Text Mining of Pre-Exam Narratives: Unveiling Clinical Indicators for Chronic Obstructive Pulmonary Disease & Heart Failure

Lasse Hyldig Hansen[1] (LHH) - Study Nr: 201908623
Nikolaj Munch Andersen[1] (NMA) - Study Nr: 201908285

1 - School of Communication and Culture, Aarhus University, DK-8000 Aarhus C

Data Science, Prediction and Forecasting - Exam Paper
Total no. characters: 50905 Corresponding to 21.21 standard pages

June 2, 2023

Abstract

Chronic obstructive pulmonary disease (COPD) and heart failure (HF) pose significant healthcare challenges due to their high prevalence, severity and symptom overlap. This study aimed to uncover distinct clinical indicators pertaining to COPD and HF by means of text mining of patients' medical histories and their correlation with radiological observations in chest x-ray examinations. To achieve this goal, a methodology based on sentence-level text encoding was employed, wherein semantic vector representations were utilized to cluster topics in the pre-examination reports, enabling an evaluation of linguistic attributes associated with the topics. Through this approach, a total of 19 distinct topics were identified. Further, we utilized the identified topics as features within a logistic regression model to determine the effect of each topic on the occurrence of the respective disease outcomes (COPD & HF). This revealed significant gender disparities in the association between topics and disease outcome, with male patients showing a higher prevalence of COPD for specific topics. To investigate the notion that a topic-based approach could provide more transparent and comprehensible insights compared to complex deep learning models, we fine-tuned a transformer-based model for text classification of disease outcomes. While the model displays decent accuracy in its predictions (AUC = 0.79), an assessment of its learned weights underlined the trade-off between prediction and explanation prevalent in deep learning systems. This raises issues of interpretability and trust in healthcare decision-support tools. Based on insights from medical literature, it is discussed how these results highlight the potential of semantic clustering in extracting clinically relevant information, and underline the importance of providing interpretable diagnostic outcome predictions. Finally, these findings underpin the necessity for future research to focus on applying the method to more extensive journals with multiple disease outcomes.

1 Introduction

1.1 Machine Learning in Healthcare

(NMA) Chronic Obstructive Pulmonary Disease (COPD) and Heart Failure (HF) are prevalent and severe chronic health conditions that impose a significant burden on healthcare systems across the world. COPD alone stands as the third leading cause of death globally and results in $\sim$ 3.2 million deaths annually (World Health Organization, 2023). Due to the similarity in symptoms for the two diseases (Theander et. al, 2014), they can be hard to distin-

guish without further physical examination of the patient and radiological imaging of the lungs (Washko, 2010). Furthermore, as these diseases require distinct treatment approaches, ensuring an accurate diagnosis is imperative for targeted interventions and the development of personalized management plans. This overlap in symptoms and the necessity for tailored treatments based on diagnosis holds true for numerous other respiratory conditions as well. Consequently, it is no surprise that chest-X rays (CXRs) rank among the most frequently conducted radiological examinations (Wu et. al, 2021). The considerable amounts of data generated from CXRs can provide valuable insights into patterns and trends across a large number of patients. By analyzing this data, researchers and healthcare professionals can identify common characteristics, abnormalities, and associations between symptoms, radiological findings, and specific respiratory conditions (Çallı et al., 2021; Mozayan et al., 2021). **(LHH)** Advancements in the field of computational physiology have highlighted the potential of utilizing data scientific methods to enhance the diagnostic process and conduct large-scale data analysis of patient data (Dash et al., 2019). The steady release of open-source datasets is gradually paving the way for more efficient and accurate clinical decision support systems (Goldberger et al., 2000). Most relevant for the diagnostic process of COPD and HF, is the development of automated systems for abnormality detection and report generation using CXRs (Li et al., 2018; Smit et al., 2020; Wu et al., 2020). These advancements leverage machine learning algorithms and deep learning techniques to analyze medical images and text data to assist healthcare professionals in diagnosing diseases and conditions more efficiently and accurately. Fuelled by the growing availability of open-source datasets such as MIMIC-CXR (Johnson et al., 2019), these models are becoming increasingly better at predicting disease outcomes from CXRs (Ait Nasser & Akhloufi, 2023). However, utilizing deep learning techniques for decision support systems is not without its challenges. One of the primary downsides when employing complex machine-learning architectures in the medical field is the lack of explainability (Samek et al., 2017). Deep learning models are often considered black boxes, meaning they provide accurate predictions but fail to provide clear explanations or insights into how they arrived at those predictions. This lack of explainability raises concerns, particularly in the healthcare sector, where doctors rely on understanding the reasoning behind a prediction to make informed decisions (Singh et al., 2020). To try and address this concern, in the area of CXR disease prediction, Wu et. al (2021) has constructed the Chest ImaGenome dataset.

(NMA) The Chest ImaGenome dataset contains information regarding the relations between CXR attributes (labeled abnormalities) and their anatomical region. Additionally, the dataset contains sentences from radiologists describing their impressions and findings in lung regions, along with pre-exam text outlining the patient's history, what the authors refer to as 'reason for exams'. The dataset has already been leveraged to build promising decision support tools that employ graph neural networks (GNNs) that learn relationships between lung attributes across individual lung regions and disease outcomes (Agu et al., 2021). Additionally, multimodal models that leverage both post-exam report sentences, attributes, and regions have been trained for diagnostic purposes (McInerney et al., 2022). These model structures allow for an explainable approach, as it becomes easier to assess which areas and attributes the model weighs highest in its prediction. As of now, models trained on this data show great promise in terms of building explainable AI in radiology, however, do not explicitly delve into the content of patient history.

(NMA & LHH) In this paper, we aim to uncover to what extent indicative information can be extracted from patient history and assess if any pat-

terns emerge when coupling linguistic patterns in patient history with radiological findings in CXRs. We propose a novel approach that leverages text mining of prior patient history, henceforth referred to as pre-exam text, to address the diagnostic challenges associated with COPD and HF. By extracting pertinent information from the textual descriptions provided by medical professionals prior to a CXR exam, we aim to uncover key clinical indicators, symptoms, and risk factors specific to each condition. Specifically, this analysis involves encoding pre-exam texts based on their semantic similarity, reducing the dimensionality of these embeddings, and clustering them into various topics. These topics are first explored in terms of their characteristics, i.e. distribution of disease outcomes in each topic. Secondly, the topics along with other demographic information are being used as predictors in a logistic regression to assess the association between each topic in regards to the disease outcome. Thereafter, we explore potential gender disparities in the association between a topic and a disease outcome. This is done to elucidate how clinical indicators from pre-exam text interact with gender to explain the disease outcome. Additionally, we discuss and examine whether model coefficients for certain linguistic patterns in pre-exam texts are consistent with radiological literature. This serves as a means to validate the efficacy of the clustering, as well as assess the impact of these topics on disease outcomes. Finally, we use a state-of-the-art pre-trained language model RadBERT (Yan et al., 2022) to predict disease outcomes from pre-exam texts. This step is performed to assess the learned attention weights of the model for given words in the pre-exam text. These attention-based explanations of model predictions can reveal biases in the model that highlight the challenges one faces when relying solely on complex models for predictions. The analysis is exploratory in its nature and primarily serves to assess the efficacy of the methods involved when applied to short pre-exam texts. This research builds upon the area of explainable disease classification of CXRs and patient reports.

1.2 Natural Language Processing

(LHH) Text mining and topic modeling are subfields within the broader discipline of natural language processing (NLP), which involves computational methods for analyzing and modeling natural language data. Text mining focuses on extracting valuable information and knowledge from large textual collections. Similarly, topic modeling, as a form of text mining, automatically identifies hidden topics and recurring themes in a document collection. Converting raw text data into numerical representations that capture the semantic and syntactic relationship of words, phrases or entire documents, is a crucial step prior to most text-based machine learning applications (Q. Liu et al., 2020). These numerical representations, often referred to as text embeddings, take the form of n-dimensional vectors that allow algorithms to learn relationships between words and phrases. The gradual development of encoders that better capture semantic and contextual content in text, has fuelled the increasing performance of natural language models on downstream tasks. Milestone models such as Word2Vec (Church, 2017), GloVe (Pennington et al., 2014), FastText (Athiwaratkun et al., 2018) and ELMo (Peters et al., 2018) have successfully tackled this problem using different neural network architectures. However, despite their notable achievements, these models still face certain limitations. One limitation is their struggle to account for the shifting semantic nuances of words used in different contexts (Devlin et al., 2018). Moreover, when processing long sequences, these models may encounter difficulties in preserving valuable contextual information, potentially resulting in a loss of accuracy or relevance.

(NMA) In recent years the transformer architecture proposed by Vaswani et al. (2017) has revolutionized the field of NLP by introducing an architec-

ture that leverages a so-called self-attention mechanism to capture long-range dependencies in text. Furthermore, this architecture allows for better contextual embeddings that take into account the surrounding words and sentence structure, allowing for a more nuanced representation of each word's meaning within its specific context. Models that utilize this architecture to encode text, such as BERT (Devlin et al., 2018) and GPT (Brown et al., 2020), have achieved state-of-the-art results in a wide range of NLP tasks. Since the release of BERT, models with similar architecture like roBERTa (Y. Liu et al., 2019), have been released, showing that with slight modifications to the training process, even higher performance can be achieved across NLP tasks. One of the key advantages of models like BERT and roBERTa is their ability to leverage transfer learning. Transfer learning refers to the process of pre-training a model on a massive corpus of text and then fine-tuning it for specific tasks with a smaller labeled dataset (Zhuang et al., 2019). This adaptable approach has shown high effectiveness in NLP, as pre-trained models can be readily shared with the research community, equipped with extensive linguistic knowledge and contextual understanding.

(LHH) BERT embeddings are generated at the token level, representing individual words or subwords (Devlin et al., 2018), which is advantageous for training the model in downstream tasks like text classification. However, this granularity exhibits limitations when attempting to cluster sentences based on their semantic similarity. To address this limitation, researchers have proposed S-BERT (Reimers & Gurevych, 2019), a variant of BERT specifically trained on a sentence similarity task. S-BERT produces sentence-level textual representations that are better suited for the task of topic clustering. Consequently, utilizing a sentence-transformer framework like S-BERT allows for a richer representation of sentence meaning and semantic relationships compared to a standard BERT model (Reimers & Gurevych, 2019). By accounting for the semantic similarity between sentences, the model embeddings better capture the semantic content of a given sentence.

1.3 NLP in Radiology

(NMA) The utility of transformer-based models in NLP has brought a lot of attention to the discipline, and various research communities in bio and healthcare informatics have started leveraging the capabilities of these NLP models, both for research purposes and direct implementations in healthcare systems (Zhou et. al, 2022). Radiology is particularly well suited for NLP-based research approaches, since the primary communication between physicians, and from physician to patient, is done through free text radiology reports (Mozayan et. al, 2021). A lot of NLP research has been conducted using these reports to uncover patterns in large amounts of text data (Pons et. al, 2016).

