Strengths and Limitations of Word-Based Task Explainability in Vision Language Models: a Case Study on Biological Sex Biases in the Medical Domain

Lorenzo Bertolini[1] **Valentin Comte**[2] **Victoria Ruiz-Serra**[2]
Lia Orfei[1] **Mario Ceresa**[1]
[1]European Commission, Joint Research Centre (JRC), Ispra, Italy
[1]European Commission, Joint Research Centre (JRC), Geel, Belgium
name.surname@ec.europa.eu

Abstract

Vision-language models (VLMs) can achieve high accuracy in medical applications but can retain demographic biases from training data. While multiple works have identified the presence of these biases in many VLMs, it remains unclear how strong their impact at the inference level is. In this work, we study how well a task-level explainability method based on linear combinations of words can detect multiple types of biases, with a focus on medical image classification. By manipulating the training datasets with demographic and non-demographic biases, we show how the adopted approach can detect explicitly encoded biases but fails with implicitly encoded ones, particularly biological sex. Our results suggest that such a failure likely stems from misalignment between sex-describing features in image versus text modalities. Our findings highlight limitations in the evaluated explainability method for detecting implicit biases in medical VLMs.

1 Introduction

Foundation and vision-language models (VLMs) have found many successful applications in the general and medical domains (Radford et al., 2021; Wang et al., 2022; Huang et al., 2023; Kim et al., 2024; Moor et al., 2023; Chen et al., 2023; Huang et al., 2023; Khattak et al., 2024; Abbaspourazad et al., 2024; Wang et al., 2024; Li et al., 2025b,a; Khan et al., 2025). While powerful, VLMs can encode harmful demographic biases and stereotypes (Berg et al., 2022; Ruggeri and Nozza, 2023; Mandal et al., 2023; Alabdulmohsin et al., 2024; Hamidieh et al., 2024; Bartl et al., 2025), that can also expand to systems that rely on them as backbone structure, such as text-to-image models (Bianchi et al., 2023; Tanjim et al., 2024). Recently, Yang et al. (2024a) found similar patterns in the medical domain, showing how general and medical VLMs can under-diagnose marginalized demographic groups, adopting bias learned from the training data. Analogous evidences were found by multiple studies, which show how different types of machine learning models used in the medical field tend to encode and produce harmful biased predictions against underrepresented demographic groups (Larrazabal et al., 2020; Seyyed-Kalantari et al., 2021; Yang et al., 2024b).

These results highlight the strong need for mechanisms to trace and quantify possible biased behaviours and knowledge encoded in VLMs, especially when a validation set is unavailable for a given task. Aside from tracing and mitigating biased distribution in training sets, and using ad-hoc metrics (see Bartl et al. (2025) for a review), instance-level explainability (XAI) methods based on saliency maps are among the most adopted methods to trace biases in VLMs (Agarwal et al., 2023; Mandal et al., 2023; Tanjim et al., 2024; Bartl et al., 2025). While instance-based XAI methods can effectively and intuitively convey their findings, they struggle to reveal broader patterns on how a model is systematically impacted during a classification task, across a full dataset.

These limitations are addressed by concept-based and task-level XAI methods (Kim et al., 2018; Ghorbani et al., 2019; Yan et al., 2023; Agarwal et al., 2023; Menon and Vondrick, 2023), which focus on gathering descriptions of the differences between a task's classes. Since visual explanations can be less effective in conveying cross-category differences, Agarwal et al. (2023) have proposed a word-based task-level XAI methodology leveraging a VLM's joint embedding space. The proposed approach aims at reconstructing the coefficients of a logistic regression, fit to discriminate between images of healthy and clinical patients, by learning a linear combination of word embeddings (see Figure 1). Intuitively, this will result in learning which subset of a pre-defined vocabulary is more descriptive of one category (e.g., disease patient) versus another (e.g., healthy patient).

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 111–123

In their work, Agarwal et al. (2023) show how this approach can capture meaningful aspects of medical diagnosis, such as the one between the *roundness* of a skin lesion and the high likelihood of it being benign, or its *asymmetry* and the high probability of such lesion being malignant. In this study, we propose to further test such an approach, to trace and quantify more implicit features and biases encoded in both individual images and overall datasets. We do so with two experiments, both injecting controlled amounts and types of biases in an X-Ray-based classification task. In the first experiment, we focus on explicitly quantifiable image characteristics, namely brightness and blurriness, while for the second experiment, we focus on controlling the association between a specific biological sex and the likelihood of such group of patients to be diseased or healthy.

