Xplainer: From X-Ray Observations to Explainable Zero-Shot Diagnosis

Chantal Pellegrini[1(✉)], Matthias Keicher[1], Ege Özsoy[1], Petra Jiraskova[2], Rickmer Braren[2], and Nassir Navab[1]

[1] Computer Aided Medical Procedures, Technical University Munich, Munich, Germany
chantal.pellegrini@tum.de

[2] Department of Diagnostic and Interventional Radiology, School of Medicine, Technical University of Munich, Munich, Germany

Abstract. Automated diagnosis prediction from medical images is a valuable resource to support clinical decision-making. However, such systems usually need to be trained on large amounts of annotated data, which often is scarce in the medical domain. Zero-shot methods address this challenge by allowing a flexible adaption to new settings with different clinical findings without relying on labeled data. Further, to integrate automated diagnosis in the clinical workflow, methods should be transparent and explainable, increasing medical professionals' trust and facilitating correctness verification. In this work, we introduce Xplainer, a novel framework for explainable zero-shot diagnosis in the clinical setting. Xplainer adapts the classification-by-description approach of contrastive vision-language models to the multi-label medical diagnosis task. Specifically, instead of directly predicting a diagnosis, we prompt the model to classify the existence of descriptive observations, which a radiologist would look for on an X-Ray scan, and use the descriptor probabilities to estimate the likelihood of a diagnosis. Our model is explainable by design, as the final diagnosis prediction is directly based on the prediction of the underlying descriptors. We evaluate Xplainer on two chest X-ray datasets, CheXpert and ChestX-ray14, and demonstrate its effectiveness in improving the performance and explainability of zero-shot diagnosis. Our results suggest that Xplainer provides a more detailed understanding of the decision-making process and can be a valuable tool for clinical diagnosis. Our code is available on github: https://github.com/ChantalMP/Xplainer

Keywords: Zero-Shot Diagnosis · Explainability · Contrastive Learning

C. Pellegrini, M. Keicher and E. Özsoy—These authors contributed equally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_41.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 420–429, 2023.
https://doi.org/10.1007/978-3-031-43904-9_41

1 Introduction

Computer-aided diagnosis systems have become a prominent tool in medical diagnosis. Yet, their adoption is limited by the need for large amounts of annotated data for training, which hinders their scalability and adaptability to new clinical findings [3,12]. Moreover, adapting to a new reporting template or clinical protocol necessitates new annotations, further reducing their feasibility in clinical settings. Recently, zero-shot [1,4,14,15,17] and few-shot [1,4,8] learning methods have been proposed as a potential solution, utilizing contrastive pretraining [13,19] on pairs of radiology reports and images, and achieving performance on par with radiologists [15]. However, these methods lack the level of detail of radiology reports and inherent explainability, impeding their adoption in clinical settings [7]. Particularly, explaining the diagnosis with image descriptors is crucial to increase trust in the system and allow radiologists to verify the results [9].

Inspired by the success of using large language models to predict image descriptors in natural images [10], we introduce Xplainer, a novel framework that enhances the explainability of zero-shot diagnosis in the clinical setting. Xplainer leverages the classification-by-description approach [10] of vision-language models and adapts it to the multi-label medical diagnosis task. Specifically, we task the model to classify the existence of descriptive observations, which a radiologist would examine on an X-Ray scan, instead of directly predicting a diagnosis. This model design imbues our framework with intrinsic explainability, as the final diagnosis prediction is predicated on the underlying descriptor predictions.

We evaluate Xplainer on two chest X-ray datasets, CheXpert [5] and ChestX-ray14 [16], and demonstrate its efficacy in enhancing the performance and explainability of zero-shot diagnosis in the clinical setting. Our results highlight that Xplainer provides a more comprehensive understanding of the diagnosis prediction process, thereby serving as a valuable tool for clinical decision-making. In summary, Xplainer presents a novel framework for zero-shot diagnosis that not only improves explainability and accuracy but also provides an invaluable tool for computer-aided diagnosis.

2 Methodology

2.1 Model Overview

We propose Xplainer, an explainable zero-shot classification-by-description approach for diagnosing pathologies from X-Ray scans. Given an image i and a list of clinical observations $o_{p_{1-n}}$ per pathology p, the goal is to make a multi-label prediction indicating the diagnosis for the patient.

