard to distinguish without additional clinical information (Oakden-Rayner, 2020). On the other hand, there may be cases of hidden stratification, i.e. within the same label, there are finer-grained abnormalities of clinical importance that are not accounted for, and are critical to deploy such a system to a medical scenario (Oakden-Rayner et al., 2020). For example, Oakden-Rayner (2020) inspected *Pneumothorax* positive cases in the ChestX-ray14 dataset (X. Wang et al., 2017), finding a majority of them were successfully treated patients, with the image showing a chest drain and no collapse of the lung; however, other cases were not treated, showing a collapsed lung, and thus were potentially lethal. Hence, the latter sub-category was far more important to detect due to the clinical relevance, but a machine learning model could much more easily learn to predict only the former sub-category due to the amount of data. In sum, it may be necessary to curate the set of target abnormalities more carefully, to consider clinically relevant labels and so the model does not attempt to generate information that is *out of reach*, i.e. that is not able to predict from the available input.

6.5. Future work

As future work, first, we will replicate implementations from some papers to evaluate and compare their performance under the same experimental conditions. Thus, we should be able to have more details on the performance differences.

Second, we will further study the clinical correctness evaluation problem by validating metrics with expert radiologists, and by inspecting essential characteristics of the reports

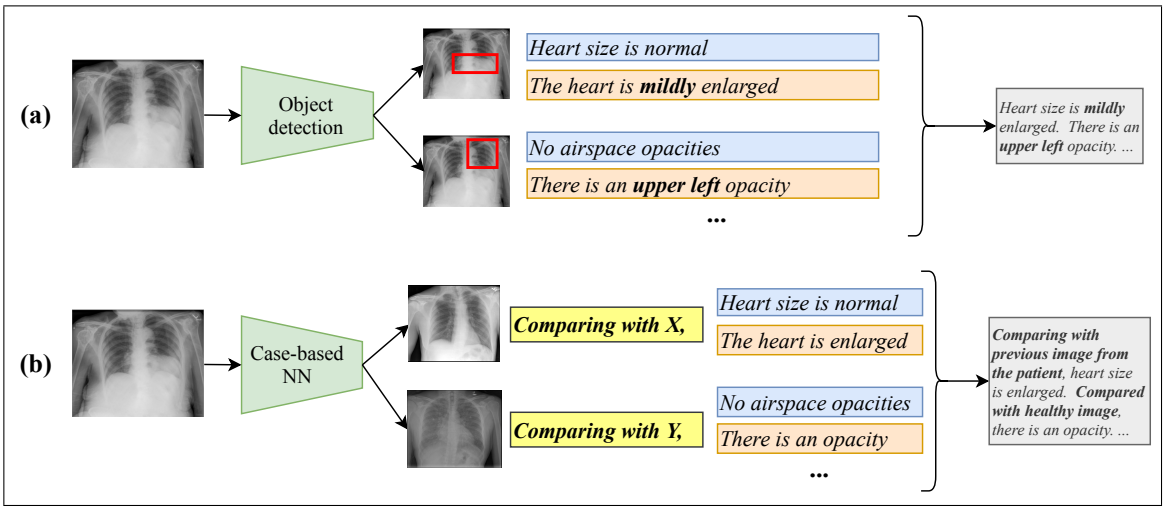


Figure 6.2. Possible model improvements. (a) Replacing the CNN with an Object detection model, to enhance the reports with visual characteristics, such as abnormality severity, location, shape, etc. (b) Replacing the CNN with a neural network that classifies with case-based reasoning (e.g. Barnett et al., 2021), to provide a comparison in the reports and present a more transparent reasoning.

that we argue should be considered in the design of the evaluation methods. For example, we want to focus on covering the abnormalities that are ascertainable on the images, since some diseases mentioned in the reports may require clinical history or additional information to diagnose them (Lukaszewicz et al., 2016; Oakden-Rayner, 2020). In addition, we want to address the uncertainty and inter-radiologist disagreement often observed in the reports, which is due to the intrinsic ambiguity in the medical domain (Cabitza et al., 2017), since this aspect is not very studied in the report generation task from a model or evaluation perspective (Messina et al., 2020).

Third, we want to enhance the reports from the template-based model by detecting more abnormalities and their visual characteristics, such as location, severity, and more. To this end, the CNN may be replaced by an object detection model (top of Figure 6.2), and we can leverage recently published datasets with enriched data, such as RadGraph (Jain, Saahil et al., 2021) and Chest ImaGenome (J. Wu et al., 2021) that add scene graphs

to the MIMIC-CXR dataset (A. E. W. Johnson et al., 2019) indicating relations between anatomic landmarks and abnormalities in the images; or other datasets that contain localization labeling on multiple abnormalities, such as *Pneumothorax*, *Cardiomegaly*, or *Pleural Effusion* (Feng et al., 2021; Zhou et al., 2021).

Fourth, and closely related to the previous one, we want to improve the template-based model by replacing the CNN with a *more inherently interpretable model*, which would allow enriching the report with extra information for the radiologists. For example, we could use a network that performs case-based reasoning for the diagnosis (Barnett et al., 2021), to provide more details on the specific reasoning on the network to the users (bottom of Figure 6.2). Another example is using the novel ENNs that emulate reasoning by concepts (Blazek & Lin, 2021).

Fifth, we want to expand our problem definition to resemble more a real clinical scenario. For example, include more patient information as input (e.g. *indication* section, *comparison* exams, etc.), since this information may be critical to make some diagnoses (Oakden-Rayner, 2020; Summers, 2021). Also, include as output the recommendation of follow-up exams, which is an important step in the communication with the referring physician (Lukaszewicz et al., 2016). Furthermore, we want to add prognosis-like information in the output reports, i.e. predict a patient condition in the future, which is typically much harder than diagnosis-like tasks (Reyes et al., 2020). All these improvements are challenging from several aspects, such as the data scarcity regarding additional inputs or outputs; the inherent difficulty in building a system addressing these tasks; and the design and use of appropriate evaluation methods.

Lastly, we want to expand our template-based proposal to other image modalities, body parts and languages beyond chest X-rays and English, which are less studied in the report generation literature (Messina et al., 2020). On the one hand, there are not many image-report datasets available in other sub-domains, and the existing ones usually contain less

images (Messina et al., 2020), since other image modalities are less common than chest X-rays in clinical practice (Jones et al., 2021). Thus, a suitable dataset may need to be procured, and a small dataset size may be an additional challenge for machine learning based systems. On the other hand, the template curation process could be generalized from our proposal to other sub-domains, by leveraging approaches in the literature that extract entities from medical reports. For example, using reports from chest CTs (Sugimoto et al., 2021), abdominopelvic from multiple modalities (Steinkamp et al., 2019), abdominal ultrasounds in Spanish (Cotik et al., 2017), or others (e.g. Báez et al., 2020). Covering other image modalities may be important in a clinical scenario, since some abnormalities cannot be properly detected using only one image type (e.g. chest X-rays) and require an additional imaging study (e.g. chest CT scans) (Hansell et al., 2008; Ginsberg, 2010). Moreover, addressing this task with different languages and different patient populations (e.g. patients from Latin America) could reveal and help solve issues with unintended biases toward protected communities, which have already been detected in similar medical imaging applications (Seyyed-Kalantari et al., 2021; I. Banerjee et al., 2021).

7. CONCLUSIONS

We address the task of automatically generating a text report from chest X-rays, by challenging the traditional NLP metrics used in the literature and favoring medical correctness metrics instead, and by proposing a template-based model much simpler than the state-of-the-art.

As experiments, we first benchmark the template-based model against naive baselines, encoder-decoder deep learning models and the SOTA in the IU X-ray and MIMIC-CXR datasets, using CheXpert to measure clinical correctness and using traditional NLP metrics, namely BLEU, ROUGE-L and CIDEr-D. The template-based model achieved the highest clinical performance, though lower NLP performance than SOTA. NLP metrics were unreliable in multiple cases, as (1) we were able to manipulate reports to improve the NLP performance of the template-based models without changing their clinical meaning, and (2) naive models were able to achieve comparable NLP scores to the SOTA.

Second, we proposed a simple stress test to check the validity of clinical correctness metrics, and we suggested more ways to create robustness tests. We tested CheXpert and MIRQI metrics, and found only the former passes the test, while the latter has important vulnerabilities to gaming effects or adversarial inputs.