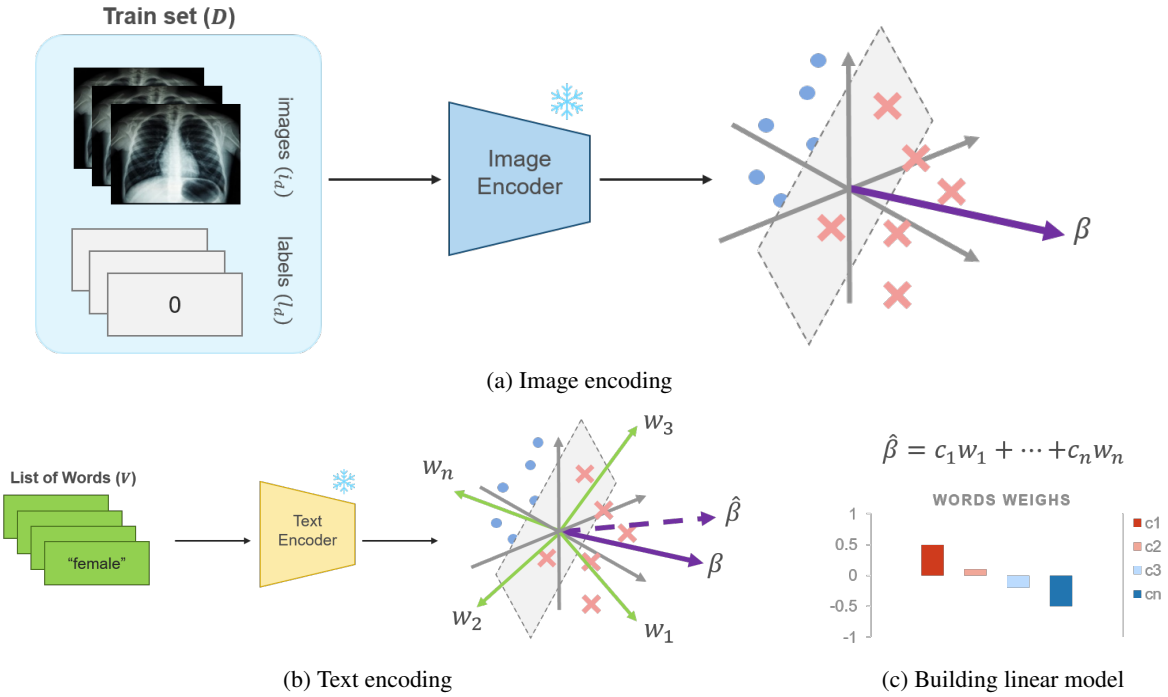
Using both a general and a medical VLM, our results show how the adopted approach can detect biases that are explicitly encoded in the images (i.e., brightness and blurriness), but fails at detecting more implicit biases connected to biological sex imbalance in the data, producing incoherent predictions, with highly variable and inconsistent patterns that resist straightforward interpretation. These findings highlight the need for more robust methodologies before making definitive claims about bias quantification in medical VLMs.

2 Related Work

Demographic biases in VLMs Ruggeri and Nozza (2023) proposed the first multimodal analysis and metrics to detect and quantify demographic biases in VLMs across the two modalities, showing how these biases are not only independently encoded in each separate modality, but can influence and propagate across modalities. Mandal et al. (2023) study the effectiveness of data-balancing methods for debiasing VLMs. Results show that fine-tuning can be effective against some type of biases, though the impact on quality is not always positive. Mandal et al. (2023) used GradCAM (Selvaraju et al., 2017), to show how CLIP (Radford et al., 2021) encodes societal gender bias, for example by associating concepts like *programmer* to male figures, and *gossipy* or *homemaker* to female ones. Yang et al. (2024a) found that a medical VLM for chest X-ray diagnosis consistently underdiagnosed marginalized groups, especially those with intersectional identities like black female patients. Crucially, the analysis of the word embedding reveals that the model consistently encoded demographic information with an accuracy exceeding human radiologists, creating bias across multiple pathologies and patient populations.

Demographic bias in medical AI Alongside research on VLMs, research on bias in medical AI systems has grown increasingly comprehensive. Larrazabal et al. (2020) demonstrated how gender imbalances in training data lead to biased convolutional neural network (CNN) classifiers for chest X-ray images. Seyyed-Kalantari et al. (2021) expanded the analysis to examine how AI systems underperform across broader demographic dimensions including age, sex, and ethnicity. Yang et al. (2024b) further revealed that CNN-based visual classifiers often exploit demographic characteristics as shortcuts when making disease classifications, compromising diagnostic accuracy.

Concept-based XAI Kim et al. (2018) introduced Concept Activation Vectors to interpret image classification by associating user-defined concept classes with neural network activations. A linear classifier separates activations of images containing the concept from those that do not, to understand how concepts influence the model's predictions.Yan et al. (2023) expanded on Kim et al. (2018) to build a human-in-the-loop diagnostic tool, based on enhancing confounding behaviours, and limiting spurious correlations, focusing on a skin cancer diagnosis task. To do so, the authors built a model learning an interpretable space able to detect concept (e.g., *darker border*) distributions in each class (e.g. *benign*). Being based on a CNN, the method still lacks any form of language knowledge, and hence, concepts are still defined post hoc, based on the CNN kernels. Agarwal et al. (2023) recently proposed to alleviate the limitation of vision-only concept discovery by leveraging VLMs, that also possess language-based knowledge. The core idea (see Figure 1) is to reconstruct the logistic classifier trained to discriminate between benign/malignant images, encoded with the a VLM's images encoder, by learning a linear combination of pre-selected words, encoded with the VLMs' text encoder. Similarly to Kim et al. (2018), this procedure will learn which concepts are more associated with a class or another, but offer more plasticity and robustness, as the only human intervention is limited to the dictionary selection, which can contain more interpretable and reliable general or medical concepts.



(a) Image encoding

(b) Text encoding

(c) Building linear model

Figure 1: Experimental method. Agarwal et al. (2023)'s method for task-level explainability is composed of three main steps: i) image encoding, and logistic regression (Figure 1a); ii) word encoding and linear modelling (Figure 1b); iii) interpretation of linear model's coefficients (Figure 1c). Diagrams adapted from Agarwal et al. (2023).

3 Bias Statement

From the medical and diagnostic perspective, we consider as bias the spurious association, created by the model, or contained in the data, between non clinically relevant traits or characteristics and disease likelihood. As demonstrated throughout the previous sections, such conditions appear to afflict medical datasets and AI models, manifesting through systematically different prediction rates across demographic groups when controlling for actual disease prevalence.