(LHH) One such technique is topic modeling which has proved valuable for structuring and describing semantic patterns in radiology reports (Hassanpour & Langlotz, 2015; Zech et al., 2018; Denti, 2022). To the best of our knowledge, topic modeling is yet to be applied to pre-exam text from CXRs in the Chest Imagenome dataset and linked to certain disease outcomes. Most previous research on topic modeling for radiology reports mainly focuses on the ability to elucidate topics by clustering the text embeddings of thousands of reports. This is also the approach we take, however, we seek to not only uncover which topics are in the pre-exam text and how these topics are distributed but also to what extent these topics can be used to explain the disease outcomes of patients. This approach allows for a more explainable overview of why a pre-exam text results in a certain diagnosis.

2 Methodology

2.1 Data & Pre-processing

(NMA) In order to unveil clinical indicators, from pre-exam text reports, in the two conditions HF and COPD we use the Chest ImaGenome dataset (Wu et al., 2021), which is derived from the MIMIC-CXR database (Johnson et al., 2019). The Chest ImaGenome dataset consists of diseases extracted from CXR final free-text reports collected from radiologists in their routine workflow. Furthermore, the dataset also includes mappings between the disease attributes and the corresponding pre-exam texts. During the preprocessing phase, we identified that out of the total 242,072 CXRs, 179,922 contained a pre-exam text. These texts include information about the patient's prior medical history and the reasons they are sent to a CXR. Further, the metadata from each patient was also available from the dataset and includes information to distinguish individual patients, their gender, and disease categories. Afterward, we filtered the dataset to only contain texts from the two diseases COPD and HF, resulting in 6,173 pre-exam texts related to HF and 6,001 texts related to COPD. The preprocessed dataset comprises a total of 7,532 unique patients, with an average of 2.02 CXRs per patient. Pre-exam texts did on many occasions include gender and age indications which we did not want to be included in the analysis, accordingly, these were stripped from the texts.

2.2 Word Frequency Analysis

(LHH) Prior to the topic modeling, we conducted a brief analysis of the word frequencies in the pre-exam texts. This entailed compiling a corpus of all the texts and subsequently tabulating the frequency of each word across the corpus. The objective of this exploratory analysis was to obtain a comprehensive understanding of the most predominant words associated with each disease outcome. To assess the prevalence of specific words in relation to each condition, the ratio of their occurrences within each condition was computed. This analysis aimed to discern words that exhibited higher frequency in one condition relative to the other, as well as those that were commonly observed in both conditions. By examining the word frequency for each disease outcome, we can gain insights into the linguistic characteristics that may differentiate the two conditions. This analysis aids in ensuring that the subsequent topic modeling is built upon a solid foundation, validating the assumption that the diseases possess discernible linguistic variations. This analysis was conducted and visualized using the Scattertext library (Kessler, 2017).

2.3 Topic Modeling

(NMA) In order to understand the large amount of text data, and unveil clinical indicators from the two diseases in question we use topic modeling. Topic modeling is a valuable approach for this task since it offers a systematic way to uncover the underlying themes within a large text corpus. In the context of our analysis, topic modeling will help identify the key clinical indicators associated with the two diseases. Specifically, we use a transformer-based topic modeling called BERTopic (Grootendorst, 2022), the method consists of three important steps: sentence embeddings, dimensionality reduction, and clustering. By organizing the pre-exam texts into topics it facilitates the analysis of key clinical indicators, symptoms, and risk factors specific to COPD and HF.

2.4 Sentence Embeddings

(LHH) In order to encode the pre-exam texts into meaningful high-dimensional representations we use BERTopic to employ the SentenceTransformers Python framework (Reimers & Gurevych, 2019, 2020). The library contains several pre-trained models that are specifically trained and evaluated based on their ability to predict cosine similarity between

pairs of sentence embeddings. Among the available models in the framework, we selected the 'all-mpnet-base-v2' model, which has demonstrated superior performance on sentence similarity tasks (Nils Reimers, 2022). The 'all-mpnet-base-v2' model is trained using the mbnet-base (Song et al., 2020) initialization, on over 1 billion sentence pairs. It consists of a 12-layer transformer architecture with a hidden layer size of 768, and ∼110 million parameters. Consequently, this architecture produces sentence embeddings of 768 dimensions. All sentences in the pre-exam texts were encoded using this model. Subsequently, the generated embeddings were mapped back to their corresponding text and patient metadata, allowing us to retain the connection between the embeddings and their original context for further analysis and interpretation.

2.5 Dimensionality Reduction

(NMA) To achieve more efficient clustering of the 768-dimensional vector embeddings, it is crucial to reduce their dimensionality. In high-dimensional spaces, the available data points become scarce, leading to difficulties in accurately measuring distances or similarities between data points, a limitation known as 'the curse of dimensionality' (Köppen, 2000). In order to reduce the dimensionality in the sentence embeddings, we used the dimensionality reduction technique UMAP (McInnes et al., 2018). The purpose of UMAP in the context of topic modeling is to transform the high-dimensional sentence vectors into a lower-dimensional space while retaining the relevant semantic relationships between the documents.

2.6 Clustering

(LHH) After the dimensions of the embeddings have been reduced with UMAP, we cluster them by similarity in order to extract meaningful topics. These clusters were determined with HDBSCAN (McInnes et al., 2017). HDBSCAN is a density-based clustering algorithm that is able to identify clusters within data. It does so by making a hierarchy of clusters based on density connectivity. Accordingly, each sentence is assigned a cluster or labeled as noise (being an outlier). In order to ensure that only topics with a sufficient number of supporting documents are considered we set the parameter 'min_topic_size' to 25. This requires that each cluster contains a minimum of 25 sentences to form a valid topic. This is done to ensure that extracted topics are meaningful and representative of the clinical context we are exploring. Further, the number of clusters was set to 20 with the 'nr_topics' parameter, indicating the desired number of clusters to be identified by the algorithm. This is done for the sake of interpretability, as having a lower number of topics helps the resulting topics become more interpretable and useful.

2.7 Outlier Reduction

(NMA) To handle instances where HDBSCAN fails to assign an embedding to a specific cluster, thus labeling it as an outlier, we implemented an outlier reduction strategy within our topic model. Initially, the clustering process identified 4574 outliers, representing sentences that could not reasonably be assigned to any of the identified topics. In order to maximize the utilization of all available pre-exam texts, we employed a probabilistic strategy for outlier reduction. The outlier reduction strategy leverages the soft-clustering capabilities of HDBSCAN and aims to determine the most appropriate matching topic for each outlier. This process was carried out using the $.reduce\_outliers$ function provided by BERTopic.

2.8 Topic Clusters

(LHH) In order to create an accurate representation of a topic the top 5 most important words for a cluster are determined. This is done based on a modification of the measure 'term frequency inverse document frequency' (TF-IDF) (Aizawa, 2003). The

modified representation is called c-TF-IDF and tries to better understand the words that differentiate a cluster of pre-exam texts. By considering the specific characteristics of each cluster, c-TF-IDF enables us to capture the intrinsic properties of the topics and facilitate a more precise representation of the most important words from one topic:

$$c\text{-}TF\text{-}IDF = \|tf_{x,c}\| * \log\left(1 + \frac{A}{f_x}\right) \quad (1)$$

Where $tf_{x,c}$ = the number of times a word (x) appears in a cluster (c), A = average number of words in a cluster, & f_x = frequency of word (x) in all clusters (c)

(NMA) Accordingly, each topic cluster is now represented by the 5 words with the highest c-TF-IDF score. These chosen words were used to create concise and informative names that capture the common themes or characteristics shared by the most important words within each cluster. After reducing outliers and conducting the topic modeling, we identified a total of 19 topics at level 2 (See Appendix A). Considering the hierarchical nature of HDBSCAN, we further reduced the 19 topics into 5 topics at a higher hierarchical level. By grouping the 19 topic cluster names into 5 higher-level topics using this hierarchical dimension, we assigned names to each topic in the second hierarchy based on the most important words that encompass the higher-level structure. These 5 topics are now referred to as level 1 topics (See Appendix B). The hierarchical structure that links level 2 topics to level 1 topics can be seen in Appendix C.

2.9 Logistic Regression

(LHH) A logistic regression was conducted to evaluate the association between topics of the pre-exam reports and disease outcomes. Specifically, we used the level 2 topics as a categorical predictor, to explain disease outcomes of individual patients. The model can be represented as follows:

$$Disease_i = logit(\beta_0 + \beta_1 Topic_i + \epsilon_i) \quad (2)$$

Where $Topic_i$ = the 19 different level 2 topics, the intercept β_0 refers to the reference topic 'Pneumonia & Respiratory Symptoms', $Gender_i$ = a binary gender category, and:

$$\text{logit}(x) = \frac{1}{(1 + \exp(-x))} \quad (3)$$

(NMA) Further, we wanted to understand if the effect between the level 2 topics and disease categories was moderated by the gender of the patient. Therefore, in order to understand if these topics lead to different predictions in disease categories for males compared to females, we build a model with an interaction between gender and level 2 topics:

$$Disease_i = logit(\beta_0 + \beta_1 Topic_i + \beta_2 Topic : Gender_i + \epsilon_i) \quad (4)$$

Where $Topic_i$ = the 19 different level 2 topics, the intercept β_0 represents a reference topic being 'Pneumonia & Respiratory Symptoms', $Gender_i$ = a binary gender category

2.10 Disease Classification Using RadBERT

(LHH) As a final step in our analysis, we fine-tune RadBERT (Yan et al., 2022) to classify pre-exam text in the two disease labels. RadBERT is a transformer-based language model that is pre-trained with ∼4,42 million radiology reports, with a BioBERT (Lee et al., 2019) initialization. This model was chosen as it achieves higher scores than any other models for NLP tasks involving radiology reports (Yan et al., 2022). The model was fine-tuned using the machine-learning framework PyTorch (Paszke et al., 2017) and 'simpletransformers' (Rajapakse, 2019) a wrapping