Third, we contrasted traditional NLP metrics against corpora with controlled clinical meanings, showing NLP metrics poorly discriminate sentences with opposite clinical meaning regarding CheXpert abnormalities. Specifically, NLP metrics are not very good at discriminating true positive sentences (ROC-AUC values between $0.61 - 0.81$ in MIMIC), and are worse for true negative sentences (ROC-AUC values $0.50 - 0.57$). From all experiments, CIDEr-D is the least fragile NLP metric.

Overall, our experiments indicate NLP metrics are not enough to be used isolated in the report generation task, since they are not clinically reliable in multiple cases. In particular, the reports can be manipulated to resemble more the ground truth and achieve better performance, without changing their underlying clinical meaning, or without presenting useful clinical information. Additionally, our proposed template-based model is much simpler and more inherently interpretable than the state-of-the-art, and surpasses its contestants in clinical performance, measured by the CheXpert labeler metric.

In conclusion, we argue medical correctness metrics are more appropriate to be used as primary evaluations in this field, to focus on assessing the clinically relevant content of the reports, and traditional NLP metrics can be used in complement or secondarily. Furthermore, evaluation methods in general in this topic need to be more extensive, to advance towards deploying CAD systems in a clinical setting in the near future.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Proceedings of the 32nd international conference on neural information processing systems* (p. 9525–9536). Red Hook, NY, USA: Curran Associates Inc.

Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., & Fahmy, A. (2021). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, *24*, 100557. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352914821000472 doi: https://doi.org/10.1016/j.imu.2021.100557

Allaouzi, I., Ben Ahmed, M., Benamrou, B., & Ouardouz, M. (2018). Automatic caption generation for medical images. In *Proc of the 3rd intl. conf. on smart city applications.* New York, NY, USA: ACM.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *Computer vision – eccv 2016* (pp. 382–398). Cham: Springer Intl. Publishing.

Ayesha, H., Iqbal, S., Tariq, M., Abrar, M., Sanaullah, M., Abbas, I., ... Hussain, S. (2021). Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, *114*, 107856. Retrieved from https://www.sciencedirect.com/science/article/pii/S0031320321000431 doi: https://doi.org/10.1016/j.patcog.2021.107856

Babar, Z., van Laarhoven, T., Zanzotto, F. M., & Marchiori, E. (2021). Evaluating diagnostic content of ai-generated radiology reports of chest x-rays. *Artificial Intelligence in Medicine*, *116*, 102075. Retrieved from https://www.sciencedirect.com/science/article/pii/S0933365721000683 doi: https://doi.org/10.1016/j.artmed.2021.102075

Báez, P., Villena, F., Rojas, M., Durán, M., & Dunstan, J. (2020, November). The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd clinical natural language processing workshop* (pp. 291–300). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.clinicalnlp-1.32 doi: 10.18653/v1/2020.clinicalnlp-1.32

Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., ... others (2021). Reading race: Ai recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*.

Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). Ann Arbor, Michigan: ACL.

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., & Rudin, C. (2021). *Interpretable mammographic image classification using case-based reasoning and deep learning.* (In IJCAI-21 Workshop on Deep Learning, Case-Based Reasoning, and AutoML: Present and Future Synergies)

Biswal, S., Xiao, C., Glass, L. M., Westover, B., & Sun, J. (2020). Clara: Clinical report auto-completion. In *The web conf.* doi: 10.1145/3366423.3380137

Blazek, P. J., & Lin, M. M. (2021, Sep 01). Explainable neural networks that simulate reasoning. *Nature Computational Science*, *1*(9), 607-618. Retrieved from https://doi.org/10.1038/s43588-021-00132-w doi: 10.1038/s43588-021-00132-w

Boag, W., Hsu, T.-M. H., Mcdermott, M., Berner, G., Alesentzer, E., & Szolovits, P. (2020). Baselines for Chest X-Ray Report Generation. In *Ml4h at neurips.*

Branko, M., Danijela, B., & Dušan, S. (2010, Jan 01). Retrieval of bibliographic records using apache lucene. *The Electronic Library*, *28*(4), 525-539.

Bustos, A., Pertusa, A., Salinas, J.-M., & de la Iglesia-Vayá, M. (2019). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv:1901.07441*.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017, 08). Unintended Consequences of Machine Learning in Medicine. *JAMA*, *318*(6), 517-518. Retrieved from https://doi.org/10.1001/jama.2017.7797 doi: 10.1001/jama.2017.7797

Castro, V., Pino, P., Parra, D., & Lobel, H. (2021). Puc chile team at caption prediction: Resnet visual encoding and caption classification with parametric relu. In *Clef2021 working notes. ceur workshop proceedings.*

Charniak, E., & Johnson, M. (2005, June). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)* (pp. 173–180). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P05-1022 doi: 10.3115/1219840.1219862

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *CoRR*, *abs/1504.00325*. Retrieved from http://arxiv.org/abs/1504.00325

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *Emnlp*. doi: 10.18653/v1/2020.emnlp-main.112

Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., & Xu, F. (2017, September). Annotation of entities and relations in Spanish radiology reports. In *Proceedings of the international conference recent advances in natural language processing, RANLP 2017* (pp. 177–184). Varna, Bulgaria: INCOMA Ltd. Retrieved from https://doi.org/10.26615/978-954-452-049-6_025 doi: 10.26615/978-954-452-049-6_025

Datta, S., & Roberts, K. (2020). A dataset of chest x-ray reports annotated with spatial role labeling annotations. *Data in Brief*, *32*, 106056. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352340920309501 doi: https://doi.org/10.1016/j.dib.2020.106056

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal stanford dependencies: A cross-linguistic typology. In *Lrec* (Vol. 14, pp. 4585–4592).

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *JAMIA*. doi: 10.1093/jamia/ocv080

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). Imagenet: A large-scale hierarchical image database. In *Cvpr*. doi: 10.1109/CVPR.2009.5206848

Denkowski, M., & Lavie, A. (2010). Extending the meteor machine translation evaluation metric to the phrase level. In *Human language technologies: The 2010 annual conf. of the north american chapter of the acl* (pp. 250–253). USA: ACL.

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proc of the sixth workshop on*

statistical machine translation (pp. 85–91). USA: ACL.

Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In *Proc of the ninth workshop on statistical machine translation* (pp. 376–380). Baltimore, Maryland, USA: ACL.

Eickhoff, C., Schwall, I., García Seco de Herrera, A., & Müller, H. (2017, September 11-14). Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. In *Clef2017 working notes.* Dublin, Ireland.

Feng, S., Azzollini, D., Kim, J. S., Jin, C.-K., Gordon, S. P., Yeoh, J., ... Wilson, B. (2021). Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, *3*(6), e210136. doi: 10.1148/ryai.2021210136

Forgues, G., Pineau, J., Larchevêque, J.-M., & Tremblay, R. (2014). Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop* (Vol. 2).

Ganeshan, D., Duong, P.-A. T., Probyn, L., Lenchik, L., McArthur, T. A., Retrouvey, M., ... Francis, I. R. (2018). Structured reporting in radiology. *Academic Radiology*, *25*(1), 66-73. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1076633217303628` doi: https://doi.org/10.1016/j.acra.2017.08.005

García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., & Müller, H. (2018, September). Overview of the ImageCLEF 2018 caption prediction tasks. In *Clef2018 working notes.* Avignon, France.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care.

The Lancet Digital Health, *3*(11), e745-e750. Retrieved from https://www.sciencedirect.com/science/article/pii/S2589750021002089 doi: https://doi.org/10.1016/S2589-7500(21)00208-9

Ginsberg, L. E. (2010). "if clinically indicated:" is it? *Radiology*, *254*(2), 324-325. Retrieved from https://doi.org/10.1148/radiol.09091736 (PMID: 20093506) doi: 10.1148/radiol.09091736

Gonzalez, C., Gotkowski, K., Bucher, A., Fischbach, R., Kaltenborn, I., & Mukhopadhyay, A. (2021). Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In M. de Bruijne et al. (Eds.), *Medical image computing and computer assisted intervention – miccai 2021* (pp. 304–314). Cham: Springer International Publishing.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120.

Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Müller, N. L., & Remy, J. (2008). Fleischner society: Glossary of terms for thoracic imaging. *Radiology*, *246*(3), 697-722. Retrieved from https://doi.org/10.1148/radiol.2462070712 (PMID: 18195376) doi: 10.1148/radiol.2462070712

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep residual learning for image recognition. In *Proc of the ieee conf. on computer vision and pattern recognition (cvpr)* (pp. 770–778). doi: 10.1109/CVPR.2016.90

Hoover, A. (1975). Stare database. *Available: http://www.ces.clemson.edu/ ahoover/stare*.

Hou, B., Kaissis, G., Summers, R. M., & Kainz, B. (2021). Ratchet: Medical transformer

for chest x-ray diagnosis and reporting. In M. de Bruijne et al. (Eds.), *Medical image computing and computer assisted intervention – miccai 2021* (pp. 293–303). Cham: Springer International Publishing.

Hou, D., Zhao, Z., Liu, Y., Chang, F., & Hu, S. (2021). Automatic report generation for chest x-ray images via adversarial reinforcement learning. *IEEE Access*, *9*, 21236-21250. doi: 10.1109/ACCESS.2021.3056175

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Cvpr.* doi: 10.1109/CVPR.2017.243

Huang, X., Yan, F., Xu, W., & Li, M. (2019). Multi-attention and incorporating background information model for chest x-ray image report generation. *IEEE Access*. doi: 10.1109/ACCESS.2019.2947134

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... others (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Aaai conf. on artificial intelligence.* doi: 10.1609/aaai.v33i01.3301590

Jain, S., Smit, A., Truong, S. Q., Nguyen, C. D., Huynh, M.-T., Jain, M., ... Rajpurkar, P. (2021). Visualchexbert: Addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the conference on health, inference, and learning* (p. 105–115). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3450439.3451862` doi: 10.1145/3450439.3451862

Jain, Saahil, Agrawal, Ashwin, Saporta, Adriel, Truong, Steven QH, Nguyen Duong, Du, Bui, Tan, ... Rajpurkar, Pranav (2021). *RadGraph: Extracting Clinical Entities and Relations from Radiology Reports (version 1.0.0).* PhysioNet. Retrieved 2021-11-12, from `https://physionet.org/content/radgraph/1.0.0/` (Type: dataset)

doi: 10.13026/HM87-5P47

Jing, B., Wang, Z., & Xing, E. (2019). Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *Acl.* doi: 10.18653/v1/P19-1657

Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports. In *Acl.* doi: 10.18653/v1/P18-1240

Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., & Horng, S. (2019). *Mimic-cxr-jpg-chest radiographs with structured labels (version 2.0.0).* PhysioNet. doi: 10.13026/8360-t248

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., ... Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*. doi: 10.1038/s41597-019-0322-0

Jones, C. M., Buchlak, Q. D., Oakden-Rayner, L., Milne, M., Seah, J., Esmaili, N., & Hachey, B. (2021). Chest radiographs and machine learning – past, present and future. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 538-544. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/1754-9485.13274` doi: https://doi.org/10.1111/1754-9485.13274

Kaur, N., Mittal, A., & Singh, G. (2021, Sep 09). Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey. *Multimedia Tools and Applications*. Retrieved from `https://doi.org/10.1007/s11042-021-11272-6` doi: 10.1007/s11042-021-11272-6

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., & Erdem, E. (2017, April). Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 1, long papers*

(pp. 199–209). Valencia, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E17-1019

Klein, D., & Manning, C. D. (2003, July). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics* (pp. 423–430). Sapporo, Japan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P03-1054 doi: 10.3115/1075096.1075150

Kougia, V., Pavlopoulos, J., Papapetrou, P., & Gordon, M. (2021, 04). RTEX: A novel framework for ranking, tagging, and explanatory diagnostic captioning of radiography exams. *JAMIA*. doi: 10.1093/jamia/ocab046

Langlotz, C. P. (2006). Radlex: a new method for indexing online educational materials. *Radiographics: a review publication of the Radiological Society of North America, Inc*, *26*(6), 1595.

Lavie, A., & Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proc of the second workshop on statistical machine translation* (pp. 228–231). USA: ACL.

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. In *Neurips.*

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Aaai conf. on artificial intelligence.* doi: 10.1609/aaai.v33i01.33016666

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out.*

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The unified medical language system. *Yearbook of Medical Informatics*, *2*(01), 41–51.

Liu, F., Ge, S., & Wu, X. (2021, August). Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 3001–3012). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.234 doi: 10.18653/v1/2021.acl-long.234

Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., & Sun, X. (2021, August). Contrastive attention for automatic chest X-ray report generation. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 269–280). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.23 doi: 10.18653/v1/2021.findings-acl.23

Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., & Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. In *Ml4h.*

Lovelace, J., & Mortazavi, B. (2020). Learning to generate clinically coherent chest X-ray reports. In *Emnlp.* doi: 10.18653/v1/2020.findings-emnlp.110

Lukaszewicz, A., Uricchio, J., & Gerasymchuk, G. (2016). The art of the radiology report: Practical and stylistic guidelines for perfecting the conveyance of imaging findings. *Canadian Association of Radiologists Journal*, *67*(4), 318-321. Retrieved from https://doi.org/10.1016/j.carj.2016.03.001 (PMID: 27451909) doi: 10.1016/j.carj.2016.03.001

Mathur, N., Baldwin, T., & Cohn, T. (2020, July). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th*

annual meeting of the association for computational linguistics (pp. 4984–4997). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.448 doi: 10.18653/v1/2020.acl-main.448

McClosky, D. (2010). *Any domain parsing: automatic domain adaptation for natural language parsing.* Brown University.

McDermott, M. B., Hsu, T. M. H., Weng, W.-H., Ghassemi, M., & Szolovits, P. (2020, 07–08 Aug). Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In F. Doshi-Velez et al. (Eds.), *Proceedings of the 5th machine learning for healthcare conference* (Vol. 126, pp. 913–927). PMLR. Retrieved from https://proceedings.mlr.press/v126/mcdermott20a.html

Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., ... Capurro, D. (2020). *A survey on deep learning and explainability for automatic image-based medical report generation.* (Accepted in ACM Computing Surveys.)

Miller, G. A. (1995, November). Wordnet: A lexical database for english. *Commun. ACM*, *38*(11), 39–41.

Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., & Jurafsky, D. (2021, June). Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 5288–5304). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.416 doi: 10.18653/v1/2021.naacl-main.416

Monshi, M. M. A., Poon, J., & Chung, V. (2020). Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, *106*, 101878.

Moradi, M., Guo, Y., Gur, Y., Negahdar, M., & Syeda-Mahmood, T. (2016). A cross-modality neural network transform for semi-automatic medical image annotation. In *Medical image computing and computer-assisted intervention – miccai 2016* (pp. 300–307). Cham: Springer Intl. Publishing.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic radiology*, *19*(2), 236–248.

Mork, J. G., Yepes, A. J. J., & Aronson, A. R. (2013). The nlm medical text indexer system for indexing biomedical literature. In *Ceur workshop proceedings* (Vol. 1094).

Najdenkoska, I., Zhen, X., Worring, M., & Shao, L. (2021). Variational topic inference for chest x-ray report generation. In M. de Bruijne et al. (Eds.), *Medical image computing and computer assisted intervention – miccai 2021* (pp. 625–635). Cham: Springer International Publishing.

Nguyen, H. T., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J., & Cheng, L. (2021). Automated generation of accurate\& fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*.

Ni, J., Hsu, C.-N., Gentili, A., & McAuley, J. (2020). Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays. In *Emnlp*. doi: 10.18653/v1/2020.findings-emnlp.176

Nishino, T., Ozaki, R., Momoki, Y., Taniguchi, T., Kano, R., Nakano, N., ... Nakamura, K. (2020, November). Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 2223–2236). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.findings-emnlp`

.202 doi: 10.18653/v1/2020.findings-emnlp.202

Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. *Academic Radiology*, *27*(1), 106-112. Retrieved from https://www.sciencedirect.com/science/article/pii/S107663321930488X (Special Issue: Artificial Intelligence) doi: https://doi.org/10.1016/j.acra.2019.10.006

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Re, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the acm conference on health, inference, and learning* (p. 151–159). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3368555.3384468 doi: 10.1145/3368555.3384468

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Acl.* doi: 10.3115/1073083.1073135

Pavlopoulos, J., Kougia, V., & Androutsopoulos, I. (2019, June). A survey on biomedical image captioning. In *Proceedings of the second workshop on shortcomings in vision and language* (pp. 26–36). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-1803 doi: 10.18653/v1/W19-1803

Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C. M., & Müller, H. (2021, September 21-24). Overview of the ImageCLEFmed 2021 concept & caption prediction task. In *Clef2021 working notes.* Bucharest, Romania: CEUR-WS.org.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., & Friedrich, C. M. (2018). Radiology objects in context (roco): A multimodal image dataset. In D. Stoyanov et al. (Eds.), *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis* (pp. 180–189). Cham: Springer Intl. Publishing.

Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., & Lu, Z. (2018). Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, *2018*, 188.

Pino, P., Parra, D., Besa, C., & Lagos, C. (2021). Clinically correct report generation from chest x-rays using templates. In C. Lian, X. Cao, I. Rekik, X. Xu, & P. Yan (Eds.), *Machine learning in medical imaging* (pp. 654–663). Cham: Springer International Publishing. Retrieved from `https://dparra.sitios.ing.uc.cl/pdfs/preprint_Pinoetal_MICCAI_2021.pdf` doi: 10.1007/978-3-030-87589-3_67

Pino, P., Parra, D., Messina, P., Besa, C., & Uribe, S. (2020). Inspecting state of the art performance and NLP metrics in image-based medical report generation. *arXiv preprint arXiv:2011.09257*. (In LXAI at NeurIPS 2020)

Pooch, E. H. P., Ballester, P., & Barros, R. C. (2020). Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In J. Petersen et al. (Eds.), *Thoracic image analysis* (pp. 74–83). Cham: Springer International Publishing.

Quintana, J. M., Florea, D., Deane, R., Parra, D., Pino, P., Messina, P., & Lobel, H. (2021). Puc chile team at tbt task: Diagnosis of tuberculosis type using segmented ct scans. In *Clef2021 working notes. ceur workshop proceedings.*

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... others (2017). *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.*

Reiter, E. (2018). A structured review of the validity of bleu. *Computational Linguistics*, *44*(3), 393–401. doi: 10.1162/coli_a_00322

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., Tengg-Kobligk, H. v., ... Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*. doi: 10.1148/ryai.2020190043

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020, July). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4902–4912). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.442 doi: 10.18653/v1/2020.acl-main.442

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... AIX-COVNET (2021, Mar 01). Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, *3*(3), 199-217. Retrieved from https://doi.org/10.1038/s42256-021-00307-0 doi: 10.1038/s42256-021-00307-0

Rodrigues, M., & Qureshi, Z. (2014). *The unofficial guide to radiology : chest, abdominal and orthopaedic x-rays, plus cts, mris and other important modalities*. London: Zeshan Quershi.

Rogers, F. B. (1963). Medical subject headings. *Bulletin of the Medical Library Association*, *51*(1), 114–116.

Rudin, C. (2019, May 01). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. Retrieved from https://doi.org/10.1038/s42256-019-0048-x doi: 10.1038/s42256-019-0048-x

Rus, V., & Lintean, M. (2012). An optimal assessment of natural language student input using word-to-word similarity metrics. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (pp. 675–676). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021).

"everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://dl.acm.org/doi/abs/10.1145/3411764.3445518

Schilling, R., Messina, P., Parra, D., & Lobel, H. (2021). Puc chile team at vqa-med 2021: approaching vqa as a classfication task via fine-tuning a pretrained cnn. In *Clef2021 working notes. ceur workshop proceedings.*

Schuit, G., Castro, V., Pino, P., Parra, D., & Lobel, H. (2021). Puc chile team at concept detection: K nearest neighbors with perceptual similarity. In *Clef2021 working notes. ceur workshop proceedings.*

Seah, J. C. Y., Tang, C. H. M., Buchlak, Q. D., Holt, X. G., Wardman, J. B., Aimoldin, A., ... Jones, C. M. (2021). Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multi-case study. *The Lancet Digital Health*, *3*(8), e496-e506. Retrieved from https://www.sciencedirect.com/science/article/pii/S2589750021001060 doi: https://doi.org/10.1016/S2589-7500(21)00106-0

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Iccv* (pp. 618–626). doi: 10.1109/ICCV.2017.74

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021, Dec 10). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*. Retrieved from https://doi.org/10.1038/s41591-021-01595-0 doi: 10.1038/s41591-021-01595-0

Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., & Lungren, M. (2020, November). Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 1500–1519). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.117 doi: 10.18653/v1/2020.emnlp-main.117

Steinkamp, J. M., Chambers, C., Lalevic, D., Zafar, H. M., & Cook, T. S. (2019, Aug 01). Toward complete structured information extraction from radiology reports using machine learning. *Journal of Digital Imaging*, *32*(4), 554-564. Retrieved from https://doi.org/10.1007/s10278-019-00234-y doi: 10.1007/s10278-019-00234-y

Sugimoto, K., Takeda, T., Oh, J.-H., Wada, S., Konishi, S., Yamahata, A., ... Matsumura, Y. (2021). Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, *116*, 103729. Retrieved from https://www.sciencedirect.com/science/article/pii/S1532046421000587 doi: https://doi.org/10.1016/j.jbi.2021.103729

Summers, R. M. (2021). Artificial intelligence of covid-19 imaging: A hammer in search of a nail. *Radiology*, *298*(3), E162-E164. Retrieved from https://doi.org/10.1148/radiol.2020204226 (PMID: 33350895) doi: 10.1148/radiol.2020204226

Syeda-Mahmood, T., Wong, K. C., Gur, Y., Wu, J. T., Jadhav, A., Kashyap, S., ... others (2020). Chest X-ray report generation through fine-grained label learning. In *Miccai.* doi: 10.1007/978-3-030-59713-9_54

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again* (1st ed.). USA: Basic Books, Inc.

van Miltenburg, E., Clinciu, M., Dušek, O., Gkatzia, D., Inglis, S., Leppänen, L., ... Wen,

L. (2021, August). Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th international conference on natural language generation* (pp. 140–153). Aberdeen, Scotland, UK: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.inlg-1.14

van Miltenburg, E., Lu, W.-T., Krahmer, E., Gatt, A., Chen, G., Li, L., & van Deemter, K. (2020, December). Gradations of error severity in automatic image descriptions. In *Proceedings of the 13th international conference on natural language generation* (pp. 398–411). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.inlg-1.45

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 5998–6008). Curran Associates, Inc.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Cvpr.* doi: 10.1109/CVPR.2015.7299087

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015, June). Show and tell: A neural image caption generator. In *Proc of the ieee conf. on computer vision and pattern recognition (cvpr)* (pp. 3156–3164).

Wang, G., Liu, X., Shen, J., Wang, C., Li, Z., Ye, L., ... Lin, T. (2021, Jun 01). A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and covid-19 pneumonia from chest x-ray images. *Nature Biomedical Engineering*, *5*(6), 509-521. Retrieved from https://doi.org/10.1038/s41551-021-00704-1 doi: 10.1038/s41551-021-00704-1

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017, July). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *The ieee conf. on computer vision and pattern recognition (cvpr)* (p. 3462-3471).

Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, *1*(2), 270–280.

Witten, I. H., Frank, E., & Hall, M. A. (Eds.). (2011). *Data mining: Practical machine learning tools and techniques*. Elsevier. Retrieved from `https://doi.org/10.1016/c2009-0-19715-5` doi: 10.1016/c2009-0-19715-5

Wu, J., Agu, N., Lourentzou, I., Sharma, A., Paguio, J., Yao, J. S., ... Moradi, M. (2021). *Chest imagenome dataset (version 1.0.0).* PhysioNet. Retrieved from `https://physionet.org/content/chest-imagenome/1.0.0/` doi: 10.13026/WV01-Y230

Wu, L., Wan, C., Wu, Y., & Liu, J. (2017). Generative caption for diabetic retinopathy images. In *2017 intl. conf. on security, pattern analysis, and cybernetics (spac)* (pp. 515–519).

Xiong, Y., Du, B., & Yan, P. (2019). Reinforced transformer for medical image captioning. In *Mlmi.* doi: 10.1007/978-3-030-32692-0_77

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proc of the 32nd intl. conf. on intl. conf. on machine learning - volume 37* (pp. 2048–2057). Lille, France: JMLR.org.