These biases are harmful because they do not necessarily reflect real-world distributions (Yang et al., 2024a), and can perpetuate or amplify existing health disparities through several mechanisms: 1) *Resource inequality*: biased predictions lead to inequitable distribution of healthcare resources, with some demographic groups receiving more accurate diagnoses and timely interventions than others (Obermeyer et al., 2019); 2) *Care quality gaps*: systematic performance differences compromise care quality for certain groups of people (Fiscella and Sanders, 2016); 3) *System distrust*: consistent misdiagnosis of certain demographic groups undermines trust in AI systems within those communities and potentially exacerbates historical mistrust in healthcare systems (Richardson et al., 2021). 4) *Policy misalignment*: if biased AI-systems were used to inform health policies (without awareness/quantification of the underlying biases), their results may fail in appropriately capturing actual population needs and in return might create regulatory gaps that undermine the goal of ensuring equitable healthcare.

4 Experimental Set-Up

This work has two main experiments, both using the method proposed by Agarwal et al. (2023). The first experiment is designed as a proof of concept or stress-test of the original work. The second experiment examines the method's ability to detect gender biases. Both experiments have the same core process, models, base dataset, and list of explainable words. These aspects are explained in more detail in the following subsections.

4.1 Method

The method is composed of three main steps: i) image encoding, and logistic regression (Figure 1a); ii) word encoding and linear modeling (Figure 1b); iii) interpretation of linear model's coefficients (Figure 1c). The method is graphically summarized by Figure 1's diagrams.

More formally, assuming a training set $D_{n=1}^{d} =$

$\{(i_1, l_1), ..., (i_d, i_d)\}$, with I_n and l_d being an image and its classification label, a pre-trained dual-encoder VLM, with an image encoder E, and a text encoder T, and a set of pre defined words $V_{n=1}^{v} = \{w_1, ..., w_v\}$, Agarwal et al. (2023)'s method use E to encode all images in D, and fit a logistic regression (Figure 1a), obtaining a vector β, containing the logistic regression's coefficients. Then, use T to embed V in the joint embedding space, and use the obtained word embedding to fit a linear model approximating β ($\widehat{\beta}$) (Figure 1b). Lastly, we interpret the linear model's coefficients (e.g., c_1 in Figure 3) for each word vector. Following Agarwal et al. (2023), we interpret positive weights as alignment with class 1 prediction. We include significance levels for each coefficient of the linear model.

Agarwal et al. (2023)'s method also includes a solution to select prototypical images for each word. The original approach calculates the residuals between the dot product computed between all images and all words, and the predicted dot product, obtained by fitting a linear regression using all images and all words but one, i.e. the "target" word. The higher the residual, the worse the fit; the image corresponding to the highest residual is considered the worst represented image by the set of words used in the linear regression and should hence be the most prototypical of the "target" word. However, since this approach considers the signed values of the residuals, the highest one would always be the largest positive residual. We therefore use the absolute value of the residuals to ensure that we capture the overall largest distance between the dot products. Aside from this minor modification, we adopt the original method and source code.

4.2 Models

The original work of Agarwal et al. (2023) adopts CLIP (Radford et al., 2021), since their method assumes a VLM with a joint embedding space and the possibility of using the frozen encoders for downstream tasks, such as image classification. In addition to CLIP, we adopt UniMedCLIP (Khattak et al., 2024), a general-purpose medical VLM trained in multiple medical fields, including X-Ray.

4.3 Data

We focus on X-Ray images due to their extensive use in AI and machine learning research, using the widely adopted CheXpert-5X200 dataset [1] (Khattak et al., 2024), which was derived from full CheXPert dataset (Irvin et al., 2019) following an established procedure (Huang et al., 2021). More in detail, CheXpert-5X200 is a dataset containing 1,000 X-ray images randomly sampled from the main dataset, comprising 200 images for each of five medical conditions: atelectasis, cardiomegaly, edema, pleural effusion, and pneumonia. To align with our binary classification approach, we selected cardiomegaly as our target condition, where 1 indicates the presence and 0 indicates the absence of the condition.

We selected cardiomegaly as our target condition because it exhibited the smallest sex disparity among positive diagnoses (class 1). Since our work focuses on studying biological sex biases, we hence added extra filtering to the data to balance the distribution of sex across the two classes. We then randomly split this data into an 80-20% ratio between training and test set.

4.4 Words

Agarwal et al. (2023)'s work adopts a list of words automatically generated with ChatGPT (Brown et al., 2020), obtained by asking the model for relevant image-property words (e.g., *color*), and subsequently requesting positive and negative adjectives describing such properties (e.g., *light*, *dark*). This approach can be effective for both general and medical purposes explanations, as it can span across diverse datasets as demonstrated in Agarwal et al. (2023)'s work. However, we focus on a single condition: cardiomegaly. For this reason, we generate a new selection of words. Mirroring Agarwal et al. (2023)'s method, we prompted Claude 3.7 Sonnet (Anthropic, 2025) to generate properties and adjectives useful to describe cardiomegaly, resulting in the list presented in Table 1. Code and data are available here[2].

