Imperial College London

Imperial College London

Department of Computing

Knowledge Distillation and its Effect on Subgroup Disparities for Disease Prediction

Author:
Jan Marczak

Supervisor:
Prof. Ben Glocker

Co-Supervisor:
Dr Stefan Winzeck

Submitted in partial fulfillment of the requirements for the MSc degree in Advanced Computing of Imperial College London

September 10, 2023

Abstract

The profound impact of Artificial Intelligence (AI) on data analytics is undeniable, especially with large neural networks enhancing predictive performances. However, their deployment often faces challenges due to significant computing and storage requirements. Knowledge Distillation (KD) has emerged as a model compression solution, enabling the transfer of knowledge from these larger networks ('teachers') to smaller models ('students'). Yet, while KD's effectiveness in boosting performance is well-acknowledged, its implication for fairness has been overshadowed. Recent studies indicate that KD can amplify biases in language processing. Similarly, a growing body of evidence shows that such biases are present in medical imaging disease prediction models. These can encode protective characteristics (e.g. racial identity), potentially leading to imbalanced performance that could amplify existing health disparities.

Motivated by these findings and the recently established model inspection framework, we explore the interplay between KD and AI fairness within medical imaging. We assess various KD techniques and their impact on model fairness for different demographic groups. Our investigation reveals that, under the right conditions, KD can counteract inherent data biases, showing its potential in AI applications. Through extensive experiments, we showcase that this effect and its characteristics are influenced, among other factors, by the chosen KD method, model capacities, and dataset intricacies.

Acknowledgments

I would like to sincerely thank my project supervisors, Ben Glocker and Stefan Winzeck, for their consistent guidance and invaluable discussions. I'm also grateful to the BioMedIA Imperial research team for their warm welcome and supportive environment. Special thank you to Charlie Jones and Fabio De Sousa Ribeiro for their assistance with my experiments and insightful feedback.

Contents

Chapter 1

Introduction

1.1 Motivation

In recent times, Deep Learning (DL) has emerged as a powerful subfield of Artificial Intelligence (AI) and has quickly gained widespread attention within both academic and mainstream circles. The rapid expansion and availability of data, coupled with advancements in hardware technologies, have paved the way for training increasingly larger models. This revolution has served as the foundation for achievements in various domains like Computer Vision [1], Natural Language Processing (NLP) [2] and finance [3]. The healthcare sector is also expected to be profoundly impacted by AI [4], and has already seen a range of innovations such as anomaly detection [5] and harnessing electronic health records for predicting patient trajectories [6].

Although deep neural networks tend to achieve superior predictive performance, they come with their own set of problems. A major concern is the intense computational demands they bring. Setting up and deploying DL models requires extensive data, powerful machines, long training time and high storage capacity. Smaller-scale operations or real-time applications on devices with limited resources, such as mobile phones or Internet of Things (IoT) devices, may find it difficult to meet these requirements.

More critically, the lack of interpretability in DL models raises concerns regarding their trust and accountability. Unfair outcomes can arise when models embody the biases of the training data, and especially in the medical domain "black-box" nature of models can further amplify health disparities. For instance, studies have shown that medical imaging models showcase highly imbalanced predictive performance for various subgroups [7, 8]. Alarmingly, recent findings suggest these models can predict attributes like a patient's age, biological sex, and even racial identity solely from medical images such as chest X-rays [9, 10].

To address the resource intensity issue, numerous effective model compression techniques such as low-rank factorization [11] or model quantization [12] have been developed to lower the complexity of the models without sacrificing their performance. Knowledge Distillation (KD) [13], which is the focus of this work, is another notable method that has gathered significant attention in recent years. It transfers learned information from a larger, more capable teacher model to a smaller student model.

Additionally, in response to the opacity and potential biases of DL models, there has been a growing emphasis on the field of AI Fairness, which seeks to ensure that AI systems operate transparently and equitably. While a lot of research focused on identifying subgroup disparities through performance metrics [14, 7, 15], recent studies have concentrated on establishing the causes of such disparities. For instance, researchers in [16] created a framework to evaluate the connection between model's representation of patients' sensitive attributes and its disease predictions.

While the effectiveness of KD has been widely validated from the performance standpoint [17], there are still questions about the exact nature of knowledge the student model captures. Worryingly, recent work in [18] has shown that student models tend to be more gender biased in language processing after KD. Moreover, experiments in [19] have shown that beyond improving task performance, other potentially harmful properties can also get indirectly distilled during this process. Considering these findings, DL challenges and the potential of KD-based solutions for problems present in healthcare [20, 21], it is crucial to examine their impact on the performance of medical models before any widespread adoption. Our review suggests that there is no preceding research that considers knowledge distillation from this perspective.