Xue, Y., Xu, T., Long, L. R., Xue, Z., Antani, S., Thoma, G. R., & Huang, X. (2018). Multimodal recurrent model with attention for automated radiology report generation. In

Miccai. doi: 10.1007/978-3-030-00928-1_52

Yang, X., Ye, M., You, Q., & Ma, F. (2021, August). Writing by memorizing: Hierarchical retrieval-based medical report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 5000–5009). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.387 doi: 10.18653/v1/2021.acl-long.387

You, D., Liu, F., Ge, S., Xie, X., Zhang, J., & Wu, X. (2021). Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In M. de Bruijne et al. (Eds.), *Medical image computing and computer assisted intervention – miccai 2021* (pp. 72–82). Cham: Springer International Publishing.

Zhang, Y., Ding, D. Y., Qian, T., Manning, C. D., & Langlotz, C. P. (2018). Learning to summarize radiology findings. In *Louhi at neurips.* doi: 10.18653/v1/W18-5623

Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., & Xu, D. (2020). When radiology report generation meets knowledge graph. *AAAI Conf. on Artificial Intelligence*. doi: 10.1609/aaai.v34i07.6989

Zhou, L., Yin, X., Zhang, T., Feng, Y., Zhao, Y., Jin, M., ... Lu, L. (2021). Detection and semiquantitative analysis of cardiomegaly, pneumothorax, and pleural effusion on chest radiographs. *Radiology: Artificial Intelligence*, *3*(4), e200172. doi: 10.1148/ryai.2021200172

APPENDIX

A. OTHER MODALITIES DATASETS

Table A.1 shows the detail on multiple publicly available datasets from the literature (Messina et al., 2020) that contain images from other modalities and body parts than chest X-rays.

Table A.1. Datasets with other image modalities and body parts. Number of pairs indicate *image-report* pairs.

Dataset	**Image Type**	**# Samples**
ImageCLEF Caption 2017 (Eickhoff et al., 2017)	Biomedical: extracted from PubMed Central papers and automatically filtered non-clinical images	Pairs: 184,614
ImageCLEF Caption 2018 (García Seco de Herrera et al., 2018)	Biomedical: same process as 2017, improved filtering	Pairs: 232,305
ROCO (Pelka et al., 2018)	Multiple radiology: CT, Ultrasound, X-Ray, Fluoroscopy, PET, Mammography, MRI, Angiography and PET-CT	Pairs: 81,825
ImageCLEF Caption 2021 (Pelka et al., 2021)	Multiple radiology modalities and body parts	Pairs: 3,700
PEIR Gross (Jing et al., 2018)	Gross lesions	Pairs: 7,442
INBreast (Moreira et al., 2012), in Portuguese	Mammography X-ray	Images: 410 Reports: 115 Patients: 115
STARE (Hoover, 1975)	Retinal fundus	Pairs: 400

Table B.1. Sentences in the *grouped* template sets. If all abnormalities are absent, the template is used; and repeat this step until all groups are consumed. Lastly, fill with individual sentences for abnormalities that have not been covered.

	Abnormalities	**Template**
IU X-ray	Cardiomegaly, Enlarged Cardiom.	*The heart size and mediastinal silhouette are within normal limits.*
	Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other	*The lungs are clear.*
	Pneumothorax, Pleural Effusion, Lung Opacity	*There is no pneumothorax or pleural effusion. No focal airspace disease.*
MIMIC-CXR	Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis	*The lungs are clear.*
	Consolidation, Pleural Effusion, Pneumothorax	*There is no focal consolidation, pleural effusion, or pneumothorax.*
	Pleural Effusion, Pneumothorax	*There is no pleural effusion or pneumothorax.*
	Consolidation, Pneumothorax	*There is no focal consolidation or pneumothorax.*

B. MATERIALS

Table B.1 shows the sentences in the *grouped* set of templates, Table B.2 shows the fallback individual sentences used for the *grouped* model in the MIMIC-CXR dataset.

Table B.2. Individual sentences to fallback in the Template *grouped* model in MIMIC-CXR. Manually curated using common sentences or words from the MIMIC-CXR training set.

Abnormality	**Absence template**	**Presence template**
Cardiomegaly	*Heart size is normal*	*Moderate cardiomegaly is stable*
Enlarged Cardiomed.	*The cardiomediastinal silhouette is normal*	*Cardiomediastinal silhouette is stable*
Consolidation	*The lungs are clear without focal consolidation*	*Underlying consolidation cannot be excluded*
Lung Opacity	*No parenchymal opacities*	*There is a persistent left retrocardiac opacity*
Atelectasis	*No atelectasis*	*There is bibasilar atelectasis*
Pleural Effusion	*No pleural effusions*	*There are small bilateral pleural effusions*
Pleural Other	*There is no evidence of fibrosis*	*There is biapical pleural thickening*
Pneumonia	*No pneumonia*	*In the appropriate clinical setting, superimposed pneumonia could be considered*
Pneumothorax	*There is no pneumothorax*	*There is a small left apical pneumothorax*
Edema	*There is no pulmonary edema*	*There is mild pulmonary vascular congestion*
Lung Lesion	*No lung nodules or masses*	*Multiple bilateral lung nodules are again demonstrated*
Fracture	*No displaced fracture is seen*	*Multiple bilateral rib fractures are again noted*
Support Devices	–	*Tube in standard placement*

C. CLINICALLY CONTROLLED CORPORA RESULTS

C.1. Sentences statistics

Table C.1 shows the amount of sentences per classification of CheXpert in each dataset.

Table C.1. Number of sentences labeled by CheXpert output in each dataset.

	IU X-ray (6,435 total)				**MIMIC-CXR** (361,440 total)			
Abnormality	**None**	**Neg**	**Unc**	**Pos**	**None**	**Neg**	**Unc**	**Pos**
Enlarged Cardiom.	5,905	320	99	111	351,495	1,305	2,832	5,808
Cardiomegaly	5,888	263	29	255	333,551	3,243	2,709	21,937
Lung Lesion	6,186	72	15	162	352,679	702	1,287	6,772
Lung Opacity	5,380	291	59	705	299,540	2,715	3,256	55,929
Edema	6,296	75	37	27	319,724	9,027	10,573	22,116
Consolidation	6,117	279	8	31	343,118	3,523	3,440	11,359
Pneumonia	6,336	43	22	34	322,119	7,626	17,595	14,100
Atelectasis	6,142	5	91	197	308,840	1,547	10,528	40,525
Pneumothorax	6,034	365	11	25	343,141	8,308	1,052	8,939
Pleural Effusion	5,796	486	40	113	300,997	7,726	5,508	47,209
Pleural Other	6,374	7	13	41	358,538	119	725	2,058
Fracture	6,262	49	12	112	356,119	621	621	4,079
Support Devices	6,175	24	0	236	297,284	4,137	383	59,636

C.2. Sampling strategy to create a corpus

Given two sets of sentences, S_{gt} and S_{gen}, to be used as ground truth and generated sentences, respectively, the sampling strategy goes as follows. Sample N random sentences from S_{gt} (without replacement) to be used as ground truth. Then, for each ground truth sentence sample k sentences from S_{gen} (without replacement), and form all possible pairs. End up with $N \times k$ pairs of sentences.

C.3. All score matrices

Figures C.1, C.2, C.3 and C.4 present 4×4 and 2×2 score matrices for all abnormalities with all NLP metrics in both datasets.