5 Experiment 1: Image Feature Bias

Agarwal et al. (2023) provided evidence that their method can efficiently model explicit or semantic image properties, such as "round". While an object's roundness *can* be mathematically quantified, this becomes challenging with images depicting skin lesions due to factors like camera angle. Evaluating such properties would require human experts

[1] https://github.com/mbzuai-oryx/UniMed-CLIP/blob/main/local_data/chexpert-5x200.csv

[2] https://github.com/jrcf7/GeBNLP_25

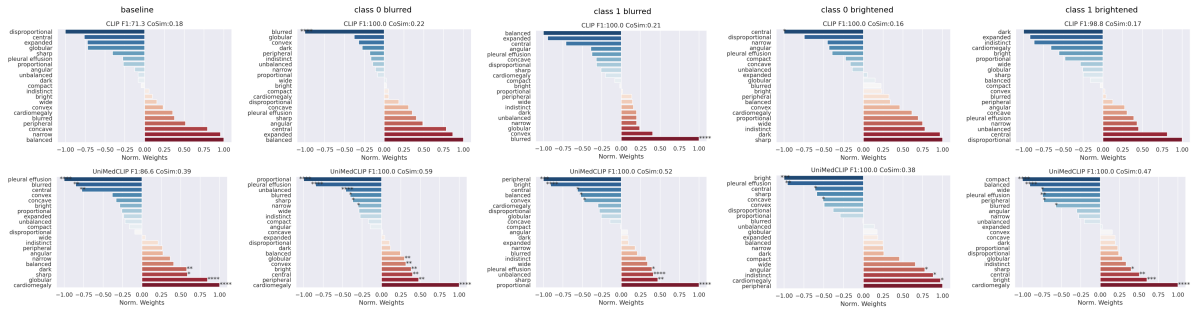


Figure 2: Experiment 1 results. Normalized word coefficients for CLIP (top row), and UniMedCLIP (bottom row) for original images (baseline, leftmost column) and systematically altered images (blurred: columns 2-3; brightened: columns 4-5). Plots display words (y-axis) and corresponding normalized coefficient values (x-axis). Positive coefficients (red bars) and negative coefficients (blue bars) indicate the direction of association. Panels' header displays performance metrics (F1 and Cosine Similarity). Asterisks indicate statistical significance (~*: $.08 \leq p \leq .05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$; ****: $p < .0001$).

Property	Adjective 1	Adjective 2
Size	narrow	wide
Shape	angular	globular
Border	indistinct	sharp
Width Ratio	proportional	disproportional
Position	peripheral	central
Contour	concave	convex
Distribution	balanced	unbalanced
Silhouette	compact	expanded

Table 1: List of selected words shared across experiments. Each row represents a visual property of cardiomegaly in X-ray images with the corresponding opposing adjective pair (adjective 1 and adjective 2).

to assess the method's effectiveness for characteristics like "roundness" or "symmetry" — an effective approach which lacks efficiency and objectivity. To better assess the method's stability, we tested its ability to detect fully controllable biases by applying quantifiable transformations to images: light alteration and blurriness.

Words The experiment includes the addition of specific words to the original set: "bright","dark", "blurred", "sharp", "cardiomegaly", and "pleural effusion". These words were chosen to evaluate the models' performance based on both visual attributes and clinically relevant features.

Dataset manipulation A new dataset was created to introduce controlled variations in brightness and sharpness. This dataset includes images with added blur and altered light intensity to assess the models' robustness to these perturbations and their ability to associate textual concepts with visual alterations. See Appendix A for more details.

5.1 Results

The results of the experiment on altered brightness and blurriness are presented in Figure 2.

UniMedCLIP outperforms CLIP on baseline images (unaltered) with higher F1-score and cosine similarity, which is expected given that it has been trained on the same dataset of radiography images (Irvin et al., 2019). This alignment allows UniMedCLIP to correctly associate the words "cardiomegaly" and "pleural effusion" with their corresponding classes. Furthermore, UniMedCLIP assigns statistically significant weights to the most influential words, whereas none of the word associations appear statistically significant for CLIP.

When blurred images from classes 0/1 (no cardiomegaly/cardiomegaly) are analyzed, (second and third columns of Figure 2 respectively), CLIP assigns greater weight to the word "blurred", indicating stronger visual feature alignment. In contrast, UniMedCLIP shows minimal, and non-significant association with this term. With brightness alterations (Figure 2, fourth and fifth columns), both models respond to these manipulations. CLIP associates "dark" with relatively reduced brightness in either class, while UniMedCLIP links "bright" with relatively increased brightness.

Collectively, the results of this experiment show that the method proposed by Agarwal et al. (2023) is sensitive to induced visual biases in CLIP and UniMedCLIP for the set of X-ray cardiomegaly images, showing the expected alignment between the relevant words and the modified image features.