Our zero-shot approach leverages the alignment of image and text embeddings provided by contrastive language-image pretraining (CLIP) [13] and therefore does not require any labeled data. We built upon BioVil [1], a CLIP model pretrained on pairs of radiology reports and images. Employing the text and image encoders from BioVil, we calculate the cosine similarity between an X-ray image and each of N pre-defined clinical observations $o_{p_{1-N}}$ describing a

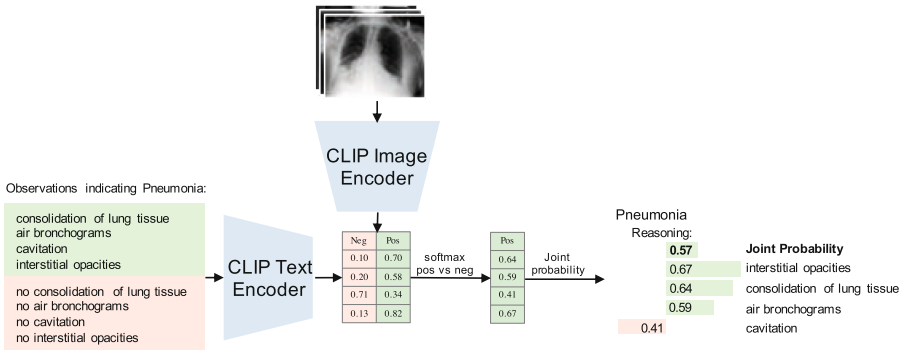

Fig. 1. Overview of Xplainer: In the first step, observation probabilities are calculated based on contrastive CLIP prompting. These are then used to make an explainable diagnosis prediction. The figure depicts an example for Pneumonia.

pathology. Then we calculate observation probabilities $P_{pos}(o_{p_i})$ for every observation. Analogously, we calculate probabilities for the absence of all observations $P_{neg}(o_{p_i})$ by defining negated prompts for all observations. Using the softmax over the positive and negative probability, we calculate the final probability of the presence of an observation $P(o_{p_i})$. Given these observation probabilities $P(o_{p_i}), i \in 1, ..., N$, we estimate a joined probability to determine the likelihood of the presence of a pathology $P(p)$:

$$P(p) = \sum_{i=1}^{N} log(P(o_{p_i})) \div N \tag{1}$$

We repeat this process for all pathologies we want to diagnose in the image. As the prediction of a pathology diagnosis is directly extracted from the observation probabilities, our method is explainable by design, producing a diagnosis prediction and the detected X-ray observations leading to that prediction. Moreover, the observation probabilities show which observations the model mainly considers for its diagnosis. Figure 1 shows an overview of our framework.

To integrate multiple images of one patient, we calculate positive and negative observation probabilities for each image and average them before calculating the pathology probability.

2.2 Prompt Engineering

Successful zero-shot inference relies on a good alignment between the contrastive pretraining and the downstream task [13]. As BioVil [1] was trained on pairs of radiological images and reports, we need to keep our observation prompts close to the style of medical reports. To initialize our prompts, we employ ChatGPT [11] and query it to describe observations in X-ray images that would occur in a radiology report indicating specific pathologies. We further refined the prompts with the help of an experienced radiologist, who manually verified and adapted the descriptors. Human refinement cost was low, taking the radiologist only a few hours. We provide a complete list of the descriptors in the supplementary.

Radiology reports often include both presence and absence of particular observations. When comparing a prompt with an image embedding, it is hard for the model to differentiate between an observation's positive and negative occurrence, as their formulation can be very similar. Previous work [14,15] has shown that introducing negative prompts can circumvent this problem. Therefore, instead of thresholding the similarity between a positive prompt and an image, we prompt the model with both a positive and a negated version of each observation prompt and compare their probabilities. We adapt our prompts in two additional steps to align them with the text in radiology reports. First, we add a disease indication, as radiology reports usually contain observations paired with conclusions. Further, this reduces the ambiguity of our prompts, as in radiology, one sign (e.g., Lung Opacity) can indicate multiple pathologies (e.g., Pneumonia, Atelectasis, or Edema). Additionally, we frame all our observations in a sentence structure sounding more like an actual report by adding "There is/are" before every observation. Putting all of this together, we define the following prompt structure: "There is/are (no) <observation> indicating <pathology>." Lastly, we define contrastive pathology-based prompts to compare to our observation-based prompting. In this setting, only two prompts, one positive and one negative prompt, are used per pathology. Overall, we compare the following styles of prompting to show the benefit of observation-based, contrastive prompting with disease indication and report style:

- **Pathology-based:** (No) <pathology>
- **Basic:** Only positive prompt per pathology: <observation>
- **Contrastive:** (No) <observation>
- **Pathology Indication:** (No) <observation> indicating <pathology>
- **Report Style:** There is/are (no) <observation> indicating <pathology>