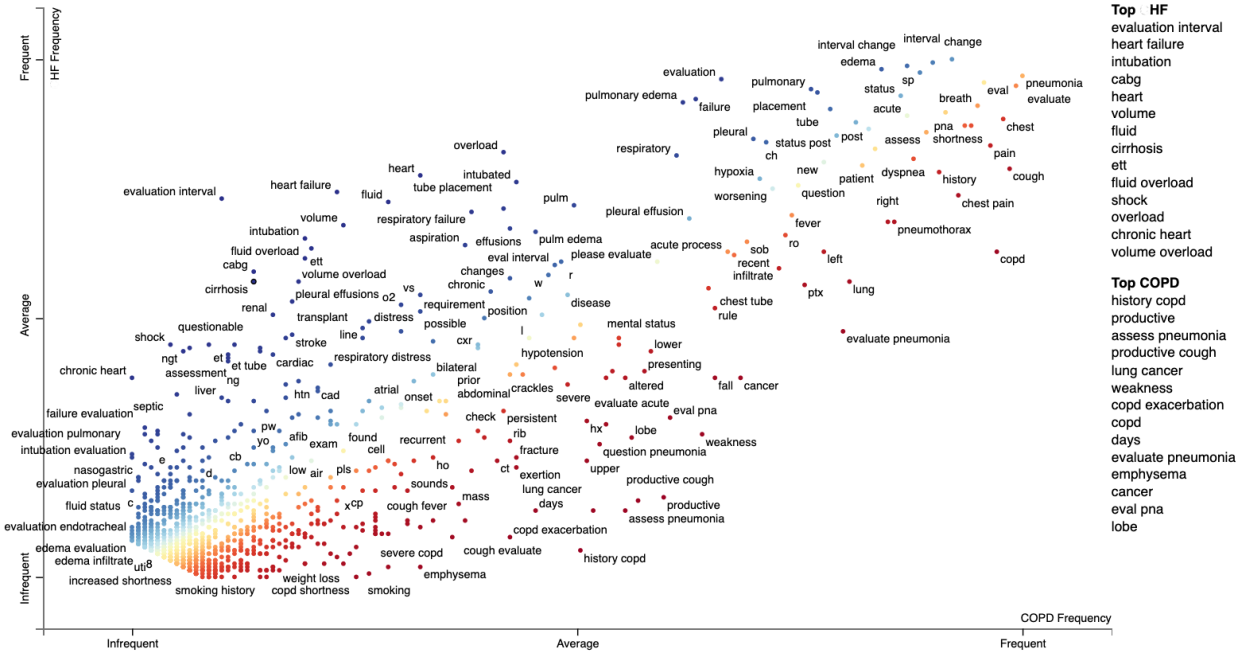

Figure 1: Scatter plot of word frequencies in pre-exam texts from patients with HF (y-axis) and COPD (x-axis).

library for huggingface's transformers library (Wolf et. al, 2020) that simplifies the process of importing pre-trained models and adding training arguments. We use 90% of the pre-exam text data for training, leaving ∼1.200 pre-exam texts in the test set to assess the model accuracy. Our model was fine-tuned using the following hyperparameters: batch size: 8, learning rate: 4e-5, no. of train epochs: 4. **(NMA)** In order to gain deeper insights into the inner workings of our trained classification model, we utilized the SHAP library (Lundberg & Lee, 2017) to evaluate the individual token contributions toward the model's prediction of disease labels. SHAP works by employing a token masking technique, where it systematically masks individual tokens, i.e. words, within a given pre-exam text input and observes the resulting change in the model's output. By comparing the predictions before and after token masking, SHAP is able to determine the specific impact of each token on the model's disease label prediction. Using the SHAP library we computed the average influence exerted by every word on each disease outcome. Furthermore, we applied shap explanations to individual pre-exam reports to gain sentence-level insights into cases where the model was very certain of a specific disease outcome. This approach serves as a means to evaluate whether the model successfully learns meaningful relationships between words that are relevant and coherent from a medical perspective.

2.11 Analysis & Model Code

(NMA & LHH) Information about used packages and softwares can be found in Appendix D. The code that produced the analysis can be found here: https://github.com/Lassehhansen/CXR-Clinical-Indicators

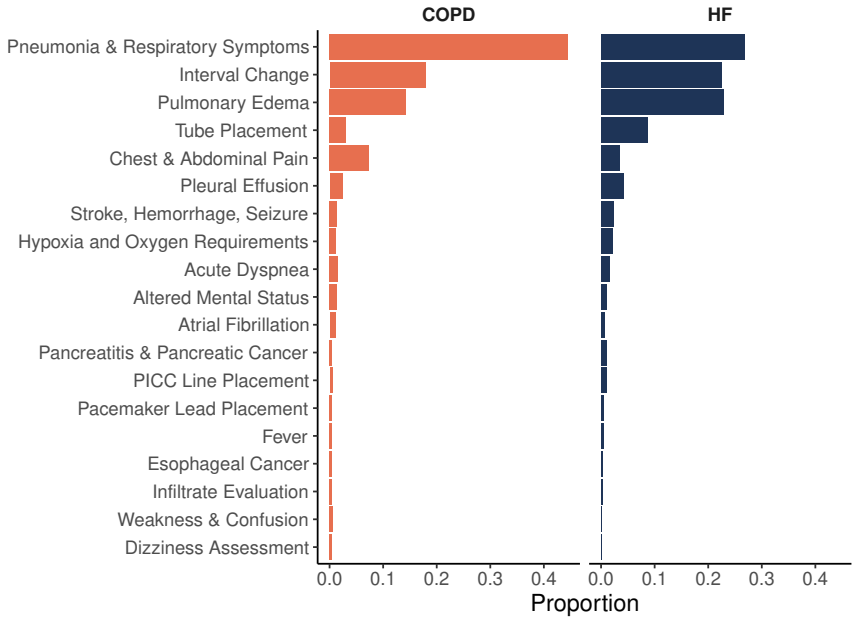

Figure 2: Bar plot comparing the proportional distribution of topics identified by a topic model for COPD and HF. Each bar represents the relative frequency of a specific topic within the respective disease.

3 Results

3.1 Word Frequencies

(LHH) The exploratory analysis using Scattertext revealed distinguishing terms between COPD and HF. The top-scoring words for each disease are calculated by comparing their relative occurrence in each disease, i.e. how often they appear in one condition compared to the other. The top-scoring words and n-grams for COPD were 'History COPD', 'Productive', 'Assess Pneumonia', and 'Productive Cough.' The frequency of these terms highlights the clinical focus on the patient's history of COPD, symptoms such as productive cough, and the need to assess for pneumonia. On the other hand, the most frequent terms found for HF included 'Evaluation interval', 'Heart Failure', 'Intubation', 'CABG', and 'Heart'. These terms reflect the emphasis on monitoring patients, a prior diagnosis of heart failure, and specific interventions like intubation and CABG (Coronary Artery Bypass Grafting).

3.2 Topic Modeling

(NMA) Topic modeling revealed that the three most frequent topics in the pre-exam narratives were: 'Pneumonia & Respiratory Symptoms,' 'Interval Change,' and 'Pulmonary Edema.' The topic 'Pneumonia & Respiratory Symptoms' was the most prevalent, appearing in 4330 documents, followed by 'Interval Change' (2459) and 'Pulmonary Edema' (2270). Further analysis was conducted to examine the distribution of these topics across the two diseases. Simple count analysis revealed that the three most frequent topics were more evenly distributed in HF compared to COPD (Figure 2). In HF, the proportions of the three most frequent topics were

as follows: 'Pneumonia & Respiratory Symptoms' (prop = 0.27), 'Interval Change' (prop = 0.22), and 'Pulmonary Edema' (prop = 0.23). Conversely, for COPD, the most dominant topic was 'Pneumonia & Respiratory Symptoms' (prop = 0.45). Showing that while 'Pneumonia & Respiratory Symptoms' is prominent in both diseases, the other two topics, 'Interval Change' and 'Pulmonary Edema,' are more evenly distributed and relatively more prevalent in HF.

3.3 Logistic Regression

(LHH) The logistic regression was conducted to determine the significance of specific topics in between COPD and HF. The model revealed that several topics were significantly more prevalent among patients with HF compared to COPD. These topics include Interval Change: OR = 0.48, 95%CI $[0.44, 0.53]$, $p < 0.001$, Pulmonary Edema: OR = 0.38, 95%CI $[0.34, 0.42]$, $p < 0.001$, Stroke, Hemorrhage, Seizure: OR = 0.35, 95%CI $[0.26, 0.46]$, $p < 0.001$, Hypoxia and Oxygen Requirements: OR = 0.34, 95% CI$[0.25, 0.45]$, $p < 0.001$, Acute Dyspnea: OR = 0.59, 95% CI$[0.44, 0.79]$, $p < 0.001$, Tube Placement: OR = 0.35, 95% CI$[0.26, 0.46]$, $p < 0.001$, PICC Line Placement: OR = 0.34, 95%CI $[0.22, 0.53]$, $p < 0.001$, Pleural Effusion: OR = 0.34, 95% CI$[0.28, 0.42]$, $p < 0.001$, and Pancreatitis & Pancreatic Cancer: OR = 0.25, 95% CI$[0.15, 0.39]$, $p < 0.001$. The presence of these topics in the narratives indicates their association with HF. Conversely, two topics were significantly associated with COPD compared to HF, these were: Chest & Abdominal Pain: OR = 1.28, 95% CI$[1.08, 1.53]$, $p < 0.001$ and Weakness & Confusion: OR = 4.48, 95% CI$[2.09, 14.1]$, $p < 0.001$. Model coefficients can be found in Figure 3a, and estimates in Appendix E.

(NMA) Furthermore, exploring the interaction between gender and topic, revealed that four topics showed stronger associations with COPD among males compared to females: Pneumonia & Respiratory Symptoms: OR = 1.17, 95% CI$[1.04, 0.42]$, $p < 0.001$, Chest & Abdominal Pain: OR = 1.46, 95% CI$[1.05, 2.03]$, $p = 0.024$, PICC Line Placement: OR = 3.23, 95% CI$[1.30, 8.53]$, $p = 0.014$ and Fever: OR = 3.61, 95% CI$[1.18, 11.9]$, $p = 0.028$. Model coefficients can be found in Figure 3b. Model comparison using the Akaike information criterion (AIC) supports including a gender interaction with the topics, as the model incorporating this interaction had a lower AIC value (16074) compared to the model without the interaction (16079).

3.4 Embeddings-based Disease Classification

(LHH) The results of the model evaluation on the test data demonstrated that the model was able to distinguish between the two diseases COPD and HF. The fine-tuned RadBERT model had an Area Under the Curve (AUC) value of 0.79. This indicates that the fine-tuned RadBERT effectively learned disease-specific patterns and features, enabling moderately accurate disease classification. The subsequent assessment of the average SHAP values, i.e. word importance towards predicted disease, shows that the top 5 most salient words for the outcome of COPD were 'preop', 'Rales', 'myalgias', 'bronchopleural' and 'copd' (Appendix H). For HF, they were: 'intubation', 'chf', 'temp', 'elevation', and 'tachypnea'(Appendix G). Individual inspection of sentence-level SHAP values was conducted to assess how words in individual pre-exam reports influence the prediction (Figure 4).