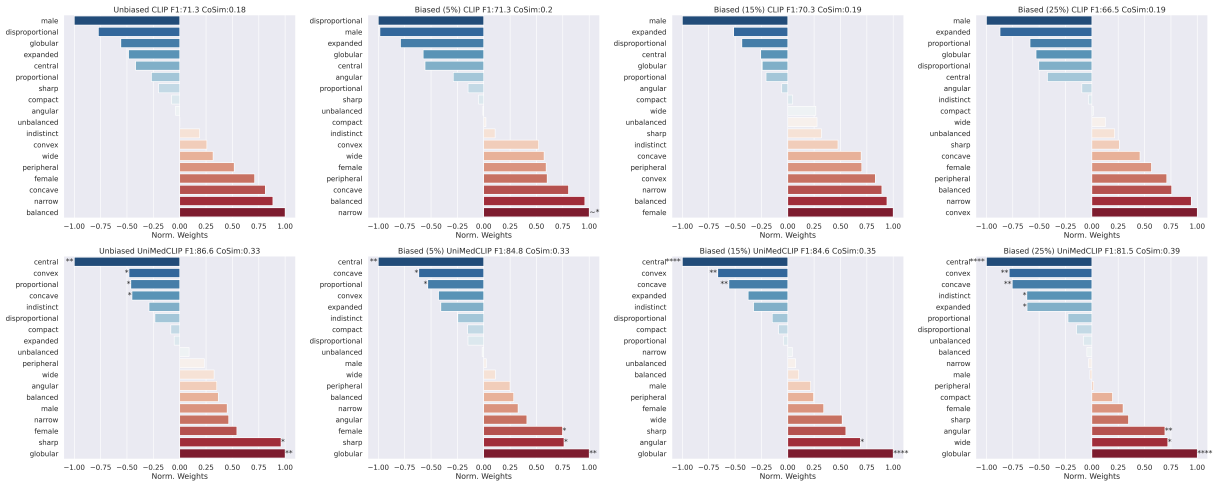


Figure 3: Experiment 2 results. Normalized word coefficients for CLIP (top row), and UniMedCLIP (bottom row) for unbiased (leftmost column) and sex-biased (in a proportion of 5, 15 and 25%) datasets. Plots display words (y-axis) and their corresponding normalized coefficient (x-axis). Positive coefficients (red bars) and negative coefficients (blue bars) indicate the direction of association. Panels' header displays performance metrics (F1 and Cosine Similarity). Asterisks indicate statistical significance (~*: $.08 \leq p \leq .05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$; ****: $p < .0001$).

6 Experiment 2: Biological Sex Bias

In this experiment, we test the ability of Agarwal et al. (2023)'s method to trace sex-based stereotypes. While biological sex may not be as immediately obvious as characteristics like roundness in images, certain sex-based anatomical features may still be detectable in chest X-rays, such as differences in breast tissue.

Words We added "female" and "male" to refer to biological sex rather than gender. This distinction follows established guidelines for scientific precision (DG RTD, European Commission, 2020).

Dataset manipulation In Experiment 1, we injected the bias by manipulating images belonging to one of the two classes. For Experiment 2, we create a disparity in the proportion of sex distribution within each class. To do so, we manipulate the starting dataset, described in Section 4.3, so that a specific sex is more represented in class 1 by increasing percentages. To mimic real-world distributions (Fairweather et al., 2023), we increase the percentage of males with pathology instances while simultaneously decreasing the instances of healthy males. In other words, we built a series of datasets with a bias toward male sex being a predictor for sickness (class 1) and female sex being a predictor for the absence of the cardiomegaly condition (class 0). See Appendix B for more details.

6.1 Results

Following the same format of results as in Section 5.1, the results for normalized word coefficients for different models (rows) and datasets (columns) are presented in Figure 3. More in detail, UniMedCLIP shows higher, more stable performance across datasets with consistently higher cosine similarity scores than CLIP. This indicates how well the linear model built with word embeddings ($\widehat{\beta}$, Figure 1b) approximates the logistic classifier β (Figure 1a). Only the linear models built with UniMedCLIP embeddings produce significant coefficients. These results suggests that UniMedCLIP is more reliable for this approach—expected given its training on X-Ray data.

Single coefficients analysis leads to similar conclusions. To reiterate, positive coefficients for a word indicate alignment with class 1 prediction (i.e., cardiomegaly). UniMedCLIP results show coherence, with relevant adjectives like "globular" and "sharp" having the highest positive scores and significance compared to CLIP. However, both models show unexpected sex-describing words results. We expected no impact in the unbiased dataset, with increasing "male" and decreasing "female" coefficients as bias increased. Instead, both models show little to no impact on the two coefficients across datasets and attribute higher coefficients to "female" than "male", with CLIP showing "male" as the most negative coefficient.

These findings might suggest that models do not

use sex information in the inference process despite our bias injection. However, results from Agarwal et al. (2023), our previous experiment, and the reported significance in one of the UniMedCLIP test, where "female" showcases a strongly positive *and* significant coefficient, might suggest that the system may simply fail to detect the models' use of sex bias. To clarify these findings, we conduct in the following subsections quantitative and qualitative analyses of textual and visual encodings associated with sex-related words.

6.1.1 Quantitative analysis: prototypical images

As mentioned in Section 4.1, we adopt a modified version of Agarwal et al. (2023)'s system, to extract the N most prototypical images for each word. We compared the system's prediction of male/female images (i.e., that a given image is prototypical of, and hence belongs to, a male/female patient) with patient's actual biological sex. This helps determine whether models are able to extract sex information implicitly or whether the inconsistencies in Figure 3 stem from poor sex encoding. As the original work does not indicate a strategy for determining the optimal number of prototypical images per dataset, we retrieve the top 100 prototypical images for "male" and "female", and evaluated their alignment with metadata.

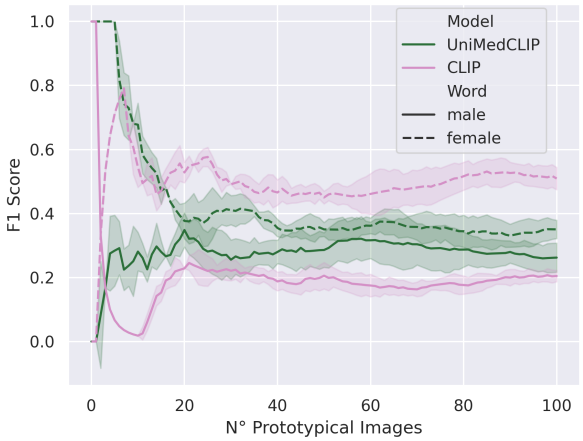


Figure 4: Experiment 2, prototypical image analysis. F1 scores (y-axis) as a function of the number of top N prototypical images (x-axis) extracted for words "male" (solid line) and "female" (dashed line) for UniMedCLIP (green) and CLIP (pink) models. Shades indicate standard deviation across tested datasets.