C.4. Score distributions

Figures C.5 and C.6 show 4×4 matrices and NLP score distributions for *Cardiomegaly* in MIMIC-CXR and *Fracture* in IU X-ray. Distributions are shown in box plots for easier

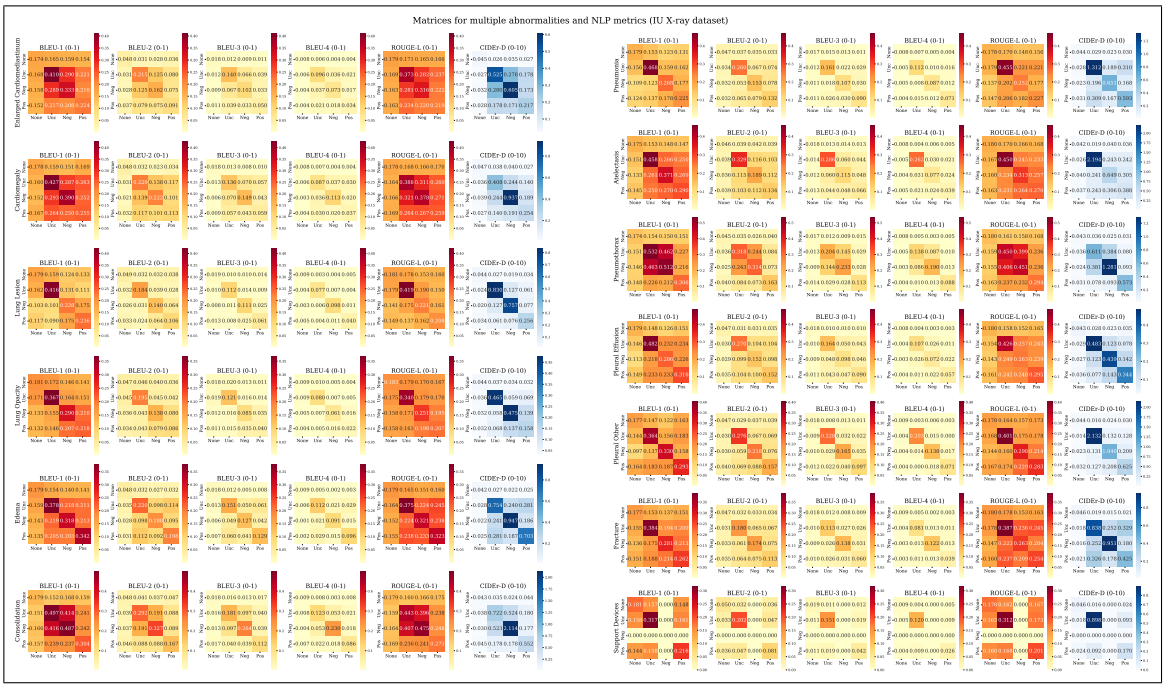


Figure C.1. 4×4 matrices with scores for all abnormalities and all NLP metrics in the IU X-ray dataset.

visualization. In most cases, NLP metrics do not discriminate well sentences with different meanings. The rest of abnormalities and metrics are omitted for brevity (these plots were randomly picked, not cherry picked).

Figures C.7 and C.8 show 2×2 matrices and NLP scores distributions for some abnormalities in both datasets. In most cases, histograms for the positive ground truth cases are separated somewhat better than for the negative ground truth.

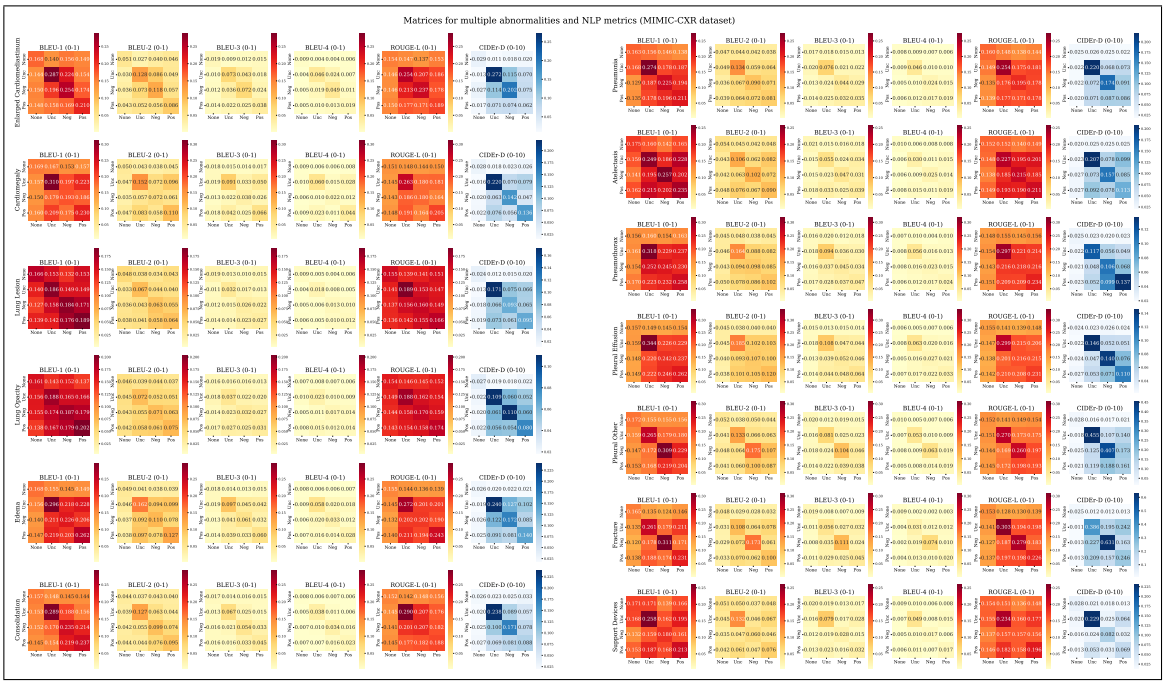


Figure C.2. 4×4 matrices with scores for all abnormalities and all NLP metrics in the MIMIC-CXR dataset.

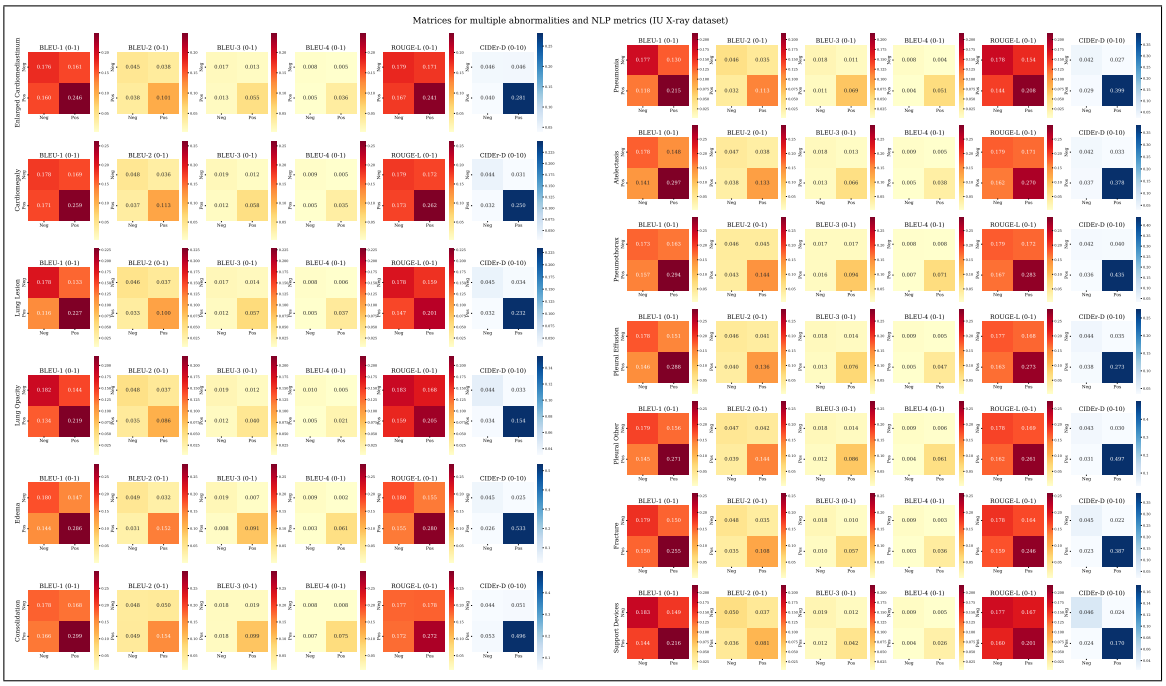


Figure C.3. 2×2 matrices with scores for all abnormalities and all NLP metrics in the IU X-ray dataset.