1.2 Objectives

This project seeks to investigate KD through the lens of AI fairness. In particular, we analyse how different KD techniques affect the performance of medical imaging prediction models on distinct subgroups. We intend to use the recent framework from [16] to establish any patterns that might exist and appear between models. We are primarily interested in the relative comparison of students and their teachers by examining the differences in their predictions and in the way the knowledge is encoded, with a focus on health disparities. Whether the distillation process amplifies biases and introduces harmful "shortcuts" in the student model, compromising its safety, or if KD can enhance the model's fairness is what we wish to explore.

Consequently, our study is fundamentally research-intensive and experimental in nature. Our objective is not only to provide insights into KD but also to guide the future trajectory of research in this domain. The experimental design was iterative, with each subsequent research question set by the outcomes of the previous one. While we have ensured seamless integration of various KD techniques with the fairness framework, our primary focus has remained on the broader implications and findings rather than the ease of implementation.

1.3 Contributions

The primary contributions of our study are organised as follows:

- We examine the behaviour, performance, and disparities of medical imaging models post-KD in both the standard and capacity-gap settings. Our findings indicate that students often emulate their teachers, although they seem constrained by their capabilities, with outcomes differing based on the KD method, dataset traits and model capacities.

- We utilise a newly established framework for unsupervised model inspection [16] to showcase on a deeper level the differences in overall and subgroup behaviour between KD methods. Furthermore, we illustrate the challenges and inconsistencies that arise when comparing performance disparities with unsupervised analysis.
- Contrary to prior findings that KD amplifies bias, we show that student models, when trained appropriately by a fairer teacher, can leverage KD to overcome bias originally present in the standalone training. This highlights KD's potential for practical applications.
- We emphasised the importance of understanding utilised medical datasets during model evaluation and deployment, showing that distinct data characteristics can significantly influence outcomes of AI fairness research.

1.4 Outline

This report begins with a review of relevant academic literature, focusing on KD (section 2.1), AI Fairness (section 2.2), and a combination of both topics (section 2.3). Subsequently, Chapter 3 explains our design choices and preliminary testing. Within this section, we describe the experimental framework that was consistently employed when addressing our primary research questions in Chapter 4. Each study is supported by experiments and presented in its entirety, from setup to results discussion. Our findings are summarised in Chapter 5, where we revisit all research questions and discuss their broader implications, suggesting potential avenues for future research.

Chapter 2

Background

This chapter serves as a review of existing literature and methodologies employed within the scope of this project. It starts by introducing the concept of KD and explaining its different knowledge types and distillation schemes. Next, the discourse shifts towards an exploration of AI fairness, together with its distinct evaluation techniques and findings, with a specific emphasis on medical imaging. Lastly, the chapter delves into the existing body of work that intersects these two domains.

2.1 Knowledge Distillation

Knowledge distillation is one of several model compression techniques in ML and involves training a smaller student network under the supervision of a larger teacher model. The seminal work of [22] introduced this notion, wherein complex ensembles of models were leveraged to annotate an extensive unlabeled dataset. This annotation process helped in training a model that is both fast and compact, which was able to maintain comparable performance levels.

This idea was later revisited and formally popularised in [13], which falls under the category of response-based KD. In general, as summarised in [23], a KD system consists of three main elements: knowledge, the distillation algorithm, and the teacher-student architecture. The following provides a review of these concepts and discusses the state of the current KD research. For a more comprehensive analysis and comparison between multiple methods, we direct the reader to [23].

2.1.1 Knowledge Types

Knowledge is a key ingredient in the KD framework and characterises the type of information transferred between models. In this section, we explain the three most popular forms of knowledge: response-based, feature-based and relation-based, providing examples of different implementations.

Response-Based

In response-based systems, the student model learns to mimic the behaviour of the teacher model by matching its output logits. These are typically transformed into class probabilities using the softmax function. However, the original softmax often gives the highest-rated class an overwhelming likelihood, which makes the chances of other classes almost negligible, hence reducing insights about them. Because of that in the original and most popular implementation of response-based KD [13], authors introduce the concept of *soft targets*. Specifically, for a given class i, the model's output logits z_i are transformed into the class probability p_i, using a temperature-controlled softmax function:

$$p_i(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{2.1}$$

The temperature parameter T adjusts the softness of the probability distribution. As shown in Figure 2.1, for hypothetical logit scores in $1-5$ range, when $T = 1$, we get a standard softmax function. If $T > 1$ the probability distribution becomes softer, giving additional insight into the similarity between predicted and other classes according to the teacher model.