3 Experiments and Results

We evaluate Xplainer in a zero-shot setting on the commonly used chest X-ray datasets, CheXpert [5], and ChestX-ray14 [16]. The CheXpert dataset provides a manually labeled validation and test set with 200 and 500 patients, respectively,

Table 1. AUC for zero-shot pathology classification on CheXpert and ChestX-ray14 datasets. *in-domain, as the underlying CLIP model was trained the ChestX-ray14

	CLIP pretraining data	CheXpert		ChestX-ray14
		val	test	test
CheXzero [15]	MIMIC	–	74.73	–
Seibold et al. [14]	MIMIC	78.86	–	71.23
Seibold et al. [14]	MIMIC, PadChest, ChestX-ray14	83.24	–	78.33*
Xplainer	MIMIC	**84.92**	**80.58**	**71.73**

Table 2. AUC per disease on both datasets

	CheXpert Val	CheXpert Test	ChestX-ray14
No Finding	88.82	89.94	–
Enlarged Cardiomediastinum	79.23	80.60	–
Cardiomegaly	78.62	83.32	79.71
Lung Opacity	88.18	91.76	–
Lung Lesion	91.46	69.33	–
Edema	84.84	84.55	81.46
Consolidation	91.56	85.89	71.87
Pneumonia	85.68	83.73	70.83
Atelectasis	84.64	85.46	66.86
Pneumothorax	78.09	83.75	72.18
Pleural Effusion	88.72	89.30	79.11
Pleural Other	83.92	58.67	–
Fracture	–	60.47	–
Infiltration	–	–	68.81
Mass	–	–	70.28
Nodule	–	–	64.74
Emphysema	–	–	74.02
Fibrosis	–	–	62.25
Pleural Thickening	–	–	67.44
Hernia	–	–	74.60
Support Devices/Foreign Objects	80.25	81.15	–

and 14 classes, including "No Finding", "Support Devices/Foreign Objects", and 12 pathology labels. ChestX-ray14 is evaluated on 14 pathology labels on a test set of 25.596 images. For both datasets, we use the official validation and test splits. We perform a multi-label classification for both datasets and evaluate the performance via the Area Under the ROC-curve (AUC) between the positive pathology probabilities and the labels.

Table 1 shows our results compared to previously proposed zero-shot pathology prediction approaches. On CheXpert, we compare with Seibold et al. [14] on the validation set, as they only reported validation performance. For the comparison with CheXzero [15], as well as the ChestX-ray14 dataset, we compare test set results. We outperform both previous works in an out-of-domain setting, where the zero-shot inference is performed on a different dataset than CLIP was trained on. The state-of-the-art results on both datasets show the effectiveness of our observation-based modeling. Further, in Table 2, we provide a detailed breakdown of our results per pathology and dataset.

Table 3. Comparison of different prompting styles on the validation set of CheXpert

	AUC
Contrastive pathology-based Prompting	76.14
Observation-based Prompting:	
Basic Prompt	58.65
Contrastive Prompt	77.00
+ pathology Indication	84.35
+ Report Style	84.92

Table 4. Comparison of ChatGPT prompts vs. refinement with the help of a radiologist

	CheXpert Val	CheXpert Test	ChestX-ray14
ChatGPT prompts	83.61	79.94	71.40
Refined Prompts	**84.92**	**80.58**	**71.73**

Ablation Studies. In our ablation studies, we investigate the impact of our prompt design and the effect of using multiple images. Table 3 shows the results on the CheXpert validation set using different prompting styles. We observe that pathology-based prompting, which reaches an AUC of 76.14%, is significantly worse than observation-based prompting, which reaches an AUC of 84.92%, again highlighting the benefit of observation-based prompting. Comparing the basic observation-based prompting, using only positive prompts per observation, to contrastive prompting, we see a substantial performance gap, showing the importance of using negative prompts to differentiate between positive and negative occurrences. We also show the effect of formulating our prompts unambiguously and in the style of an actual radiology report by adding pathology indication and report style. Adding pathology indication to the contrastive observation-based prompting significantly improves performance, achieving an AUC of 84.35%. Finally, incorporating report style in the prompts leads to the highest AUC of 84.92%, indicating that a contrastive observation-based prompt with pathology indication and report style is the most effective for zero-shot X-ray pathology classification.