Figure 4 shows the weighted F1 scores as a function of the number of top 100 prototypical images. The models produce remarkably different results, which appear specular *within* each model. For UniMedCLIP, "male" and "female" start at opposite extremes (0 and 1 respectively) before converging to similar scores at around 20 prototypical images. CLIP exhibits comparable initial boundary conditions (1 for "male" and 0 for "female"), followed by rapid inversions that eventually stabilize with scores remaining distinctly separated beyond 20 images. Overall, performance generally remains poor, even when considering 20 or fewer prototypical images. The near-perfect or near-zero initial results suggest the system is essentially guessing the sex of patients. This indicates that the method fails to detect the injected sex bias due to its inability to extract sex information encoded in the multimodal embeddings. Overall, these results suggest that the method is inconsistent for detecting biological sex bias, as evidenced by the unstable performance metrics and the system's apparent inability to reliably extract injected imbalanced sex information encoded in the multimodal embeddings.

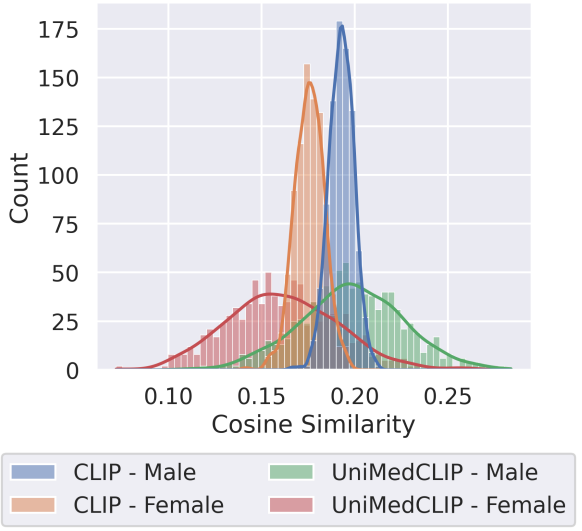
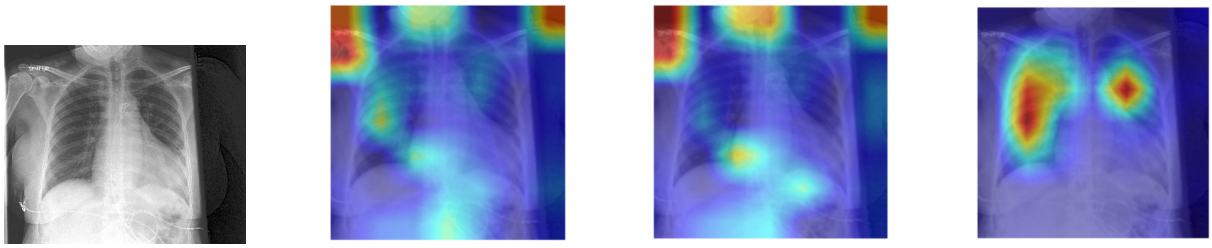


Figure 5: Experiment 2, cosine similarity analysis. Distributions of the cosine similarity scores obtained comparing each image from the unbiased train and test set with the word "male" and "female".

6.1.2 Quantitative analysis: similarity scores

To further investigate the limitations of the prototype-based approach for detecting gender bias, we analysed the underlying similarity distributions between image embeddings and gender-specific textual representations. Figure 5 provides a potential partial explanation for the method's shortcomings by summarising the distribution of the cosine similarity scores between each image and



(a) Original (b) Female (c) Male (d) Lungs

Figure 6: Experiment 2, qualitative analysis: CLIP attention maps. Each diagram summarizes the internal activation of CLIP when the image encoder is prompted with the same image (6a) (female patient), and the textual encoder is prompted with "female" (6b), "male" (6c), and "lungs" (6d).

the words "female" or "male".

As shown in the figure, despite having drastically different shape, both models demonstrate a marked preference for one of the two word, in this case "male". This imbalance in the similarity distribution suggests an inherent bias in how the models encode gender-related concepts, regardless of the actual gender information present in the medical images. The skewed distributions could explain why the prototypical image extraction process yields inconsistent F1 scores as observed in our previous analysis.

6.1.3 Qualitative analysis: attention maps

To complement our quantitative findings and gain deeper insights into how these models process sex information, we conducted a qualitative analysis of model attention. By visualizing where the model focuses when prompted with "female" and "male' terms, we can better understand potential disconnects between human anatomical understanding and model representation. We applied the attention visualization method from Chefer et al. (2021) to study the activation patterns in the image encoder. We analyzed the same chest X-ray image from a female patient using three different input words: "female", "male", and "lungs". Due to implementation constraints in the code, we limit the analysis to CLIP. Results are presented in Figure 6.

The results reveal that attention patterns for "male" and "female" prompts are strikingly similar, which is not entirely unexpected. However, these patterns do not seem to align with anatomical expectations for gender recognition in chest X-rays, such as focus on the breast area. Conversely, the attention pattern for "lungs" appears coherent and anatomically appropriate, suggesting that the model may have learned meaningful representations for organ structures but not for sex-specific features in this medical imaging context. These findings further support previous results and suggest that the selected VLMs may not be encoding biological sex information in ways that align with human anatomical understanding. This misalignment between model attention and expected anatomical features could explain the poor performance in detecting injected sex data imbalance observed in our previous experiments.