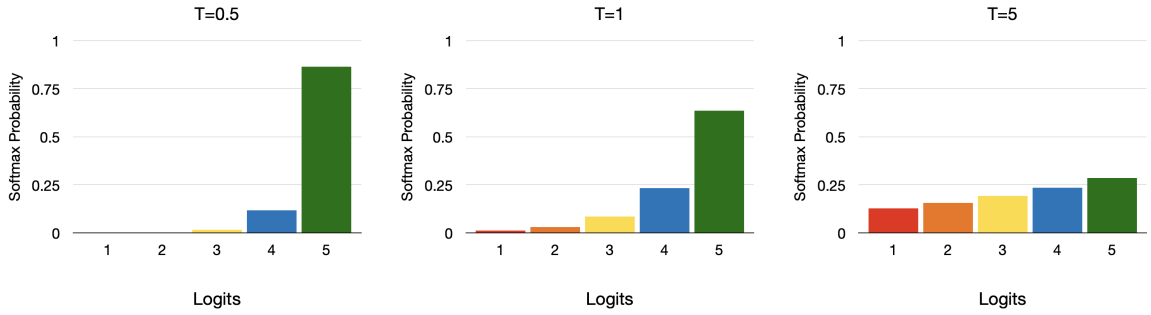

Figure 2.1: Temperature-controlled softmax outputs given an example set of logits.

Consequently, the so-called *distillation loss* in response-based KD system is concerned with minimizing the divergence metric between the logits of the teacher and student produced by Equation 2.1. Both models commonly employ the same T, usually being greater than 1. This usually refers to the Kullback-Leibler (KL) Divergence loss L_{KL}. Additionally, for labelled datasets, [13] suggests using the ground truth for training the student model. This is referred to as *student loss*, which is defined as a standard cross entropy loss L_{CE} between logits (produced with $T = 1$) and labels. The overall response-based KD loss function is a linear combination of both student and distillation losses:

$$\begin{aligned} L_{ResKD} &= \alpha * L_{CE}(y, p(z_s, 1)) + (1-\alpha) * L_{KL}(p(z_s, T), p(z_t, T)) \\ &= \alpha * \left(-\sum_{c=1}^{C} y_c \log(p_c(z_s, 1)) \right) + (1-\alpha) * \left(\sum_c p_c(z_s, T) \log \left(\frac{p_c(z_s, T)}{p_c(z_t, T)} \right) \right) \end{aligned} \tag{2.2}$$

where C is the number of classes, s and t indicate student and teacher networks respectively, y is the ground truth, p is the temperature-controlled softmax from Equation 2.1, and α is the hyperparameter of the linear combination between losses. As explained in [13], it is generally

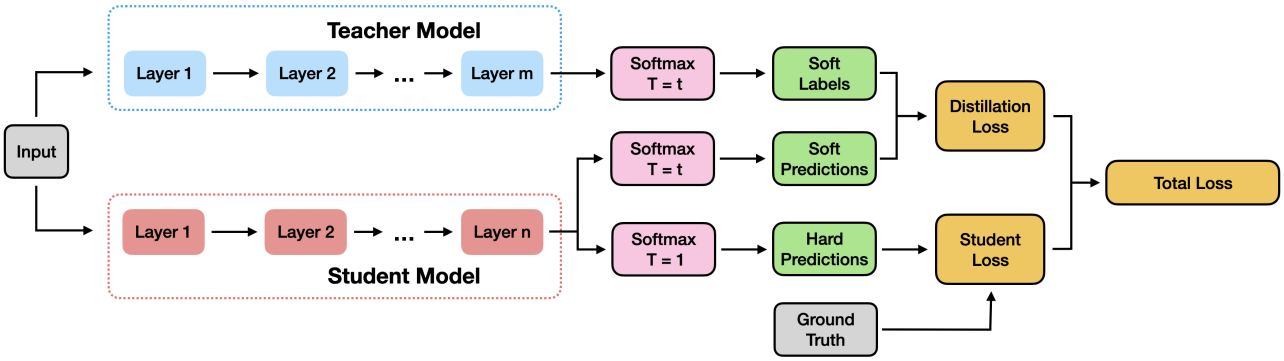

Figure 2.2: A typical response-based KD framework as described in [13]

best to use considerably low α, subsequently putting more emphasis on the distilled knowledge. The overall framework for response-based KD is present in Figure 2.2

Feature-Based

After the original work by [13], subsequent research has explored different distillation extensions aimed at leveraging internal feature representations. In feature-based KD, in addition to the logits, the output of the hidden layers (i.e. *feature-maps*) of the teacher are used for the supervision of the student.

The idea originated in *FitNets* [24], where the intermediate outputs, referred to as *hints*, from the deep and wide teacher serve as targets for training the hidden layers of an even deeper, yet thinner, student (termed *guided layers*). This approach is complemented by the standard response-based KD, leading to models that, as the authors argue, offer improved generalization. Given that the teacher model might be wider, *regressors* are introduced after the guided layers. These regressors compensate for size discrepancies between the teacher and student models, and their parameters are trained in tandem with those of the primary network. They can be designed as simple convolutional modules that align the feature maps, as demonstrated in [24].