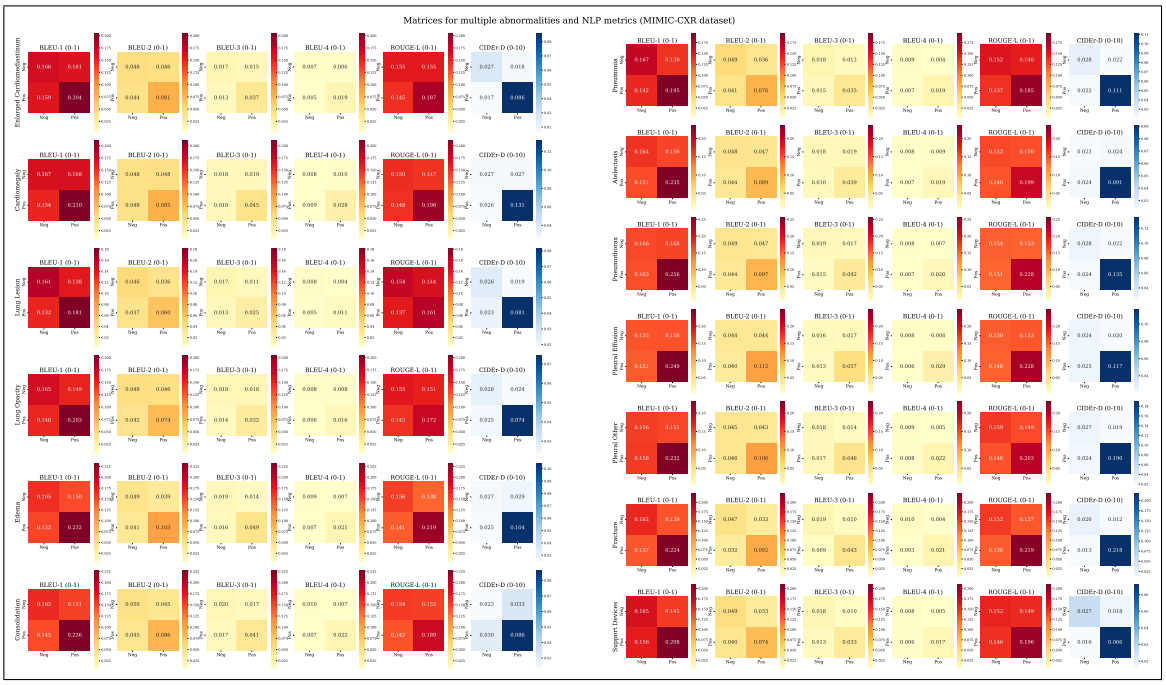


Figure C.4. 2×2 matrices with scores for all abnormalities and all NLP metrics in the MIMIC-CXR dataset.

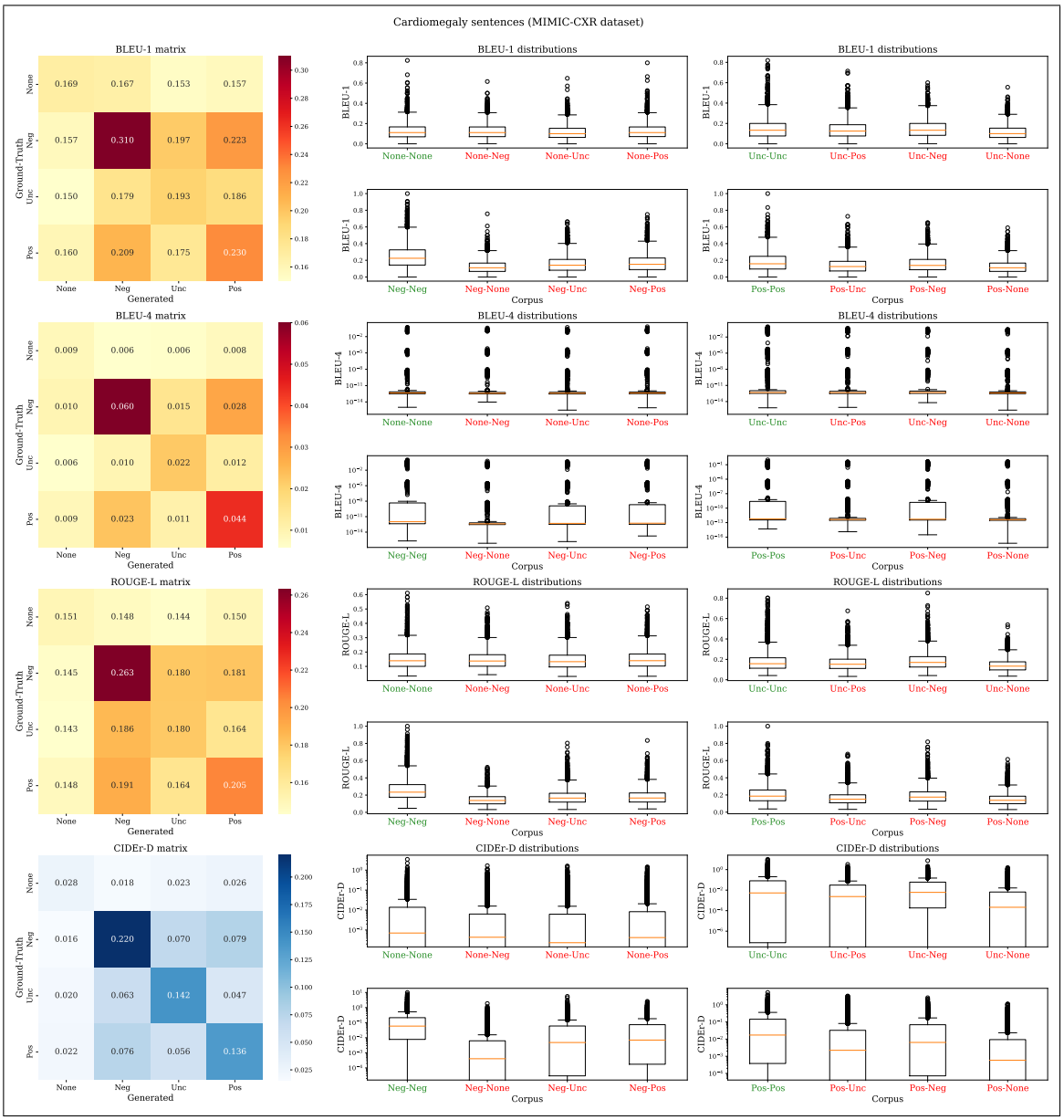


Figure C.5. 4×4 matrices and NLP scores distributions for *Cardiomegaly* in the MIMIC-CXR dataset. BLEU-4 and CIDEr-D histograms are shown in log-scale.