4 Discussion

4.1 Clinical Indicators

(NMA) The objective of this study was to investigate the degree to which clinical indicators can be derived from pre-exam texts and assess the presence of any discernible patterns when correlated with disease outcomes observed in CXRs. The analysis of word

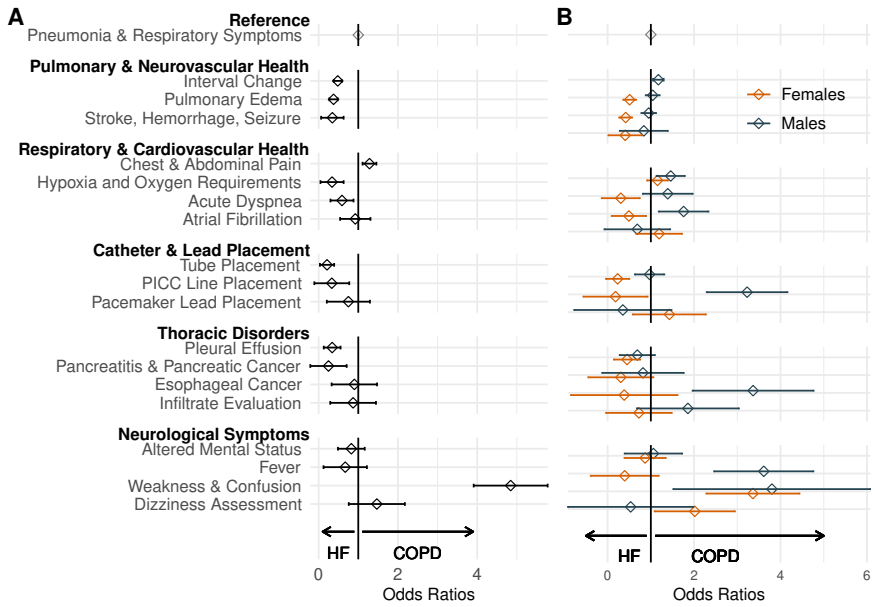

Figure 3: This figure displays the coefficient estimates obtained from two logistic regression models, A) Highlights the coefficient estimates from Eq. 3, while B) Highlights the estimates from Eq. 5. The error bars in both figures represent the Standard Error of the Mean

frequencies revealed terms that differentiate COPD from HF (Figure 1). The most frequent terms for COPD included 'History COPD', 'Productive', 'Assess Pneumonia', and 'Productive Cough'. This highlights the focus on disease history, symptoms, and necessary assessments. In contrast, the most frequent terms for HF were 'Evaluation interval', 'Heart Failure', and 'Intubation'. 'CABG', and 'Heart'. This reflects the distinct clinical characteristics of the two diseases and underlines the value of a further investigation into linguistic clinical indicators between them. Further, it also reveals that many of the pre-exam texts contain information on patient history.

(LHH) Topic modeling identified three predominant topics in the pre-exam narratives: 'Pneumonia & Respiratory Symptoms', 'Interval Change', and 'Pulmonary Edema'. Here 'Pneumonia & Respiratory Symptoms' was the most prevalent topic, followed by 'Interval Change' and 'Pulmonary Edema' The prominence of 'Pneumonia & Respiratory Symptoms' aligns with the fact that the comorbidity between COPD/HF and Pneumonia is high (Restrepo et al., 2018; Shen et al., 2021). The topic "Interval Change" refers to the investigation of any alterations in imaging findings over a given time window (Braileanu et al., 2019), which is necessary for disease progression monitoring and treatment response in both diseases. Literature further shows that 'Pulmonary Edema' is related to both diseases (Braileanu et al., 2019; Khalid et al., 2021).

(NMA) Further exploration revealed the distribution of topics across the two diseases (Figure 2). The three most frequent topics were more evenly distributed for patients diagnosed with HF compared to those with COPD. While the topic 'Pneumonia & Respiratory Symptoms' was the most prominent in both diseases, 'Interval Change' and 'Pulmonary Edema' were more prevalent in HF patients. This

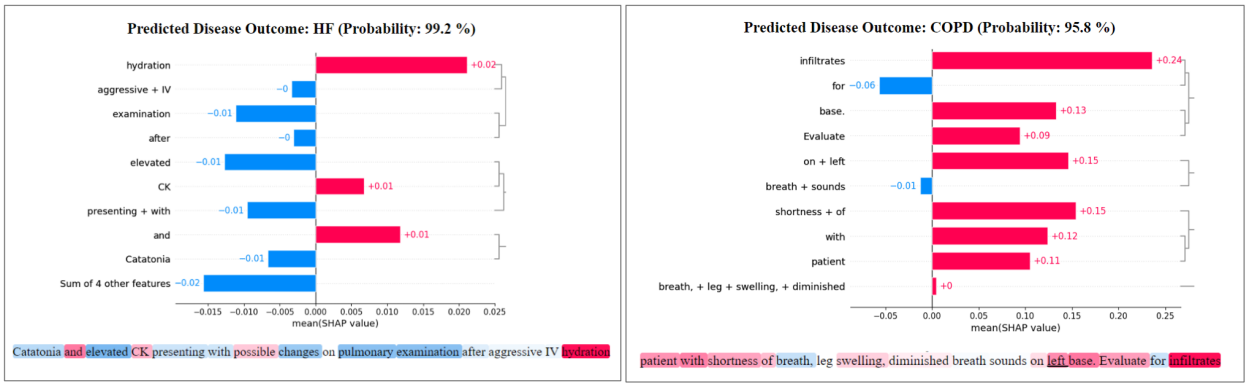

Figure 4: Bar plots illustrating the importance of individual words for each disease outcome in the RadBERT classifier. The left plot depicts the impact of words on the predicted outcome of HF, while the right plot presents the corresponding analysis for COPD. Positive values on the horizontal axis denote a higher degree of importance towards the assigned outcome label. Generated using the SHAP library.

suggests that monitoring changes over time and the presence of pulmonary edema may be more characteristic of HF, while respiratory symptoms are prominent in both diseases. This higher presence of pulmonary edema in heart failure makes sense in the context of literature, stating that Pulmonary edema is often caused by HF (Chen, 2022).

(LHH) Based on our findings, the notion that topics extracted from pre-exam texts might contain linguistic clinical indicators, show promising prospects. Logistic regression provided insights into the statistical explainability of specific topics in distinguishing between COPD and HF (Figure 3). Topics: 'Pulmonary Edema', 'Stroke, Hemorrhage, Seizure', 'Hypoxia and Oxygen Requirements', 'Acute Dyspnea', 'Tube Placement', 'PICC Line Placement', 'Pleural Effusion', and 'Pancreatitis & Pancreatic Cancer', were significantly more prevalent in patients with HF compared to those with COPD.

(NMA) This observation aligns with literature suggesting that these attributes are generally associated with HF. Notably, pulmonary edema is often caused by HF (Chen, 2022). Further, HF increases the risk of stroke, which can lead to seizures (Kim & Kim, 2018; Nei, 2009). Hypoxia, characterized by inadequate oxygen supply, is prevalent in cardiovascular disorders, including HF (Abe et al., 2017). Dyspnea, difficult or labored breathing, is commonly associated with HF (Kupper et al., 2016). The topics 'Tube Placement' and 'PICC Line Placement' are likely linked to HF as these procedures are employed to manage arrhythmias, which can be both symptoms and causes of HF (Cleveland Clinic, 2022). Additionally, while the direct mechanisms remain unclear, pancreatitis has been associated with HF (Dams et al., 2022).

(LHH) Conversely, the topics 'Chest & Abdominal Pain' and 'Weakness & Confusion' had a higher association with COPD compared to HF. Chest pain is a known symptom experienced by patients with COPD (Healthline, 2021). Furthermore, COPD is associated with increased confusion, as well as muscle wasting and resultant weakness (Greenlund et al., 2016; Wüst & Degens, 2007).

Overall, the topics from the identified clusters align with the diagnostic attributes from medical handbooks and literature on the two diseases at hand. This is an indication that the clusters capture se-

mantic information related to COPD and HF. This methodology highlights the promising potential of semantic clustering in extracting clinically relevant information to explain which topics in the pre-exam text lead to a given diagnosis.

4.2 Gender Disparities

(NMA) The model that considered an interaction between gender and topics, revealed a significant gender moderation of the association between topics and COPD/HF. In particular, given that a patient was a male the probability of having COPD was increased when contained in the topics 'Pneumonia & Respiratory Symptoms', 'Chest & Abdominal Pain', 'PICC Line Placement', and 'Fever'. This association is most likely due to the higher prevalence of COPD and its major risk factors such as smoking among male populations (Ntritsos et al., 2018). Another factor that might contribute to the disparity found here is the differences in symptoms experienced and reported by males and females. Males are more likely to present with symptoms such as chronic cough, sputum production, and exertional dyspnea, which are commonly associated with COPD (Zhang et al., 2021).

Another possible explanation for the results is that societal perceptions and expectations surrounding gender roles and health behavior shape the representation and interpretation of symptoms. Traditionally, COPD was primarily associated with elderly males who had a smoking history. However, emerging evidence indicates that the mortality rate of COPD in women is approaching those of males (Ulrik, 2003; Watson et al., 2004). A study by Chapman et al. (2001) supports the notion of implicit bias influencing the diagnosis of COPD.

(LHH) Their findings demonstrated that female patients with a chronic cough and a smoking history were more likely to receive a diagnosis of asthma or a non-respiratory problem, while male patients with identical symptoms were more likely to be diagnosed with COPD. Further, it is thought that because of the traditional view of COPD as a male disease, women are underdiagnosed with the disease (Chapman et al., 2001). This underlines the fact that stereotypical gender norms can influence healthcare providers' understanding and documentation of symptoms. The model with a gender interaction had a lower AIC than the one without an interaction, indicating that the inclusion of the gender interaction term improved the model's fit. This underscores the importance of considering the interaction between gender and topics when examining the association between topics and COPD/HF diagnosis.

Overall, the observed gender disparities in the association between topics and COPD/HF diagnosis can be attributed to multiple factors. These include the higher prevalence of COPD and its risk factors among males, differential symptom patterns experienced by males and females, and implicit biases in healthcare professionals' decision-making. Addressing these disparities requires raising awareness about gender-related factors influencing symptom presentation, and promoting equitable healthcare practices that consider individual differences beyond gender stereotypes.

4.3 Predict or Explain?

(NMA) In the field of radiology, trust in diagnostic tools is crucial and the ability to explain the reasoning behind a prediction is paramount. Classification of disease outcomes with autonomous intelligent systems leads to concerns about legal risks, responsibility, data ownership, consent, and biased predictions (Carter et al., 2020). The General Data Protection Regulation (GDPR) law introduced by the European Union states that it is a right to have a meaningful explanation of the logic involved when using automated decision-making tools (European Union, 2016). Consequently, the concern of black-box algorithms being biased towards specific demographics is concerning (Obermeyer et al., 2019).

(LHH) The embedding-based disease classifier, utilizing pre-trained RadBERT embeddings, demonstrated promising performance in distinguishing between COPD and HF, as evidenced by the Area Under the Curve (AUC) value of 0.79. This indicates that the model effectively learned disease-specific linguistic patterns, enabling an accurate disease classification.

The results from the SHAP-based feature importance analysis are promising in terms of explaining why an individual sentence falls into their given disease category, as we can directly calculate how the masking of words in pre-exam reports affects the predicted outcome probabilities of the model. From Figure 4, it becomes apparent that different words within the same sentence will affect the predicted outcome in opposite directions. It is interesting that in the phrase 'shortness of breath', the word 'breath' is more important towards HF, while the 'shortness of' adds to the certainty of the COPD outcome. Firstly, this highlights that the decisions of the model do not reflect the way any medical professional would approach a pre-exam report. Second, though model outcomes are explainable, these masking-based explanations do not necessarily relate to the overall semantic context of the sentence, as a topic label would.