The success of FitNets has led to the development of several alternative approaches that align features in unique ways. One notable method was introduced in [25] called ***attention-transfer***. This mechanism compares the activation patterns of CNN layers between models. The underlying assumption is that how strongly a neuron reacts can show its importance to a particular input. By examining these reactions, an attention map is created which highlights important areas of the input data. The authors propose adding absolute values of feature maps over their channel dimensions and raising them to a power to underscore regions of high significance.

Contrary to the attention-transfer approach, the study presented in [26] introduces a method that imitates the teacher's decision boundaries by exclusively matching the activation sign of neuron responses while entirely omitting its magnitude. Other notable technique includes the integration of auxiliary networks [27], termed as the *paraphraser* and the *translator*. The paraphraser is trained autonomously to distil salient features from the teacher. Subsequently, these representations are harnessed to train the translator, which is responsible for helping the student understand and adapt the teacher's knowledge.

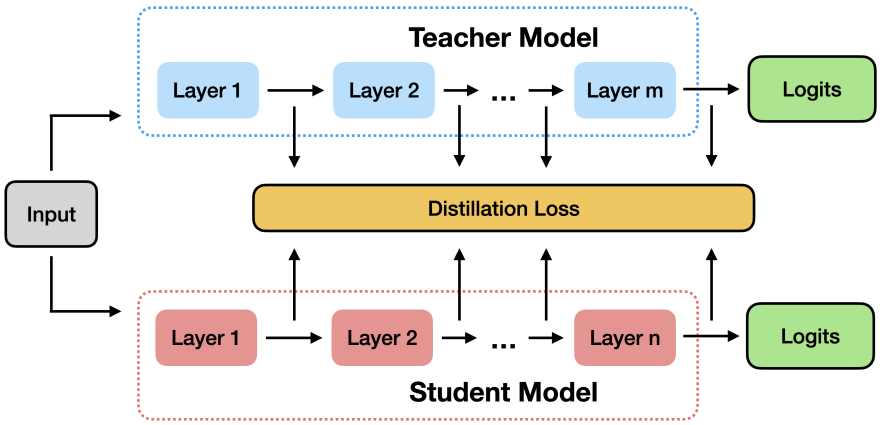

Figure 2.3: A typical feature-based KD framework.

Typically, the formulation of the distillation loss for feature-based KD can be written as:

$$L_{Fea} = L_S(\Phi_s(F_s), \Phi_t(F_t)) \tag{2.3}$$

where F_s and F_t are the hidden layers' feature maps of student and teacher respectively, whereas the function $\Phi(.)$ transforms them to a particular representation through regressors. Subsequently, they are matched by the similarity function $L_S(.)$ (most often l_2-norm distance or Mean Squared Error).

It is generally common to employ response-based training alongside feature-based for better distillation. This can be done by adding and weighting L_{Fea} to the loss function from Equation 2.2 as employed in [28, 29]. Alternatively, [24] use a multiple-stage training process, where the feature-based and response-based distillations are disentangled. A generic feature-based model is shown in Figure 2.3.

Relation-Based

In relation-based KD, the student is trained to mimic not only the final prediction but also the relationships between feature maps or input data from the teacher.

For example, [30] proposed a method to capture the relationships between feature maps in a neural network, through the Flow of Solution Process (FSP) matrix. It is formed by taking the inner product of feature maps from two different layers. During training, the student model aims to minimise the l_2-distance between its and the teacher's FSP matrix. More recent developments in relation-based KD have captured the knowledge by considering the relationships among different input points. For instance, [31] introduces a locality-preserving loss. It ensures that if two input points are close or have a specific correlation in the teacher network's high-dimensional feature space, they retain that relationship in the student network's low-dimensional feature representation.

We can distinguish two different cases for relation-based loss function [23]. First is concerned with the pair of feature maps (F_s^i, F_s^j), (F_t^i, F_t^j) that go through a relationship function $\Psi(.)$

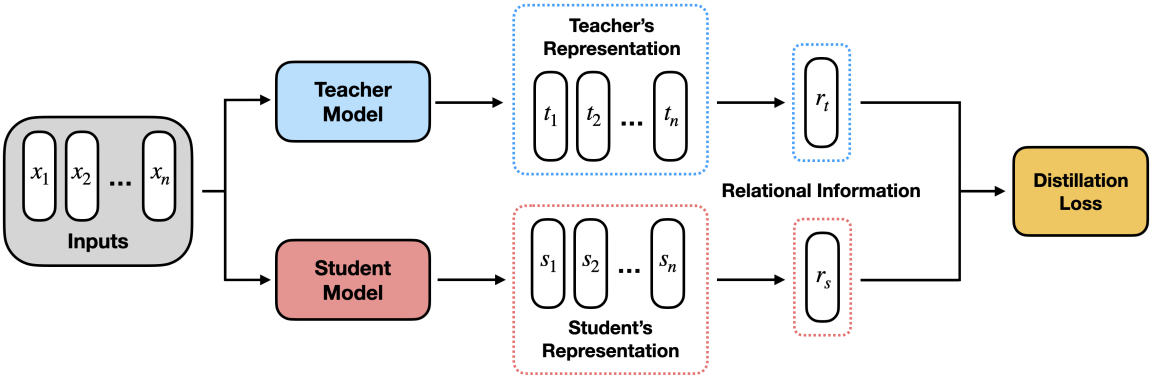

Figure 2.4: A typical instance relation-based KD framework.