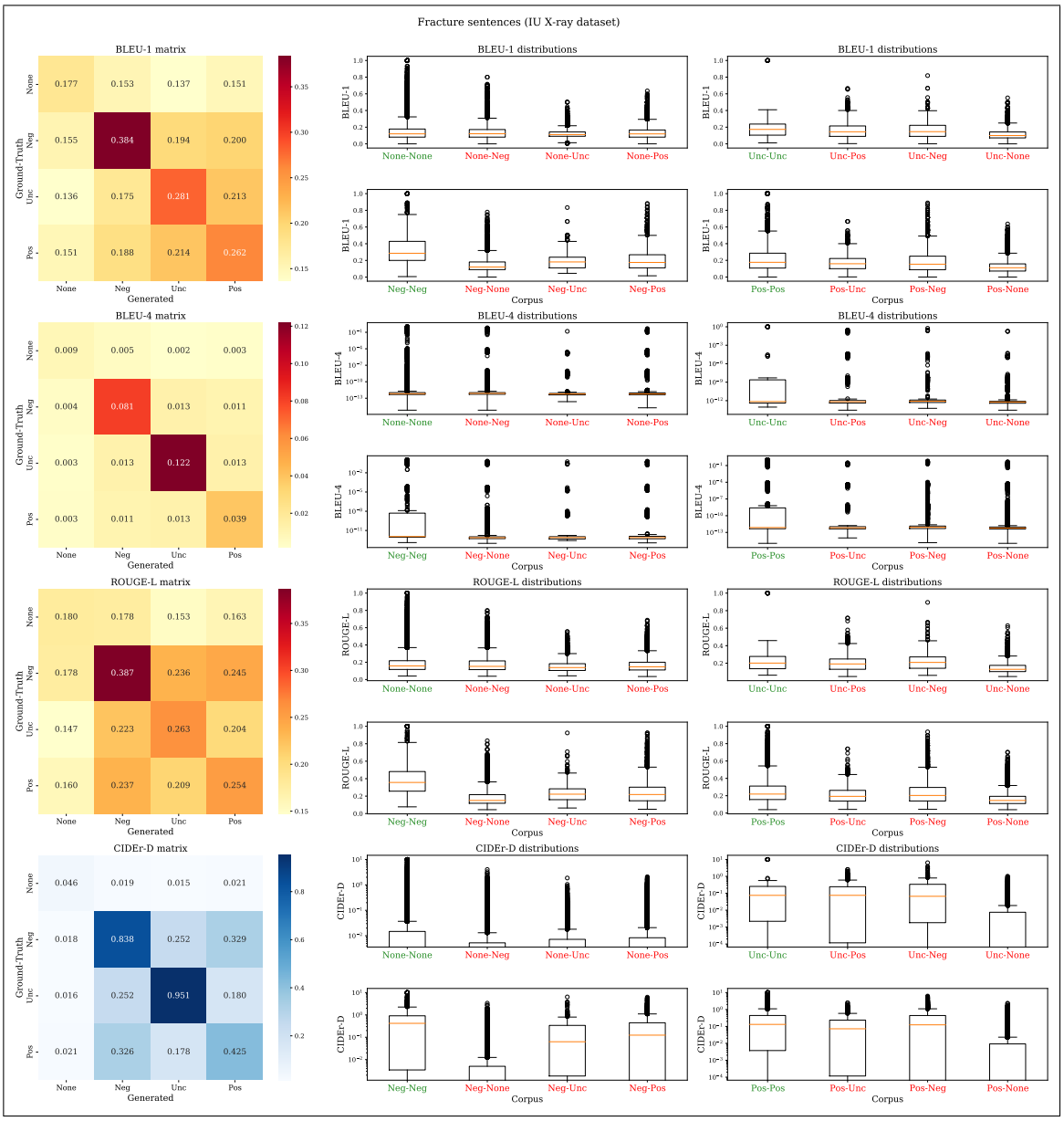


Figure C.6. 4×4 matrices and NLP scores distributions for *Fracture* in the IU X-ray dataset. BLEU-4 and CIDEr-D histograms are shown in log-scale.

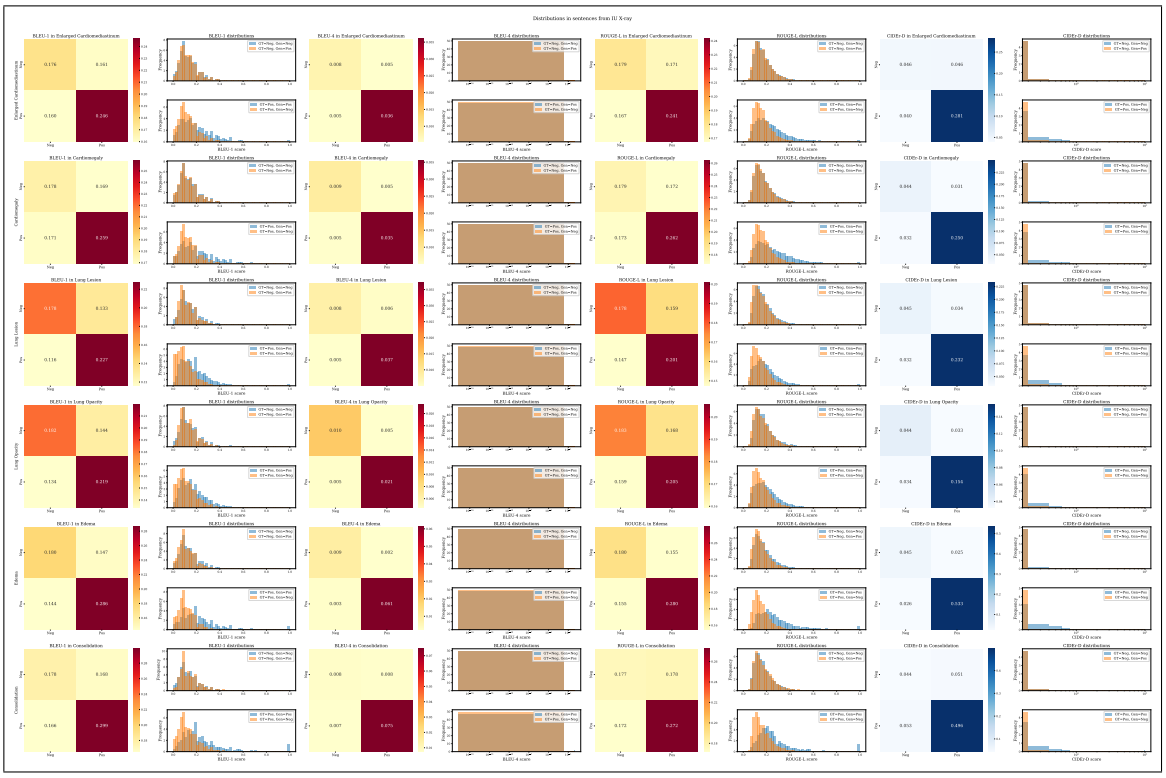


Figure C.7. 2×2 matrices and NLP scores distributions for some abnormalities in the IU X-ray dataset. BLEU-4 and CIDEr-D histograms are shown in log-scale.

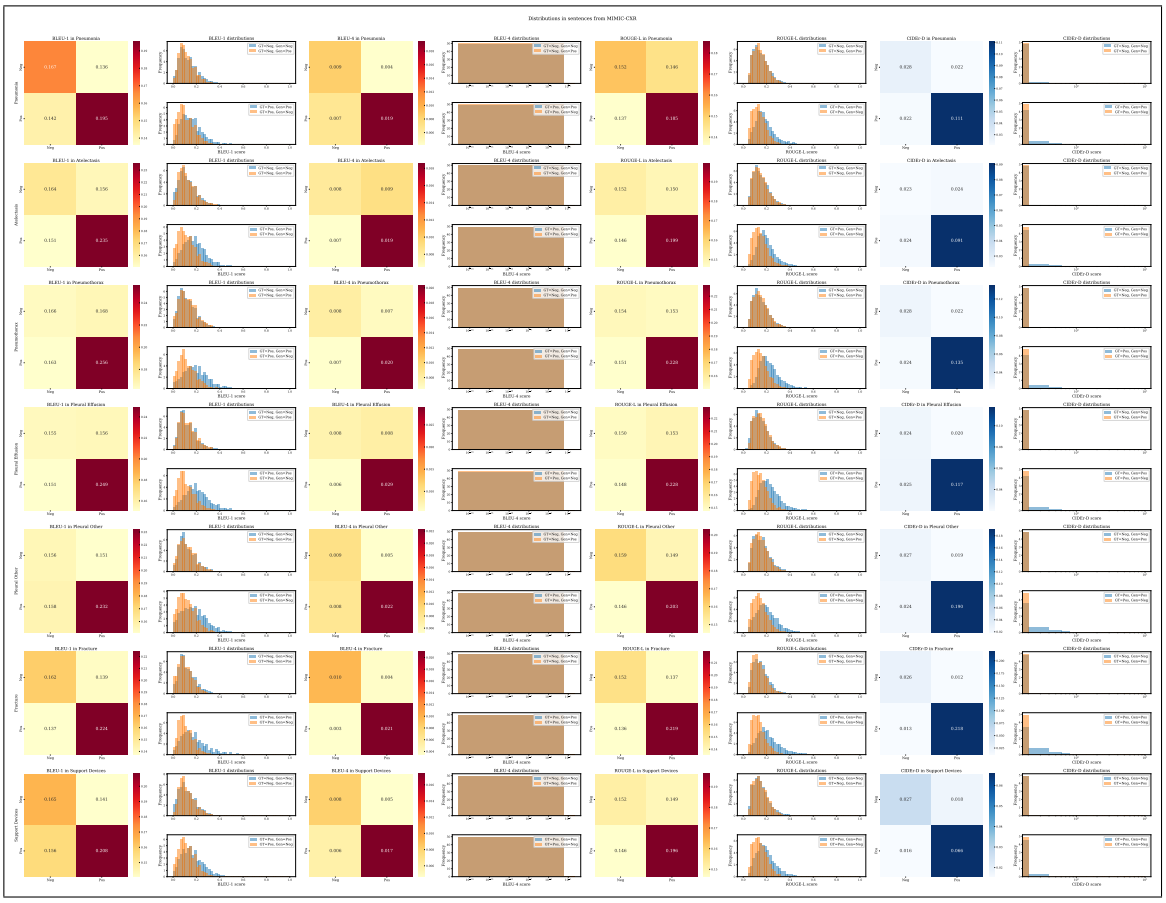


Figure C.8. 2×2 matrices and NLP scores distributions for some abnormalities in the MIMIC-CXR dataset. BLEU-4 and CIDEr-D histograms are shown in log-scale.