(NMA) Additionally, the words calculated to have the highest effect on an outcome, on average, across the entire test dataset, reveal certain linguistic indicators associated with each condition (Appendix G/H). It is no surprise that the word 'copd' is among the most salient words for the outcome of COPD, and 'chf' for (congestive) heart failure, as some patients in the dataset have previous diagnoses listed in their report. The other terms, deemed most salient by the model, reveal other linguistic indicators for the disease outcomes, however, to assess if these are substantial findings it would require the expertise of radiologists. It is also important to note that solely focusing on two diseases may result in a misinterpretation of our disease classifier, as the evaluation of the model assumes that a patient will certainly fall into one of the two disease categories. Therefore, it would be valuable to determine if a model can equally distinguish between all nine diseases in the original dataset, to assess its validity, when linguistic features can fall into other categories than COPD and HF. This would result in a more nuanced assessment of the linguistic features that lead to certain diagnoses.

(LHH) The comparison made between the efficacy of a topic versus a transformer-based approach underpins the concerns regarding explainability, as we not only try to predict disease outcomes but instead explain the underlying reasons and patterns that contribute to these predictions. This enables healthcare professionals to have a clear understanding of the factors influencing diagnostic outcomes. The use of topic modeling as a clinical decision support tool offers several advantages over black-box algorithms. First, it provides interpretability by identifying the specific topics or themes present in the pre-exam texts that are associated with certain diseases. Also, it sheds light on the potential disparities that might exist in disease classification. The transparency provided by a topic-based model allows healthcare professionals to comprehend the basis of the model's predictions and validate its reasoning.

(NMA) A potential use case of the methodology would be to utilize it as an automated screening and alert system. By integrating such, in collaboration with healthcare professionals, we can flag high-risk cases, highlight critical findings, and assist in prioritizing the review of CXRs based on the information reported prior to a CXR. Our approach aligns with the static nature of the pre-exam reports, that precede the conduction and evaluation of a CXR. This would ultimately save time and aid in prioritizing patients based on the urgency and severity of their condition.

4.4 Limitations & Future Work

(LHH) One major limitation of this study is the short length of the pre-exam texts in the dataset. To create semantically meaningful clusters, it would be beneficial to have more extensive reports for each patient. This would likely enable the identification of larger and more distinguishable topics with clear disease-related meanings, as more semantic information would be available for each patient. Although outlier reduction was implemented to address the issue of sentences not being successfully placed in a cluster, we speculate that more detailed reports would help address this confounder. Therefore, future research should strive for the incorporation of a patient's full journal history to overcome this limitation.

(NMA) The topic naming process relied on using only the top 5 words from each topic, and individuals without medical expertise in lung disease diagnosis performed the naming. To enhance this process, it would be beneficial to collaborate with doctors experienced in diagnosing lung diseases. This collaboration would enable the utilization of domain knowledge from doctors and the expertise of data scientists in NLP and topic modeling, resulting in more informed analysis. Importantly, it is essential to acknowledge that the findings presented in this study are of a correlational nature, and caution should be exercised in inferring causal relationships.

(LHH) Several factors warrant careful consideration when interpreting the coefficient results. First, there exist numerous confounding variables that may influence the two disease outcomes, apart from the identified topics and their association with gender. These unaccounted variables can potentially introduce bias and confound the observed associations. Furthermore, an issue of reverse causality should be acknowledged. It is apparent in the findings that patients who have already been diagnosed with COPD or HF are more likely to undergo CXRs to monitor changes in their respective diseases.

A limitation in the dataset is the lack of clear distinguishability between texts that represent the reason for undergoing a CXR and those that contain a patient's diagnostic history. Many reports treated as 'reasons for the exam' also contain previous diagnoses, such as "with COPD, pulmonary". By including texts that explain COPD but clearly state that the patient already has the condition, it becomes uncertain whether the topic contains a clinical indicator based on the doctors' use of words or merely represents a group of texts where the patient has already been diagnosed. Addressing this limitation would require a dataset with better differentiation between 'reason for exam texts' and patients' diagnostic histories.

5 Conclusion

(NMA & LHH) This study demonstrated the potential of employing text mining techniques to discern and elucidate linguistic patterns present in pre-exam texts. The application of topic modeling not only holds promise for enhancing the diagnostic process pertaining to COPD and HF but also facilitates a deeper comprehension of the distribution of specific semantic patterns within these texts in relation to disease outcomes. Our findings underlined the distinct clinical characteristics of COPD and HF and validated the association between patient history and radiological findings in CXRs. The observed gender disparities in the associations between topic clusters and disease outcomes emphasized the need for considering individual demographic information in the diagnosis and treatment of COPD and HF. This advocates for a more personalized and inclusive approach in healthcare delivery. While the application of a state-of-the-art language model like RadBERT demonstrated promising disease prediction capabilities, it also underscored the trade-off between prediction and explanation prevalent in machine learning applications. This comparison high-

lighted the challenges of employing black-box algorithms in healthcare and emphasized the importance of interpretability and trust in clinical decision-support tools. Ultimately, our methodology provides the groundwork for future topic-based decision-support tools in healthcare. A potential application of this methodology could be an automated screening and alert system, enhancing patient prioritization, and ultimately improving overall healthcare delivery. In conclusion, our research paves the way for further investigation into harnessing the power of patient history and linguistic patterns in the diagnosis and treatment of diseases.

References

Abe, H., Semba, H., & Takeda, N. (n.d.-a). The roles of hypoxia signaling in the pathogenesis of cardiovascular diseases. , *24*(9), 884–894. Retrieved 2023-05-25, from https://www.jstage.jst.go.jp/article/jat/24/9/24_RV17009/_article doi: 10.5551/jat.RV17009

Abe, H., Semba, H., & Takeda, N. (n.d.-b). The roles of hypoxia signaling in the pathogenesis of cardiovascular diseases. , *24*(9), 884–894. Retrieved 2023-05-25, from https://www.jstage.jst.go.jp/article/jat/24/9/24_RV17009/_article doi: 10.5551/jat.RV17009

Agu, N. N., Wu, J. T., Chao, H., Lourentzou, I., Sharma, A., Moradi, M., ... Hendler, J. (n.d.). AnaXNet: anatomy aware multi-label finding classification in chest x-ray. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, strasbourg, france, september 27–october 1, 2021, proceedings, part v 24* (pp. 804–813). Springer.

Ait Nasser, A., & Akhloufi, M. A. (n.d.). A review of recent advances in deep learning models for chest disease detection using radiography. , *13*(1), 159. Retrieved 2023-05-26, from https://www.mdpi.com/2075-4418/13/1/159 doi: 10.3390/diagnostics13010159

Athiwaratkun, B., Wilson, A. G., & Anandkumar, A. (n.d.). Probabilistic fasttext for multi-sense word embeddings.

Braileanu, M., Crawford, K., Key, S., & Mullins, M. (n.d.). Assessment of explicitly stated interval change on noncontrast head CT radiology reports. , *40*(7), 1091–1094. Retrieved 2023-05-26, from http://www.ajnr.org/lookup/doi/10.3174/ajnr.A6081 doi: 10.3174/ajnr.A6081

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (n.d.). Language models are few-shot learners. Retrieved 2023-05-27, from https://arxiv.org/abs/2005.14165 (Publisher: arXiv Version Number: 4) doi: 10.48550/ARXIV.2005.14165

Chambon, P., Cook, T. S., & Langlotz, C. P. (n.d.). Improved fine-tuning of in-domain transformer model for inferring COVID-19 presence in multi-institutional radiology reports. doi: 10.1007/s10278-022-00714-8

Chapman, K. R., Tashkin, D. P., & Pye, D. J. (n.d.-a). Gender bias in the diagnosis of COPD. , *119*(6), 1691–1695. Retrieved 2023-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0012369215523143 doi: 10.1378/chest.119.6.1691

Chapman, K. R., Tashkin, D. P., & Pye, D. J. (n.d.-b). Gender bias in the diagnosis of COPD. , *119*(6), 1691–1695. (Publisher: Elsevier)

Chen, M. A. (n.d.). *Pulmonary edema - symptoms and causes.* Retrieved 2023-05-25, from https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to

-z/pulmonary-edema

Church, K. W. (n.d.). Word2vec. , *23*(1), 155–162. (Publisher: Cambridge University Press)

Clinic, C. (n.d.). *Heart failure surgery: Options, outlook & risks.* Retrieved 2023-05-25, from https://my.clevelandclinic.org/health/treatments/12905-heart-failure-surgery

Dams, O. C., Vijver, M. A. T., Van Veldhuisen, C. L., Verdonk, R. C., Besselink, M. G., & Van Veldhuisen, D. J. (n.d.). Heart failure and pancreas exocrine insufficiency: Pathophysiological mechanisms and clinical point of view. , *11*(14), 4128. Retrieved 2023-05-25, from https://www.mdpi.com/2077-0383/11/14/4128 doi: 10.3390/jcm11144128

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (n.d.). Big data in healthcare: management, analysis and future prospects. , *6*(1), 54. Retrieved 2023-05-26, from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0217-0 doi: 10.1186/s40537-019-0217-0

Denti, T. (n.d.). Topic modeling of prostate cancer radiology reports.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (n.d.). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved 2023-05-27, from https://arxiv.org/abs/1810.04805 (Publisher: arXiv Version Number: 2) doi: 10.48550/ARXIV.1810.04805

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... Stanley, H. E. (n.d.). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. , *101*(23). Retrieved 2023-05-26, from https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215 doi: 10.1161/01.CIR.101.23.e215

Greenlund, K. J., Liu, Y., Deokar, A. J., Wheaton, A. G., & Croft, J. B. (n.d.). Association of chronic obstructive pulmonary disease with increased confusion or memory loss and functional limitations among adults in 21 states, 2011 behavioral risk factor surveillance system. , *13*, 150428. Retrieved 2023-05-25, from http://www.cdc.gov/pcd/issues/2016/15_0428.htm doi: 10.5888/pcd13.150428

Hassanpour, S., & Langlotz, C. P. (n.d.). Unsupervised topic modeling in a large free text radiology report repository. , *29*, 59–62. (ISBN: 0897-1889 Publisher: Springer)

Healthline. (n.d.). *COPD and chest pain: The link and tips to manage.* Retrieved 2023-05-25, from https://www.healthline.com/health/copd/copd-and-chest-pain

Honnibal, M., & Montani, I. (n.d.). *spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.*

Kessler, J. S. (n.d.). Scattertext: a browser-based tool for visualizing how corpora differ. (Place: Vancouver, Canada Publisher: Association for Computational Linguistics)

Khalid, K., Padda, J., Komissarov, A., Colaco, L. B., Padda, S., Khan, A. S., ... Jean-Charles, G. (n.d.). The coexistence of chronic obstructive pulmonary disease and heart failure. Retrieved 2023-05-26, from https://www.cureus.com/articles/65345-the-coexistence-of-chronic-obstructive-pulmonary-disease-and-heart-failure doi: 10.7759/cureus.17387

Kim, W., & Kim, E. J. (n.d.). Heart failure as a risk factor for stroke. , *20*(1), 33–45. Retrieved 2023-05-25, from http://j-stroke.org/journal/view.php?doi=10.5853/jos.2017.02810 doi: 10.5853/jos.2017.02810

Kupper, N., Bonhof, C., Westerhuis, B., Widdershoven, J., & Denollet, J. (n.d.). Determinants of dyspnea in chronic heart failure. , *22*(3), 201–209. Retrieved 2023-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S1071916415011227 doi: 10.1016/j.cardfail.2015.09.016

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (n.d.). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Retrieved 2023-05-25, from https://arxiv.org/abs/1901.08746 (Publisher: arXiv Version Number: 4) doi: 10.48550/ARXIV.1901.08746

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (n.d.). *Hybrid retrieval-generation reinforced agent for medical image report generation* (No. arXiv:1805.08298). arXiv. Retrieved 2023-05-26, from http://arxiv.org/abs/1805.08298

Liu, Q., Kusner, M. J., & Blunsom, P. (n.d.). A survey on contextual embeddings.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (n.d.). RoBERTa: A robustly optimized BERT pretraining approach. , *abs/1907.11692*. Retrieved from http://arxiv.org/abs/1907.11692 (_eprint: 1907.11692)

Lundberg, S. M., & Lee, S.-I. (n.d.-a). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Lundberg, S. M., & Lee, S.-I. (n.d.-b). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

McInerney, D. J., Young, G., van de Meent, J.-W., & Wallace, B. C. (n.d.). *That's the wrong lung! evaluating and improving the interpretability of unsupervised multimodal encoders for medical data* (No. arXiv:2210.06565). arXiv. Retrieved 2023-05-26, from http://arxiv.org/abs/2210.06565

McInnes, L., Healy, J., & Astels, S. (n.d.). hdbscan: Hierarchical density based clustering. , *2*(11), 205.