(e.g. generation of FSP matrix [30]) and then a correlation function $L_C(.)$:

$$L_{Rel}(F_s, F_t) = L_C(\Psi_s(F_s^i, F_s^j), \Psi_t(F_t^i, F_t^j)) \tag{2.4}$$

In the case of relation knowledge based on relations between inputs we can write:

$$L_{Rel}(R_s, R_t) = L_C(\psi_s(s_i, s_j), \psi_t(t_i, t_j)) \tag{2.5}$$

where $(s_i, s_j) \in R_s$, $(t_i, t_j) \in R_t$ and R_s, R_t are sets of input representations. $\psi(.)$ and $L_C(.)$ are similarity and correlation functions respectively. For a visual representation of relation-based KD refer to Figure 2.4

2.1.2 Distillation Methods

A second vital element of KD is the mechanism in which the knowledge gets transferred between the models, i.e. the distillation scheme. This section discusses three of the most commonly used learning methods, which are summarised in Figure 2.5

Offline

In offline distillation, we assume that the teacher network is pre-trained and then frozen. It means that while the student network gets trained, the teacher model is not updated and is only used to provide knowledge. This follows the setup described in the original work [13]. The majority of research on offline KD has primarily centred around altering the type of knowledge utilised and the associated distillation loss. However, some studies have specifically focused on the configuration of the teacher and student networks. While the multi-teacher approach was used already in the original paper [13] through ensembles, [32] proposed an alternative method, where for each batch a random teacher is selected from a pool of candidates. According to the authors, adopting this method would allow the student to perceive the input data from various perspectives, leading to better generalisation. Researchers in [33] took a different approach after finding that "student network performance degrades when the gap between student and the teacher is large". To alleviate the issue, they propose incorporating "assistant teachers",

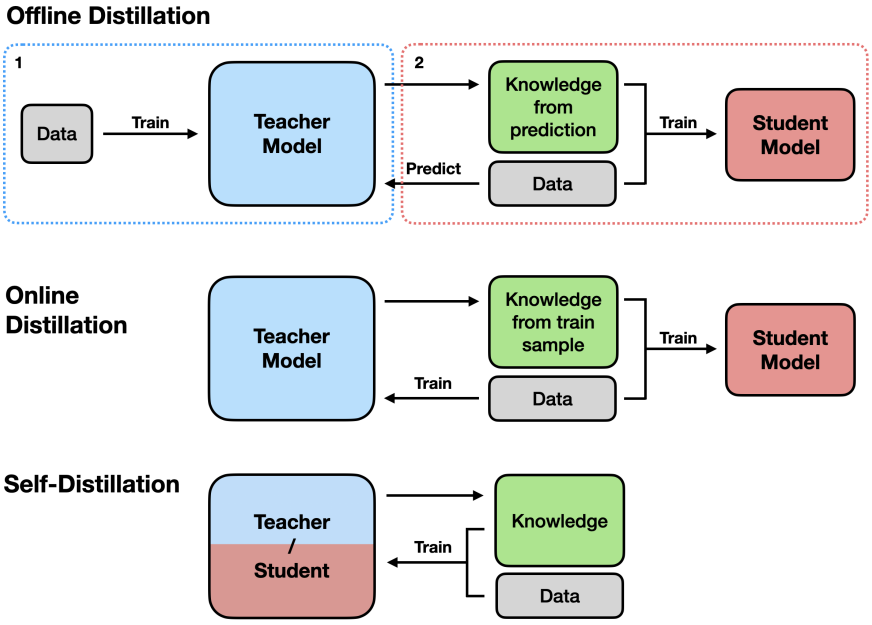

Figure 2.5: Graphical representation of most common distillation schemes.

which are intermediate-sized networks. These assistant teachers serve as the means to bridge the gap between the networks, establishing a seamless chain of KD across multiple models.

Generally, offline KD suffers from the two-phase training procedure, i.e. there is no way to avoid training a complex, highly capable teacher model, which is often a time-consuming process. Furthermore, the student is heavily dependent on the teacher and the capacity gap (see subsection 2.1.3) between them can hurt the performance [33]. Regardless, the simplicity of offline KD and the flexibility in picking a teacher model make it an attractive choice.