Additionally, we compare the initial ChatGPT output to our refined prompts (Table 4). Refinement was performed by deleting irrelevant, redundant, or incorrect descriptors. We observe an improvement through the refinement, indicating that including domain knowledge further improves our method. Nevertheless, the original ChatGPT prompts already perform quite well, showing the impressive potential of combining large generic language models with large domain-specific contrastive models.

For the "No Finding" class, we compare to either define specific prompts such as "Clear lung fields" or "Normal heart size and shape" to classify "No Finding" or model it as the absence of all of the other 13 labels (Rule-based). Table 5

Table 5. Modeling of "No Finding" label with explicit prompts or rule-based definition as lack of other findings

	AUC - No Finding
Explicit Prompting	79.64
Rule-based	**88.82**

Table 6. Comparison of single-view inference to different methods for multi-image processing

	AUC
Only single Frontal View	84.19
All - Max Aggregation	84.77
All - Mean Aggregation	**84.92**

shows that a rule-based modeling of this class leads to better results. A reason for this could be that there is no clearly defined list of observations that indicate a healthy X-ray scan, which a radiologist would mention in his report.

Lastly, we investigate different image aggregation methods for pathology prediction. We compare only using a single frontal view X-ray to using all images available for a patient. For aggregation, we compute positive and negative observation probabilities for every image. In Max aggregation, we then use the highest observation probability. The intuition behind this approach is that an observation might be seen much better from one perspective than another, and then only the perspective where the model is most confident should be used. On the other hand, different views give different insights about which kind of observation a visual cue on the image indicates. To leverage this multi-view information, we test Mean aggregation, where all observation probabilities are averaged over multiple images. The results shown in Table 6 indicate Mean aggregation to be superior, while both aggregation methods outperform using just a single image.

Qualitative Results. Figure 2 shows qualitative examples of our model's predictions. For the true positive prediction, it can be seen that most of the descriptors are detected, and the model recognizes the descriptor "Mass in the mediastinum" as the main indication for the Enlarged Cardiomediastinum. For the True Negative case, the model, correctly, detected none of the descriptors. For the false positive example, one can clearly see that the model made a mistake because it detected an air bronchogram with relatively high certainty and no consolidation. Therefore, this false positive finding is easily falsified by the radiologist since an air bronchogram is a finding that co-occurs with consolidation (i.e., air-filled bronchi in consolidated areas). Thus, knowing which combination of descriptors leads to such a decision substantially improves explainability. In the false positive case, the model misses the pacemaker but detects some implant, showing the model understands there is some foreign object, but can not identify it, which is easily detected by the radiologist. Overall the classification-by-description may facilitate a plausibility check of a specific inference result and an understanding of the source of errors.

Discussion. One downside of modeling a joint probability is that it assumes that all descriptors appear simultaneously and gives all descriptors the same

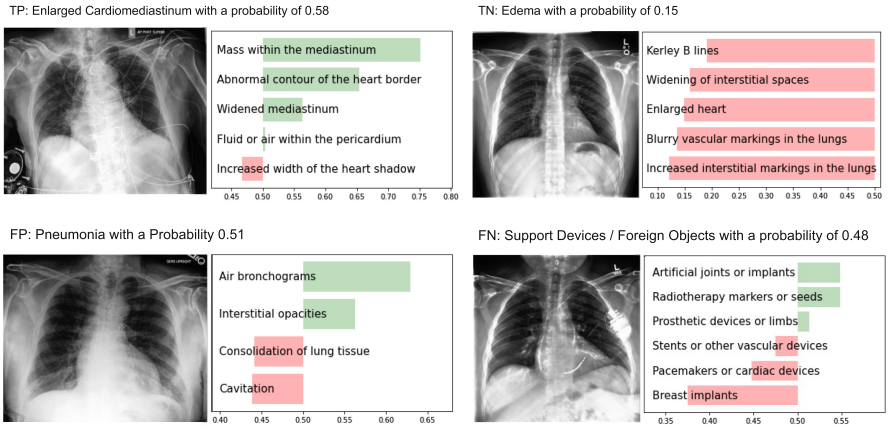

Fig. 2. Qualitative results of Xplainer

importance. While this estimation leads to good results, the assumption does not always hold, as a pathology does not always present with the same signs. Further, there might be inter-dependencies between the descriptors, e.g., there can be descriptors that strongly correlate with the presence of a disease when combined with one descriptor but much less when combined with another. As a first try to model the importance of descriptors, we look into a supervised, out-of-domain approach to model these inter-dependencies. For this, we train a Naive Bayes [2,18] CheXpert classifier on MIMIC-CXR [6], predicting a diagnosis given the descriptor probabilities, allowing the model to focus more on more relevant descriptors. While this approach relies on labels for MIMIC, these labels can be automatically generated by the CheXpert labeler [5], still not requiring human effort for labeling. We observe a slight performance increase on the test set from 80.58% to 81.37% AUC. This shows that the descriptor importance learned on MIMIC can partially be transferred to an out-of-domain dataset. We believe investigating methods to consider varying importance and complex relations between the descriptors is an essential and exciting direction to investigate in future work. Moreover, as Xplainer is not tied to specific image and text encoders, orthogonal works that lead to better encoders can be used to improve our results further.