McInnes, L., Healy, J., & Melville, J. (n.d.). Umap: Uniform manifold approximation and projection for dimension reduction.

Mozayan, A., Fabbri, A. R., Maneevese, M., Tocino, I., & Chheang, S. (n.d.). Practical guide to natural language processing for radiology. , *41*(5), 1446–1453. Retrieved 2023-05-26, from http://pubs.rsna.org/doi/10.1148/rg.2021200113 doi: 10.1148/rg.2021200113

Nei, M. (n.d.). Cardiac effects of seizures. , *9*(4), 91–95. Retrieved 2023-05-25, from http://journals.sagepub.com/doi/10.1111/j.1535-7511.2009.01303.x doi: 10.1111/j.1535-7511.2009.01303.x

Ntritsos, G., Franek, J., Belbasis, L., Christou, M. A., Markozannes, G., Altman, P., ... Evangelou, E. (n.d.). Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis. , *Volume 13*, 1507–1514. Retrieved 2023-05-25, from https://www.dovepress.com/gender-specific-estimates-of-copd-prevalence-a-systematic-review-and-m-peer-reviewed-article-COPD doi: 10.2147/COPD.S146390

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... Lerer, A. (n.d.). Automatic differentiation in PyTorch. In *NIPS-w.*

Pennington, J., Socher, R., & Manning, C. D. (n.d.). Glove: Global vectors for word representation. In (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (n.d.). Deep contextualized word representations. Retrieved 2023-05-27, from https://arxiv.org/abs/1802.05365 (Publisher: arXiv Version Number: 2) doi: 10.48550/ARXIV.1802.05365

Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (n.d.). Natural language processing in radiology: a systematic review. , *279*(2), 329–343. (ISBN: 0033-8419 Publisher: Radiological Society of North America)

Price, W. N. (n.d.). Big data and black-box medical algorithms. , *10*(471), eaao5333. Retrieved 2023-05-25, from https://www.science.org/doi/10.1126/scitranslmed.aao5333 doi: 10.1126/scitranslmed.aao5333

R Core Team. (n.d.). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rajapakse, T. C. (n.d.). *Simple transformers.* Retrieved from https://github.com/ThilinaRajapakse/simpletransformers

Reimers, N. (n.d.). *Pretrained models — sentence-transformers documentation.* Retrieved 2023-05-27, from https://www.sbert.net/docs/pretrained_models.html

Reimers, N., & Gurevych, I. (n.d.-a). Making monolingual sentence embeddings multilingual using knowledge distillation.

Reimers, N., & Gurevych, I. (n.d.-b). Sentence-bert: Sentence embeddings using siamese bert-networks.

Restrepo, M. I., Sibila, O., & Anzueto, A. (n.d.). Pneumonia in patients with chronic obstructive pulmonary disease. , *81*(3), 187. Retrieved 2023-05-26, from http://e-trd.org/journal/view.php?doi=10.4046/trd.2018.0030 doi: 10.4046/trd.2018.0030

Samek, W., Wiegand, T., & Müller, K.-R. (n.d.). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models* (No. arXiv:1708.08296). arXiv. Retrieved 2023-05-26, from http://arxiv.org/abs/1708.08296

Shen, L., Jhund, P. S., Anand, I. S., Bhatt, A. S., Desai, A. S., Maggioni, A. P., ... McMurray, J. J. (n.d.). Incidence and outcomes of pneumonia in patients with heart failure. , *77*(16), 1961–1973. Retrieved 2023-05-26, from https://linkinghub.elsevier.com/retrieve/pii/S0735109721005775 doi: 10.1016/j.jacc.2021.03.001

Singh, A., Sengupta, S., & Lakshminarayanan, V. (n.d.). Explainable deep learning models in medical image analysis. , *6*(6), 52. Retrieved 2023-05-26, from https://www.mdpi.com/2313-433X/6/6/52 doi: 10.3390/jimaging6060052

Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., & Lungren, M. P. (n.d.). *CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT* (No. arXiv:2004.09167). arXiv. Retrieved 2023-05-26, from http://arxiv.org/abs/2004.09167

Team, P. C. (n.d.). *Python: A dynamic, open source programming language.* Python Software Foundation.

Theander, K., Unosson, M., Karlsson, I., Eckerblad, J., Luhr, K., & Hasselgren, M. (n.d.). Symptoms and impact of symptoms on function

and health in patients with chronic obstructive pulmonary disease and chronic heart failure in primary health care. , 785. Retrieved 2023-05-23, from http://www.dovepress.com/symptoms-and-impact-of-symptoms-on-function-and-health-in-patients-wit-peer-reviewed-article-COPD doi: 10.2147/COPD.S62563

Ulrik, C. (n.d.). Smoking and mortality in women:" smoke like a man, die (at least) like a man". , *8*, 103–117. (Publisher: European Respiratory Society)

Union, E. (n.d.). Regulation (EU) 2016/679 of the european parliament and of the council. , *679*, 2016.

Washko, G. (n.d.). Diagnostic imaging in COPD. , *31*(3), 276–285. Retrieved 2023-05-26, from http://www.thieme-connect.de/DOI/DOI?10.1055/s-0030-1254068 doi: 10.1055/s-0030-1254068

Watson, L., Vestbo, J., Postma, D. S., Decramer, M., Rennard, S., Kiri, V. A., ... Soriano, J. B. (n.d.). Gender differences in the management and experience of chronic obstructive pulmonary disease. , *98*(12), 1207–1213. Retrieved 2023-05-26, from https://linkinghub.elsevier.com/retrieve/pii/S0954611104001866 doi: 10.1016/j.rmed.2004.05.004

WHO. (n.d.). World health organization: Chronic obstructive pulmonary disease (COPD). Retrieved 2023-05-26, from https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

Wickham, H. (n.d.). *stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (n.d.). Welcome to the tidyverse. , *4*(43), 1686. doi: 10.21105/joss.01686

Wu, J. T., Agu, N. N., Lourentzou, I., Sharma, A., Paguio, J. A., Yao, J. S., ... Moradi, M. (n.d.). *Chest ImaGenome dataset for clinical reasoning* (No. arXiv:2108.00316). arXiv. Retrieved 2023-05-23, from http://arxiv.org/abs/2108.00316

Wu, J. T., Syed, A., Ahmad, H., Pillai, A., Gur, Y., Jadhav, A., ... Syeda-Mahmood, T. (n.d.). AI accelerated human-in-the-loop structuring of radiology reports. In *AMIA annual symposium proceedings* (Vol. 2020, p. 1305). American Medical Informatics Association.

Wüst, R. C. I., & Degens, H. (n.d.). Factors contributing to muscle wasting and dysfunction in COPD patients. , *2*(3), 289–300.

Yan, A., McAuley, J., Lu, X., Du, J., Chang, E. Y., Gentili, A., & Hsu, C.-N. (n.d.). RadBERT: Adapting transformer-based language models to radiology. , *4*(4), e210258. (Publisher: Radiological Society of North America)

Zech, J., Pain, M., Titano, J., Badgeley, M., Schefflein, J., Su, A., ... Oermann, E. K. (n.d.). Natural language–based machine learning models for the annotation of clinical radiology reports. , *287*(2), 570–580. (ISBN: 0033-8419 Publisher: Radiological Society of North America)

Zhang, H., Wu, F., Yi, H., Xu, D., Jiang, N., Li, Y., ... Wang, K. (n.d.). Gender differences in chronic obstructive pulmonary disease symptom clusters. , *Volume 16*, 1101–1107. Retrieved 2023-05-25, from https://www.dovepress.com/gender-differences-in-chronic-obstructive-pulmonary-disease-symptom-cl-peer-reviewed-article-COPD doi: 10.2147/COPD.S302877

Zhou, B., Yang, G., Shi, Z., & Ma, S. (n.d.). Natural language processing for smart healthcare. (ISBN: 1937-3333 Publisher: IEEE)

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y.,

Zhu, H., ... He, Q. (n.d.). A comprehensive survey on transfer learning. Retrieved 2023-05-27, from https://arxiv.org/abs/1911.02685 (Publisher: arXiv Version Number: 3) doi: 10.48550/ARXIV.1911.02685

Çallı, E., Sogancioglu, E., Van Ginneken, B., Van Leeuwen, K. G., & Murphy, K. (n.d.). Deep learning for chest x-ray analysis: A survey. , *72*, 102125. Retrieved 2023-05-26, from https://linkinghub.elsevier.com/retrieve/pii/S1361841521001717 doi: 10.1016/j.media.2021.102125