Online

In an online distillation setting, both the teacher and the student models are trained concurrently, enabling end-to-end training of the entire network. This helps to overcome the main limitation of offline KD: the training of the static teacher model. Arguably the first online KD was proposed in Deep Mutual Learning [34], wherein a group of student networks, all initialised with different starting conditions, learn together and teach each other. During training, each network not only aims to predict the correct label but also matches the probability estimates of other networks. Their experimental results suggest that training models in this way outperforms traditional offline distillation.

Similar findings were reported in [35], where researchers further explored the idea of online KD by introducing an attention-based mechanism (inspired by the Transformer model [36]). In their work, they propose a two-level distillation process with multiple auxiliary networks and one group leader. Initially, each auxiliary model generates its training targets through the attention mechanism that weights and combines the predictions of other auxiliary networks. Then their logits are averaged and serve as knowledge passed to the teacher.

The main advantage of online distillation is that it is end-to-end trainable and provides an

opportunity to utilise parallel computing. However, the problem of training a high-capacity teacher or a whole group of networks still exists. Hence, the introduction of self-distillation.

Self-Distillation

Self-distillation can be viewed as a special variant of online distillation wherein the teacher and student models employ the same network architecture [23]. This concept was most likely introduced as Born-Again Networks (BANs) [37]. At first, a teacher model is trained from the ground up and is later used to supervise a new identical model initialised with a different random seed. After the convergence of this model, it becomes the teacher for later generations, and the cycle continues. Ultimately, the final network can either be an individual or an ensemble of models learned throughout training. Although the authors report superior performance, upon conducting further experiments, researchers in [38] claim that training and ensemble independently from scratch remains a more effective approach.

A different variation of self-distillation was discussed in [39, 40], where only a single model is trained during the whole process. In the case of snapshot-distillation [39], instead of using previous models, information from earlier epochs in the same generation is used as knowledge during distillation. A more complex approach was proposed in [40], where the deeper layers of the network provide the knowledge to supervise the shallower ones. The model is partitioned into sections, each acting as a classifiers with extra fully connected layers. These intermediate classifiers are trained using labels, distillation from deeper layers, and hints (feature maps from deeper sections). In predictions, one can choose individual section-classifiers or ensembles based on speed and accuracy needs.

2.1.3 Capacity Gap Issue

The concept of the *capacity gap* arises when the capacity of a student model is significantly smaller than that of the teacher model. Here, "capacity" refers to the number of parameters a network possesses. Such a disparity can hurt the student model's ability to effectively replicate the performance of the teacher. For instance, [33] illustrated that as the teacher's capacity is progressively increased, the student's performance initially improves before starting to worsen. This poses a nuanced challenge in striking the right balance between model resemblance and the benefits of KD, especially when aiming to deploy compact models that offer comparable performance.

In the domain of KD research, this issue has gathered significant attention, as evidenced by works in [41, 23], with numerous methodologies developed to specifically address this challenge [33, 42]. Notably, some of the early feature-based KD techniques were also centred on this problem. While FitNets [24] aimed to train deeper student models, the attention-transfer method [25] was created to facilitate the learning of shallower networks, thereby assisting in mitigating the capacity gap issue.

2.2 AI Fairness

Fairness in AI is a second vital component of this project. In an ML context, it emphasises the need for establishing trustworthiness and mitigating any potential discriminatory behaviours exhibited by DL models [43]. The significance of fairness becomes evident in the field of healthcare, which is witnessing a growing adaptation and clinical trials of ML solutions [44, 45]. Hence, it is not surprising that more research has been focused on analysing and ensuring fairness in the medical field, given that biased behaviours of models can have detrimental consequences in these instances.

In this section, we give an overview of AI fairness topics, providing essential definitions, the exploration of unfairness sources, methods for enforcing fairness as well as evaluating it. Given the focus of the project on medical imaging, we place greater emphasis on discussing relevant research and advancements in this particular domain.

2.2.1 Definition

Fairness in statistics is an emerging field that aims to reduce both deliberate and unintentional discrimination against protected subgroups through quantifiable metrics [46]. This can be defined in two ways. First, as *group fairness* or *statistical parity*, which calls for consistent outcomes across different protected groups [47]. Second, as *individual fairness*, which asserts that similarly situated individuals with respect to a particular task should be treated alike [47]. We define group fairness as in [48]:

Consider an ML setting with Dataset $\mathcal{D}$ containing N samples denoted as $d_1, d_2, ..., d_N$. Each data point d_i consists of task-relevant medical image X_i, a ground truth label Y_i and a set of task-irrelevant *sensitive attributes* $A_i = A^1, A^2, ..., A^L$ (e.g. race, gender, age etc.). Continuous variables like age are most commonly discretised. Each sensitive attribute A^l can have V_l possible values, i.e. $A^l = a_1, a_2, ..., a_v$. Hence, a dataset D can be split into $G = \prod_{l=1}^{L} V_l$ groups, with each group having N_g samples ($\sum_{g=1}^{G} N_g = N$). Given a typical model f, its output $\hat{Y}$ and a distance criterion M we have:

$$f(X_i) = \hat{Y} \tag{2.6}$$

$$D_i = M(Y_i, \hat{Y}_i) \tag{2.7}$$

To achieve absolute fairness, we would like the average performance of each group

$$D^g = \frac{1}{N_g} \sum_{i=1}^{N_g} D_i, \quad g \in [1, G] \tag{2.8}$$

to be the same for all groups (i.e. $D^1 = D^2 = ...D^G$). This is also usually referred to as *intersectional fairness* and measures biases across groups formed with combined attributes (e.g. young white male). However, this approach presents notable challenges. The sheer volume of potential intersectional groups can result in data sparsity, and their disparate sizes might yield

skewed outcomes [49]. Furthermore, as the number of groups increases exponentially with each sensitive attribute, their interpretation and comparative analysis become increasingly complex. Consequently, fairness research often focuses on single-attribute fairness (e.g., Male vs. Female, White vs. Black). Although this method has its own set of limitations, it allows for a clearer and more direct analysis.

It is important to highlight that sensitive attributes can sometimes correlate with specific objectives, such as cancer risk increasing with age. However, during training, models may prioritise straightforward features as noted by [50], leading to *shortcut learning*. This can result in biases when relying on sensitive attributes for predictions.

2.2.2 Unfairness Origins in ML

Unfairness in ML primarily stems from data issues. For example, many medical datasets are imbalanced in labels and skewed across subgroups, causing predictive model disparities [51, 52]. Data annotation and collection can also introduce bias. Doctors may label the same patients differently, and the use of different medical equipment can result in unwanted artefacts on the scans [53]. Algorithmic components can further bias outcomes. Metrics for model comparison may skew final model selection [54, 18], especially since the same model can show consistent overall performance but vary significantly in subgroup results [54]. It is also worth noting that human interpretation of model fairness evaluation can also introduce biases [55].

2.2.3 Fairness Metrics

Measuring fairness is challenging, partly because of its many dimensions, the multitude of different metrics, and their context-sensitive nature [56]. This can result in different outcomes that may be hard to interpret. Nonetheless, it is common to employ fairness metrics that capture the extent of impartial treatment among different subgroups. Relevant to our study and most common within fairness research, we define them with binary sensitive attributes. For multi-value cases, we refer the reader to [57].

Firstly, there exist several criterion that can be imposed as constraints or incorporated into a loss function. For instance *Demographic Parity* (DP) ensures predictions $\hat{Y}$ and a binary sensitive attribute A are independent [58]:

$$P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|A = a), \tag{2.9}$$

whereas *Equality of Odds* [59] requires that $\hat{Y}$ is independent of A conditional on the true outcome Y. This means that for any outcome Y, the probability $P(\hat{Y} = \hat{y}|Y, A = a)$ should be the same across all values of the sensitive attribute A:

$$P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|A = a, Y = y) \ \ \forall y, a. \tag{2.10}$$

This way, the Equality of Odds criterion ensures consistent True Positive Rates (TPR) and False Positive Rates (FPR) across all groups defined by the sensitive attribute.

For a binary sensitive attribute A, it is common to obtain an overall measurement of unfairness

by comparing the criterion M for subgroups, such as by calculating their absolute difference:

$$M_A^{GAP} = |M_{a_1} - M_{a_2}| \tag{2.11}$$

The metric is often denoted as $Metric^{GAP}$, such as F1^{GAP}. While demographic parity and equality of odds are popular fairness metrics, many researchers focus on subgroup performance disparities. For instance, in medical diagnostics, [7] used TPR^{GAP} for Chest X-ray model fairness. Other metrics include Youden's J statistic differences [16], F1 score [15], and area under the receiver operating characteristic curve (AUC) [60, 16]. Notably, AUC^{GAP} is prevalent due to its data imbalance independence, ensuring consistent evaluations [61].

Given the multifaceted nature of fairness, it is not always straightforward to integrate it seamlessly into models without some trade-offs. According to recent research in [62], there is evidence to suggest that as fairness in a model increases, there can be a corresponding decrease in the model's overall performance. A similar trend has been noted in [63], where efforts to enhance fairness can inadvertently lead to a reduction in overall welfare. It is also important to be aware of a phenomenon called *Fairness Gerrymandering* [64], where ensuring fairness for individual attributes or intersectional groups in isolation can lead to unfair outcomes when multiple attributes or groups are considered together.

In summary, while traditional fairness metrics offer limited insight into model behaviour, they serve as an initial step to identify potential biases. As stressed by [62] for a comprehensive understanding of fairness, a nuanced examination of DL models is required.