The use of descriptors in Xplainer provides a flexible and adaptive approach to automated diagnosis prediction. By identifying and classifying the presence of descriptive observations, our model can capture the underlying characteristics of a disease without relying on labeled data. This means that our system can easily adapt to new settings with different clinical findings, including new conditions where the symptoms are known, but there is no training data available yet. Additionally, using descriptors allows for adapting the system to specific populations, where the essential descriptors can differ. This is because the model is

not constrained by pre-defined labels but rather by the meaningful underlying features of a given diagnosis.

4 Conclusion

In this work, we present a novel and effective zero-shot approach for chest X-ray diagnosis prediction, which provides an explanation for the model's decision. We leverage BioVil, a pretrained, domain-specific CLIP model, and use contrastive observation-based prompting to make predictions without label supervision. Our approach significantly outperforms previous zero-shot methods on CheXpert and Chest-Xray14, showcasing the effectiveness of our approach. Furthermore, we show that designing informative prompts is crucial to improve model performance. Our ablation studies demonstrate that adding disease indication and report style formulation to observation-based prompts notably enhances performance, underscoring the importance of aligning prompts with the domain-specific language used in medical reports. Additionally, contrastive prompts significantly boost performance, suggesting that the model can benefit from explicitly contrasting positive and negative examples.

Our work highlights the potential of contrastive pretraining combined with observation-based prompting as a promising avenue for zero-shot medical image classification, where labeled data is scarce or expensive to obtain, and explainability is vital. We envision that our approach can be extended to other medical imaging domains and have practical applications in real-world scenarios. Our findings contribute to the growing body of research to improve the accuracy and interpretability of medical image diagnosis.

Acknowledgements. The authors gratefully acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) under project DIVA (FKZ 13GW0469C) and the Bavarian Research Foundation (BFS) under project PandeMIC (grant AZ-1429-20C).

References

1. Boecking, B., et al.: Making the most of text semantics to improve biomedical vision-language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part XXXVI, pp. 1–21. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_1
2. Chan, T.F., Golub, G.H., LeVeque, R.J.: Updating formulae and a pairwise algorithm for computing sample variances. In: Caussinus, H., Ettinger, P., Tomassone, R. (eds.) COMPSTAT 1982 5th Symposium held at Toulouse 1982: Part I: Proceedings in Computational Statistics, pp. 30–41. Springer, Cham (1982). https://doi.org/10.1007/978-3-642-51461-6_3
3. Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.J., Ducoffe, M.: Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng. Appl. Artif. Intell. **92**, 103678 (2020)

4. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951 (2021)
5. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
6. Johnson, A.E., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 317 (2019)
7. Kayser, M., Emde, C., Camburu, O.M., Parsons, G., Papiez, B., Lukasiewicz, T.: Explaining chest X-ray pathologies in natural language. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part V, pp. 701–713. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_67
8. Keicher, M., Mullakaeva, K., Czempiel, T., Mach, K., Khakzar, A., Navab, N.: Few-shot structured radiology report generation using natural language prompts. arXiv preprint arXiv:2203.15723 (2022)
9. McInerney, D.J., Young, G., van de Meent, J.W., Wallace, B.C.: CHiLL: zero-shot custom interpretable feature extraction from clinical notes with large language models. arXiv preprint arXiv:2302.12343 (2023)
10. Menon, S., Vondrick, C.: Visual classification via description from large language models. arXiv preprint arXiv:2210.07183 (2022)
11. OpenAI: Chatgpt. chat.openai.com. Accessed 8 Mar 2023
12. Qin, C., Yao, D., Shi, Y., Song, Z.: Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomed. Eng. Online **17**(1), 1–23 (2018)
13. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
14. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part V, pp. 690–700. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16443-9_66
15. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat. Biomed. Eng., 1–8 (2022)
16. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)
17. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
18. Zhang, H.: The optimality of Naive Bayes. In: Barr, V., Markov, Z. (eds.) Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004). AAAI Press (2004)
19. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference, pp. 2–25. PMLR (2022)