A Appendix

Level 2 Name	Top Words	Description
Pneumonia & Respiratory Symptoms	Pneumonia, Of, Cough, Breath	Pneumonia and various respiratory symptoms.
Interval Change	Interval, Change, Pneumothorax, Sp	Changes observed within specific intervals.
Pulmonary Edema	Pulmonary, Edema, deWith, Pna	The presence and assessment of pulmonary edema.
Chest & Abdominal Pain	Pain, Chest, With, Abdominal	Pain experienced in the chest and abdominal areas.
Tube Placement	Placement, Tube, ET, ETT	The placement of various tubes, such as endotracheal tubes (ETT).
Pleural Effusion	Pleural, Effusion, Effusions, For	Presence and evaluation of pleural effusion.
Acute Dyspnea	Dyspnea, Exertion, Acute, On	Sudden or acute dyspnea, characterized by difficulty in breathing or shortness of breath during exertion.
Stroke, Hemorrhage, Seizure	Stroke, Hemorrhage, Seizure, Subarachnoid	Stroke, hemorrhage, and seizure events, their evaluation, and associated subarachnoid symptoms.
Altered Mental Status	Mental, Altered, Status, Delirium	Changes in mental status, including altered cognition, delirium, or confusion.
Hypoxia and Oxygen Requirements	Hypoxia, Oxygen, Process, And	Assessment and management of hypoxia, as well as oxygen requirements in patients with respiratory or thoracic disorders.
Pancreatitis & Pancreatic Cancer	Pancreatitis, Pancreatic, Cancer, Volume	Pancreatitis and pancreatic cancer, including their evaluation, volume considerations, and associated symptoms.
PICC Line Placement	PICC, Line, Placement, IJ	This category focuses on topics related to the placement of peripherally inserted central catheter (PICC) lines and associated considerations.
Atrial Fibrillation	Atrial, Fibrillation, Palpitations, Syncope	This category involves topics related to the evaluation and management of atrial fibrillation, a common cardiac arrhythmia.
Pacemaker Lead Placement	Lead, Pacemaker, Leads, PP	This category includes topics related to the placement of pacemaker leads and associated evaluation methods.
Fever	Ever, Pneumonia, For, And	This category focuses on topics related to fever and its assessment as a symptom or manifestation of underlying conditions.
Esophageal Cancer	Esophageal, Esophagectomy, Interval, Sp	This category encompasses topics related to esophageal cancer, including evaluation, intervals, and associated surgical interventions.
Weakness & Confusion	Weakness, Confusion, Generalized, With	This category involves topics related to weakness, confusion, and generalized manifestations often associated with respiratory or thoracic disorders.
Infiltrate Evaluation	Infiltrate, Evaluation, WRising, As	This category includes topics related to the evaluation of infiltrates observed in imaging studies and associated characteristics.

Figure 5: A table with Topic Level 2 names, their top words, and a description of the topics.

B Appendix

Level 1 Name	Description
Thoracic Disorders	Topics related to disorders or conditions specifically affecting the thoracic region.
Respiratory & Cardiovascular Health	Topics related to respiratory and cardiovascular health, including symptoms, evaluations, and conditions affecting these systems.
Pulmonary & Neurovascular Health	Topics related to pulmonary health, as well as neurovascular symptoms and conditions that may be associated with respiratory issues.
Catheter & Lead Placement	Topics related to respiratory and cardiovascular health, including symptoms, evaluations, and conditions affecting these systems.
Neurolical Symptoms	Topics related to respiratory and cardiovascular health, including symptoms, evaluations, and conditions affecting these systems.

Figure 6: A table with Topic Level 1 names, and a description of the topics.

C Appendix

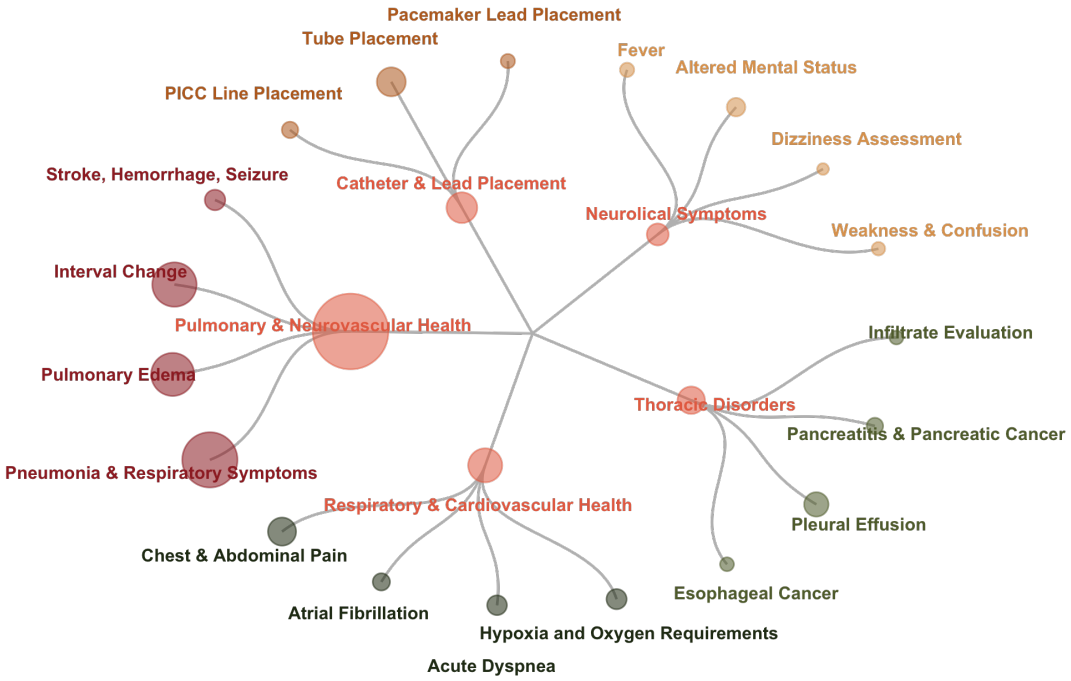

Figure 7: Dendogram of the topic hierarchy.

D Appendix

Python v. 3.11 (Python Core Team, 2021) was used for data preprocessing, topic modelling, and classification. BERTopic v. 0.11.0 was used for topic modelling (Grootendorst, 2022). Numpy v. 1.23.5 was used for mathematical transformation. Pandas v. 1.5.2, was used for data handling. Fastjsonschema v. 2.16.3 was used to read the data contained in JSON files. Sentence transformers v. 2.2.2 was used to generate s-bert embeddings. Further Rstudio (R Core Team, 2022) with R v. 4.1.3 was used for data visualizations and regression models. Here Ggplot2 was used for visualizations, tidyverse for data-wrangling, and lme4 for regression models (Bates et al., 2015; Wickham, 2016; Wickham et al., 2019). Code for reproducing the analysis can be found here: https://github.com/Lassehhansen/CXR-Clinical-Indicators

E Appendix

	Model Controls			Model Gender Interaction		
Variable	**OR**[1]	**95% CI**[1]	**p**	**OR**[1]	**95% CI**[1]	**p**
Gender						
F	—	—		—	—	
M	1.11	1.03, 1.20	0.005	1.20	1.06, 1.36	0.004
Age Decile						
>=90	—	—		—	—	
0-10	0.00		>0.9	0.00		>0.9
20-30	0.25	0.16, 0.38	<0.001	0.25	0.16, 0.38	<0.001
20-Oct	0.11	0.02, 0.32	<0.001	0.11	0.02, 0.32	<0.001
30-40	0.55	0.40, 0.75	<0.001	0.54	0.40, 0.74	<0.001
40-50	0.89	0.67, 1.19	0.4	0.89	0.66, 1.18	0.4
50-60	1.06	0.80, 1.40	0.7	1.05	0.80, 1.40	0.7
60-70	1.19	0.91, 1.58	0.2	1.19	0.90, 1.57	0.2
70-80	1.16	0.88, 1.53	0.3	1.16	0.87, 1.53	0.3
80-90	1.24	0.93, 1.66	0.14	1.24	0.93, 1.65	0.15
Topic						
Pneumonia & Respiratory Symptoms	—	—		—	—	
Interval Change	0.49	0.44, 0.54	<0.001	0.51	0.44, 0.59	<0.001
Pulmonary Edema	0.37	0.34, 0.42	<0.001	0.42	0.36, 0.49	<0.001
Stroke, Hemorrhage, Seizure	0.34	0.25, 0.45	<0.001	0.40	0.27, 0.59	<0.001
Chest & Abdominal Pain	1.29	1.09, 1.55	0.004	1.16	0.91, 1.48	0.2
Hypoxia and Oxygen Requirements	0.33	0.24, 0.44	<0.001	0.30	0.19, 0.46	<0.001
Acute Dyspnea	0.58	0.43, 0.77	<0.001	0.48	0.32, 0.72	<0.001
Atrial Fibrillation	0.89	0.61, 1.31	0.5	1.14	0.68, 1.98	0.6
Tube Placement	0.21	0.18, 0.26	<0.001	0.24	0.18, 0.31	<0.001
PICC Line Placement	0.32	0.21, 0.50	<0.001	0.18	0.08, 0.36	<0.001
Pacemaker Lead Placement	0.67	0.39, 1.17	0.2	1.30	0.58, 3.20	0.5
Pleural Effusion	0.33	0.27, 0.41	<0.001	0.44	0.32, 0.60	<0.001
Pancreatitis & Pancreatic Cancer	0.26	0.16, 0.41	<0.001	0.32	0.14, 0.67	0.003
Esophageal Cancer	0.90	0.51, 1.64	0.7	0.36	0.09, 1.21	0.11
Infiltrate Evaluation	0.87	0.49, 1.57	0.6	0.72	0.33, 1.58	0.4
Altered Mental Status	0.79	0.56, 1.12	0.2	0.81	0.50, 1.31	0.4
Fever	0.65	0.37, 1.13	0.12	0.38	0.16, 0.81	0.015
Weakness & Confusion	4.48	1.93, 13.0	0.002	2.99	1.13, 10.3	0.046
Dizziness Assessment	1.39	0.70, 2.94	0.4	1.91	0.80, 5.30	0.2
Gender * Topic						
M * Interval Change				0.91	0.74, 1.12	0.4
M * Pulmonary Edema				0.79	0.64, 0.97	0.027
M * Stroke, Hemorrhage, Seizure				0.71	0.40, 1.26	0.2
M * Chest & Abdominal Pain				1.26	0.88, 1.79	0.2
M * Hypoxia and Oxygen Requirements				1.20	0.67, 2.19	0.5
M * Acute Dyspnea				1.49	0.82, 2.70	0.2
M * Atrial Fibrillation				0.58	0.27, 1.25	0.2
M * Tube Placement				0.83	0.58, 1.20	0.3
M * PICC Line Placement				2.82	1.12, 7.53	0.032
M * Pacemaker Lead Placement				0.29	0.09, 0.89	0.034
M * Pleural Effusion				0.58	0.38, 0.89	0.012
M * Pancreatitis & Pancreatic Cancer				0.72	0.28, 1.93	0.5
M * Esophageal Cancer				3.23	0.81, 14.5	0.10
M * Infiltrate Evaluation				1.54	0.48, 5.27	0.5
M * Altered Mental Status				0.96	0.49, 1.90	>0.9
M * Fever				3.19	1.02, 10.7	0.051
M * Weakness & Confusion				3.50	0.46, 72.0	0.3
M * Dizziness Assessment				0.44	0.10, 1.91	0.3
AIC	15,884			15,880		
Deviance	15,826			15,786		

[1] OR = Odds Ratio, CI = Confidence Interval

Figure 8: Logistic Regression Model tables.