7 Discussion and Conclusions

A consistent body of evidence has shown how many AI models, including VLMs, can encode harmful biases and stereotypes based on demographic features, such as ethnicity or biological sex. These biases have been shown to negatively impact the performance of these models, and it is hence essential to trace and quantify their impact at inference time, especially in a crucial field as medical decision-making. In our work, we have focused on a task-level approach to explainability, aiming at understanding if it can coherently trace explicitly (e.g. brightness) or implicit (i.e., biological sex) bias distributions that we have injected in a medical image classification task. Our experiments, which use the task-level explainability method proposed by Agarwal et al. (2023), reveal important limitations in this method for detecting implicit biases in medical VLMs. While Experiment 1 demonstrated the method's effectiveness in detecting explicit visual modifications like brightness and blurriness (see Figure 2), Experiment 2 exposed its failure to detect sex-based biases. Despite deliberately manipulating the datasets to enhance the association between a specific biological sex and disease presence, the method failed to detect these manipulations in both CLIP and UniMedCLIP models.

Such failure could indicate that the models are not using biological sex information in the classifi-

cation process, so we performed a detailed analysis. Results strongly suggest a fundamental issue: the misalignment between how biological sex is represented in these models versus how humans would interpret it. To start, the prototypical image analysis produced remarkably poor performance (see Figure 4), indicating the system was essentially guessing patients' biological sex rather than detecting meaningful patterns. Moreover, our qualitative investigation showed how CLIP's image encoder internal activations appear remarkably similar for the two sexes. While this evidence is in line with the basic assumption behind distributional modeling (i.e., similar concepts occupy a close position in the latent space), we notice how the "behaviour" of the model appears poorly aligned with our expectation on where we might focus to make a distinction between biological-sex in a chest image (see Figure 6). Such evidence might seem in contrast with the intuition that VLMs might hold better and more grounded knowledge, thanks to their dual-modality modeling. However, recent preliminary evidence suggests that VLMs might in fact, be less aligned with human internal representations (Bavaresco and Fernández, 2025).

To conclude, this work presented an extensive analysis of the ability of a task-level explainability method based on linear combination of word embeddings to detect implicit and explicit biases by focusing on injecting quantifiable biases, such as brightness and blurriness altering, and more implicit biases, such as patients' biological sex. The first experiment's results are in line with the original work, showing that the system is able to detect imbalances in the data when they are related to explicit features. However, results from the second experiment showed how the method is not able to coherently detect implicitly encoded biases such as the biological sex. Our analysis suggested that this is likely due to a misalignment of the concept in the two modalities.

Limitations

The limitations of our study stem primarily from two fundamental sources, namely the inherent constraints of our chosen methodological approach and the characteristics of the available data, which are detailed in the following paragraphs.

Fixed vocabulary and dichotomisation Our methodology favors binary descriptors. For human interpretability though, this is not strictly required. While biological sex (male/female) and some clinical features might work in this format, demographic factors like age and ethnicity are harder to force into binary distinctions. This limitation is particularly relevant given the growing body of evidence that intersectional demographic factors significantly impact healthcare outcomes (Vohra-Gupta et al., 2022).

Disease-specific image characteristics/vocabulary Each medical condition presents unique visual characteristics that demand tailored descriptive vocabulary. The adjectives appropriate for describing cardiomegaly features (such as "enlarged", "prominent", or "distended") differ substantially from those that would effectively characterize other conditions like pneumonia or fractures. Our approach did not rely on a universal set of descriptive words across different pathologies, as the visual manifestations vary dramatically. This complicates cross-condition comparisons and demands expert knowledge to select appropriate terms for each studied condition.

Sex representation Due to the lack of metadata, or study focus on biological sex as a binary variable (male/female), which poses inherent limitations for comprehensive bias analysis. This approach fails to account for non-binary individuals and diverse anatomical variations.

Gender representation We assume that the metadata available from CheXpert corresponds to biological sex only and does not take into account gender representation. That is why we consider the potential impact of sex on our results only. However, in medical contexts, "sex" and "gender" are often used interchangeably, but we are unable to distinguish between them, so we rely on the sex variable. Additionally, since our analysis does not capture the complexities of gender identities and expressions, it may not be representative of individuals whose gender identity does not align with their assigned sex at birth.

Metadata availability The validation of our methodology heavily depends on the availability of demographic metadata in medical imaging datasets. While such information is crucial for comprehensive bias analysis, it is often not publicly available due to privacy concerns and data protection regulations. This limitation constrains the broader applicability of our approach and highlights the need for balanced solutions that address both privacy requirements and the imperative for algorithmic fairness assessment. Initiatives such as the one

developed by Luo et al. (2024), which introduced the *Harvard-FairVLMed* dataset, are highly encouraged in this aspect, since they offered a dataset that includes demographic attributes, ground-truth labels, and clinical notes.

Ethical considerations

Our research on bias detection in medical AI adheres to responsible AI principles. We used only medical images hosted in public repositories. We acknowledge the limitations of binary categorizations and recognize that bias detection itself carries assumptions. As our findings may influence clinical systems, we emphasize this work is a starting point for ongoing evaluation, not a comprehensive solution. We remain committed to developing medical AI that benefits all patients equitably, requiring continuous assessment across diverse populations.

Acknowledgments

We would like to thank the colleagues of the Digital Health Unit (JRC.F.7) at the Joint Research Centre of the European Commission for helpful guidance and support. The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. 2024. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations*.

Shobhit Agarwal, Yevgeniy R. Semenov, and William Lotter. 2023. Representing visual classification as a linear combination of words. In *ML4H@NeurIPS*.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Peter Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. 2024. CLIP the bias: How useful is balancing data in multimodal learning? In *The Twelfth International Conference on Learning Representations*.

Anthropic. 2025. Claude 3.7 sonnet and claude code.

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Comput. Surv.*, 57(6).

Anna Bavaresco and Raquel Fernández. 2025. Experiential semantic information and brain alignment: Are multimodal models better than language models? *Preprint*, arXiv:2504.00942.