F Appendix

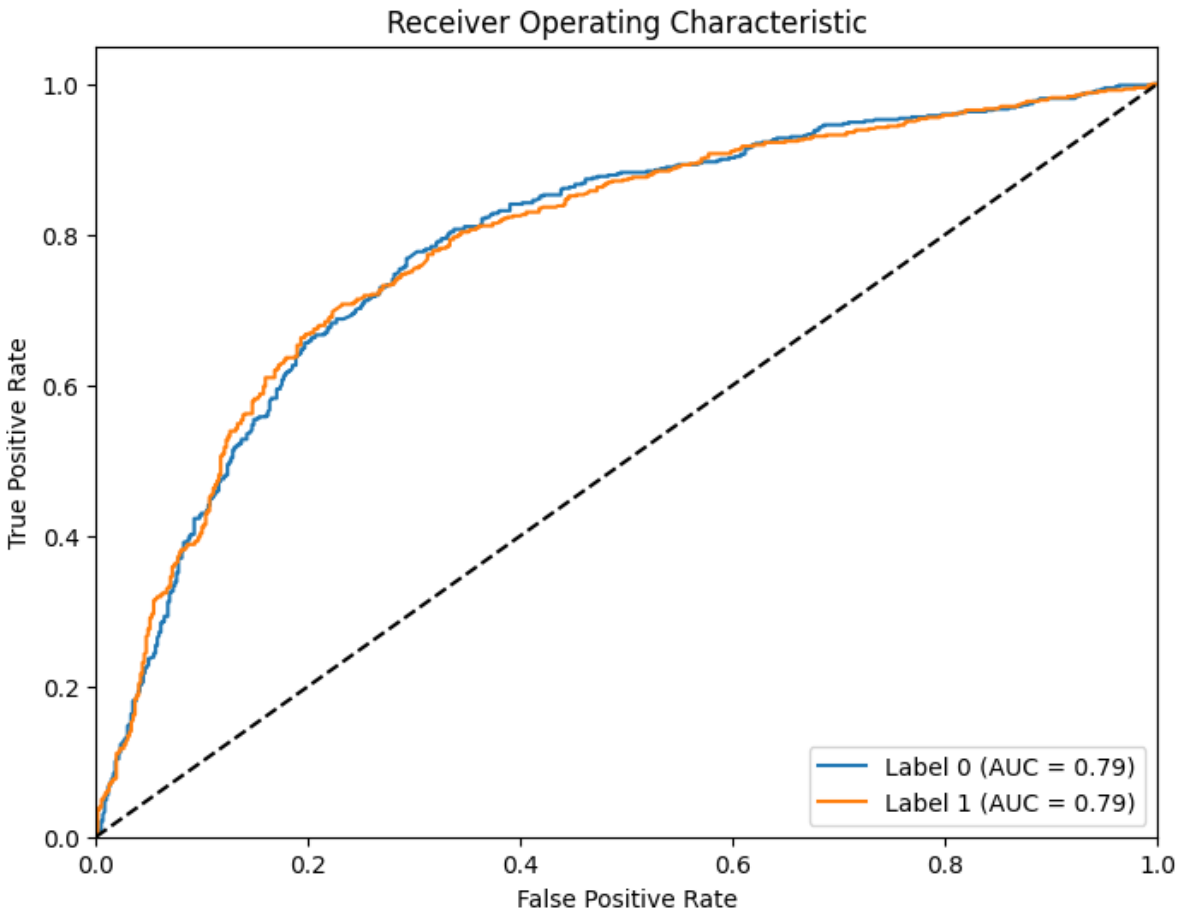

Figure 9: ROC Curve for RadBERT classifier.

G Appendix

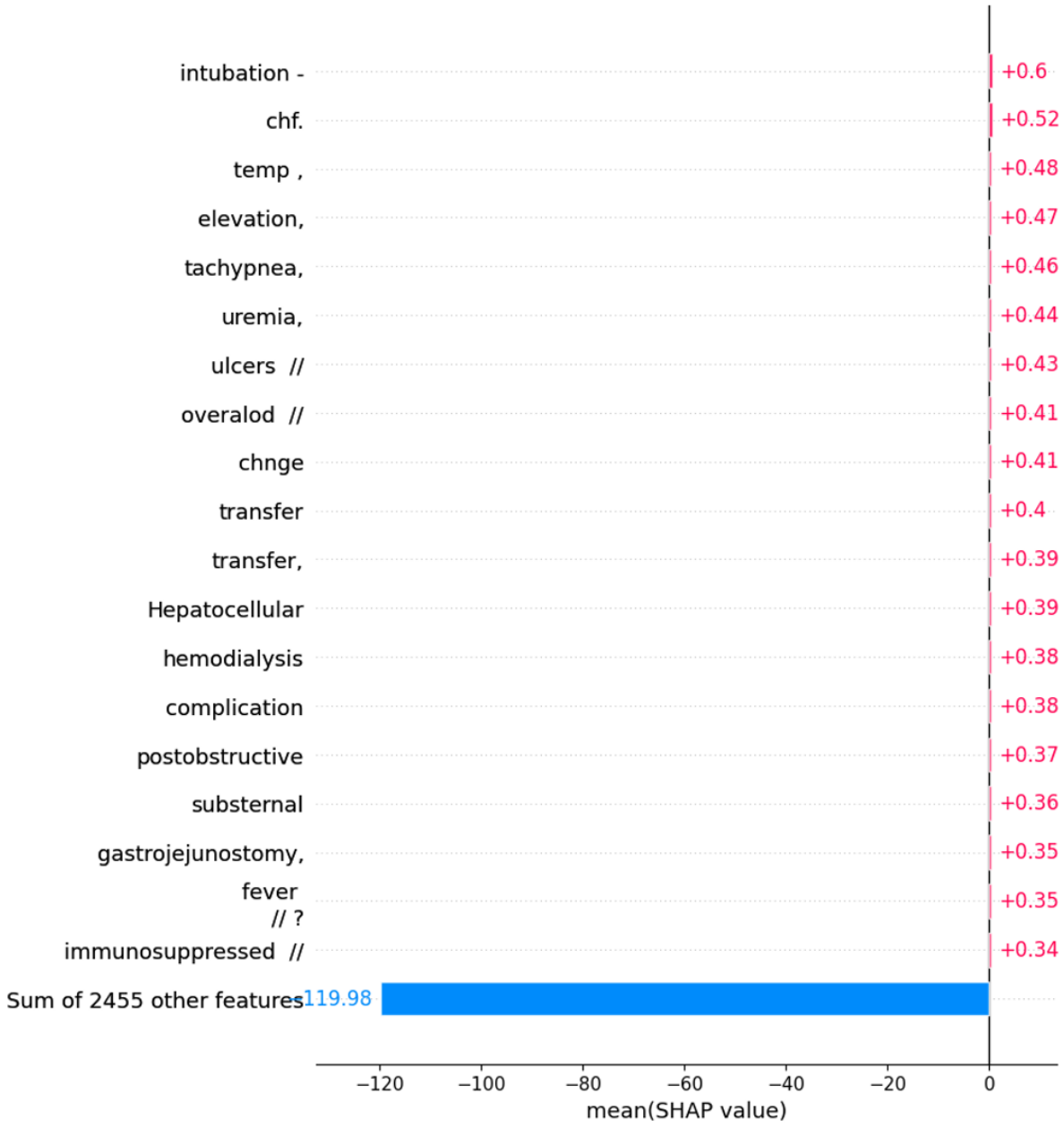

Figure 10: Top 20 words with highest average influence on disease outcome HF for the RadBERT classifier.

H Appendix

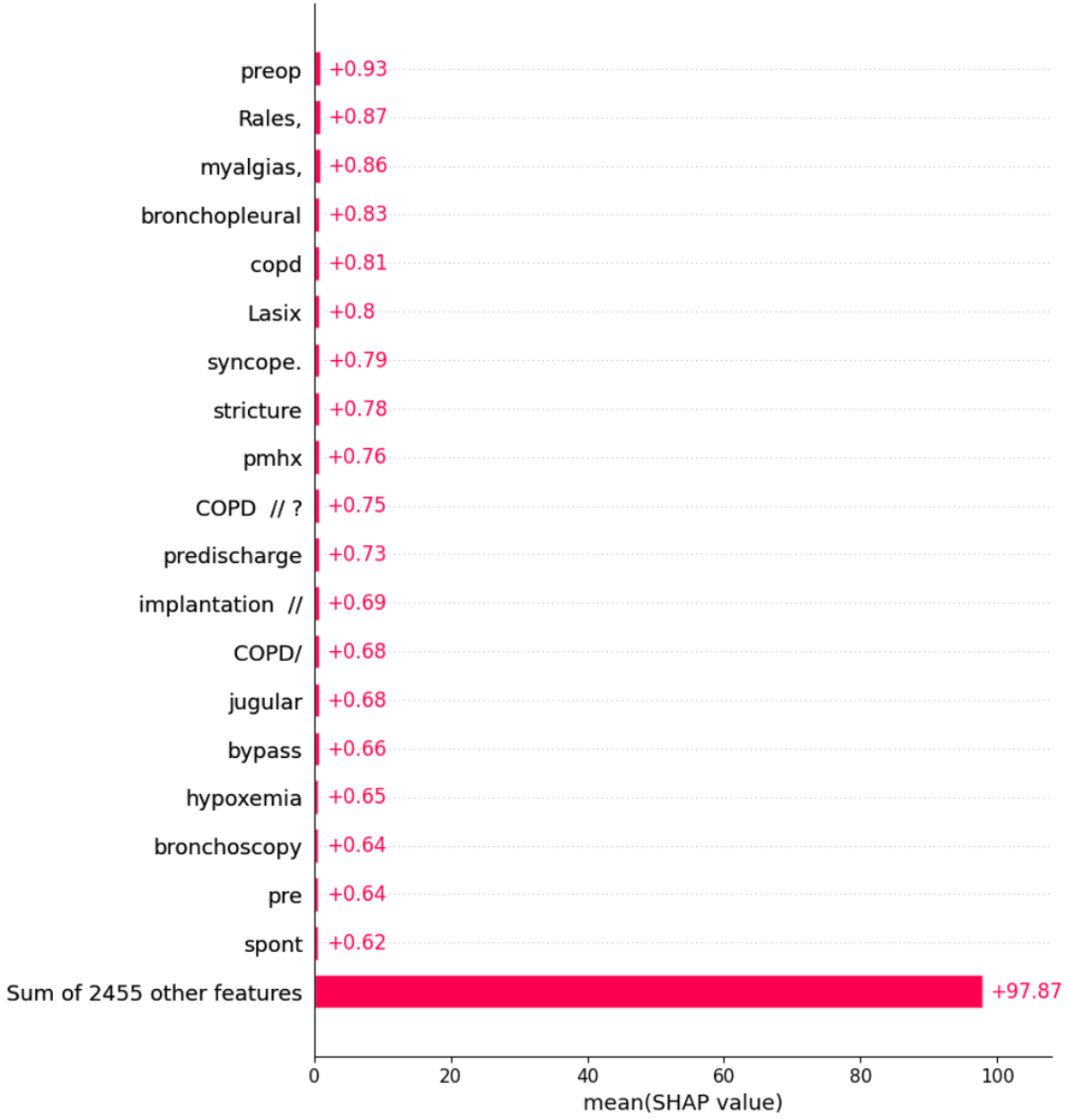

Figure 11: Top 20 words with highest average influence on disease outcome COPD for the RadBERT classifier.