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, Online only. Association for Computational Linguistics.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

DG RTD, European Commission. 2020. Gendered innovations 2: how inclusive analysis contributes to research and innovation : policy review.

DeLisa Fairweather, Danielle J. Beetler, Nicolas Musigk, Bettina Heidecker, Melissa A. Lyle, Leslie T. Cooper, and Katelyn A. Bruno. 2023. Sex and gender differences in myocarditis and dilated cardiomyopathy: An update. *Frontiers in Cardiovascular Medicine*, 10.

Kevin Fiscella and Mechelle R. Sanders. 2016. Racial and ethnic disparities in the quality of health care. *Annual Review of Public Health*, 37(1):375–394.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:547–561.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Wasif Khan, Seowung Leem, Kyle B. See, Joshua K. Wong, Shaoting Zhang, and Ruogu Fang. 2025. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, pages 1–20.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. 2024. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *Preprint*, arXiv:2412.10372.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR.

Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Jesutofunmi A. Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. 2024. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine*, 30(4):1154–1165.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594.

Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, Shih-Jen Chen, Jiing-Feng Lirng, Kai-Wei Chang, and Shih-Hwa Chiou. 2025a. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1).

Xiang Li, Like Li, Yuchen Jiang, Hao Wang, Xinyu Qiao, Ting Feng, Hao Luo, and Yong Zhao. 2025b. Vision-language models in medical image analysis: From simple fusion to general large models. *Information Fusion*, 118:102995.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. 2024. Fairclip: Harnessing fairness in vision-language learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12289–12301.

Abhishek Mandal, Suzanne Little, and Susan Leavy. 2023. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 416–424, New York, NY, USA. Association for Computing Machinery.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Jordan P. Richardson, Cambray Smith, Susan Curtis, Sara Watson, Xuan Zhu, Barbara Barry, and Richard R. Sharp. 2021. Patient apprehensions about the use of artificial intelligence in healthcare. *npj Digital Medicine*, 4(1).

Gabriele Ruggeri and Debora Nozza. 2023. A multidimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada. Association for Computational Linguistics.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182.

Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, and Garrison W. Cottrell. 2024. Discovering and mitigating biases in clip-based image editing. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2972–2981.

Shetal Vohra-Gupta, Liana Petruzzi, Casey Jones, and Catherine Cubbin. 2022. An intersectional approach to understanding barriers to healthcare for women. *Journal of Community Health*, 48(1):89–98.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun-Hsing Yu. 2024. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siyuan Yan, Zhen Yu, Xuelin Zhang, Dwarikanath Mahapatra, Shekhar S. Chandra, Monika Janda, Peter Soyer, and Zongyuan Ge. 2023. Towards trustable skin cancer diagnosis via rewriting model's decision. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11568–11577. IEEE.

Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J Wang, Dushyant Sahani, and Shwetak Patel. 2024a. Demographic bias of expert-level vision-language foundation models in medical imaging. In *GenAI for Health: Potential, Trust and Policy Compliance*.

Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi. 2024b. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30(10):2838–2848.

A Image Feature Alteration Dataset

As described in Section 5, we conducted Experiment 1 using a modified version of the original image dataset in which controlled alterations were applied to evaluate model sensitivity to specific visual features. These alterations included brightness enhancement and blurring. To increase brightness, we clipped low pixel intensity values across the image volume. Specifically, all values below a fixed threshold, set at $v_{\max} = 1.5$ above the image minimum, were raised to that threshold. To introduce blurring, we applied Gaussian filtering using a two-dimensional convolutional kernel of size 9×9 and a standard deviation of $\sigma = 5$. An example of the corresponding alteration is given in Figure 7.

Figure 7: Experiment 1 image samples. Comparison of the brightened and blurred version of an image from CheXpert-5x200 used in Experiment 1, and produced with the procedure described in Appendix A.

B Biological Sex Dataset

As described in Section 6, the datasets used for Experiment 2 inject an increasing percentage of biases based on biological sex. More formally, given a target label l, a biological sex b, and a percentage p, our procedure increases the amount of instance in class l, having biological sex b, by $p\%$, while decreasing the number of instances in the opposite class having the opposite biological sex, by

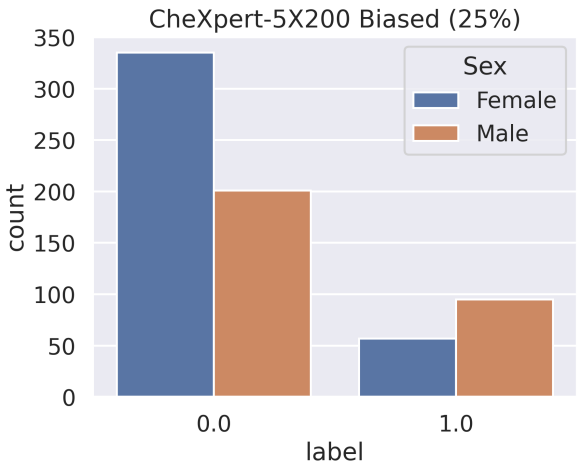


Figure 8: Experiment 2 dataset distribution. Visualisation of the biological sex distribution among the two classes in the dataset with 25% bias injection.

the same percentage p. To balance out the number of training and test instances with the baseline dataset, share across experiments, the instances are removed from, and placed in, the test set. in this work, we adopt $l = 1$, $b =$ male, and gather three dateset with $p = \{5, 15, 25\}$. As mentioned in Section 6, we do so to mimic distributions reported in the literature, showing how the selected label (i.e., cardiomegaly) (Fairweather et al., 2023). Figure 8 shows the training set obtained for $p = 25$.