2.2.4 Fairness Enforcing Training

Various methods have been introduced to enforce fairness in ML models during their training. For instance, researchers in [65] introduced a *domain-independent training* method. Here, individual classifiers are trained for each domain (e.g., colour/grayscale), all sharing a common backbone. During testing, predictions are aggregated from the decision boundaries of all domain-specific classifiers, promoting unbiased outcomes. Similarly, in *adversarial training* [66], an adversary classifier is trained to predict sensitive attributes from the primary classifier's features. The primary model is then trained to both make accurate predictions and deceive the adversary, ensuring predictions do not reveal sensitive attributes.

Another approach involves adding fairness-related penalty terms to the loss function [67]. However, this might introduce non-convexities to the overall loss function, complicating training and hyperparameter tuning. To account for this, the optimisation problem can be formulated with the use of the Lagrangian Multiplier method to incorporate the constraints within a single overall loss. The goal is to minimise the primary loss L_{NN} subject to the fairness loss L_F being less than some predefined value ϵ, i.e.

$$\text{Minimize } L_{NN}(\theta) \text{ subject to } L_F(\theta) \leq \epsilon \tag{2.12}$$

Given this, the Lagrangian can be written as:

$$L(\theta, \lambda) = L_{NN}(\theta) + \lambda(L_F(\theta) - \epsilon) \tag{2.13}$$

where λ is the Lagrangian multiplier and θ are model's parameters. Then, the overall optimisation problem is carried by performing gradient descent w.r.t θ and gradient ascent w.r.t λ at every iteration, i.e.:

$$min_\theta max_\lambda L(\theta, \lambda) \tag{2.14}$$

This approach was implemented in [68], with L_F formulated as Demographic Parity, Equalized Odds among other fairness metrics. Work in [69] uses a similar idea and provides additional fairness constraints. Notably, their results show that the Equal Loss metric concerned with minimising the difference between original loss values calculated on distinct subgroups showed the smallest AUC^{GAP} between Male and Female patients. For a specific binary sensitive attribute A, it can be formulated as:

$$L_{F_{EQ}}(\theta, A) = |L_{NN}^{a_1}(\theta) - L_{NN}^{a_2}(\theta)| \tag{2.15}$$

However, it's essential to approach these methods with caution. Empirical research has shown that these techniques often underperform across all groups in comparison to simple baselines [70]. Results presented in the MEDFAIR framework, dedicated to benchmarking fairness in medical imaging, further support this observation [71]. Thus, carelessly applying these methods may not always yield fairer models.

2.2.5 Bias Inspection Methods

To gain deeper insight into the model's predictions, researchers often use methods that necessitate a more manual, often visual and thorough evaluation, which might not always be quantifiable as a single metric. In this section, we highlight different techniques for model inspection.

Transfer Learning

Transfer learning can be used to see if sensitive attributes are encoded and used during inference time [10]. This involves training a disease detection model, freezing its parameters, and then retraining it with a new classification layer focused on predicting a sensitive attribute like race. Because the main parameters of the model are frozen during retraining, high performance on the secondary objective might suggest a correlation between the two tasks. This approach was later formalised in [16] as supervised prediction layer information test (SPLIT), but it is also argued to be insufficient when used alone for determining causal relationships.

Multitask Learning

Multitask learning evaluates how interconnected different tasks are by simultaneously training a model on several prediction objectives [72]. A shared neural network backbone is typically employed with distinct prediction layers for various attributes, like disease, biological sex, and gender. If there is a correlation between a patient's protected attributes and disease prediction, these features might align similarly in the feature space. Comparing the feature representation

of a multitask model to a single-task model can provide insights into the relationship between tasks, as demonstrated in [16].

Unsupervised Exploration of Feature Representations

Unsupervised exploration of feature representations, as used in [16], employs unsupervised ML techniques to inspect the information encoded in a model's learned features, particularly in relation to its primary task. Given the high dimensionality of these representations, evident in models like ResNet18 with 512 dimensions, dimensionality reduction techniques are essential. Principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) are the primary methods used. PCA identifies the principal components that capture the most significant variations in the feature space, ordered by the amount of explained variance. Conversely, t-SNE projects data into a lower dimensional space, striving to maintain the pairwise distances between data points, thereby preserving the inherent clustering in the original space. Both methods facilitate the creation of two-dimensional scatter plots. When overlaid with additional patient information, these plots serve as an effective tool for manual bias inspection. As noted in [16], if we observe significant differences between marginal distributions for subgroups in PCA modes or t-SNE dimensions, it can suggest the presence of biases in the model's features.

Statistical Analysis

Statistical methods can enhance model evaluations by quantifying findings present in visual bias analysis. For example, the Kolmogorov-Smirnov (KS) test compares two sample distributions by measuring the maximum distance between their cumulative curves. In unsupervised feature analysis, the KS test can be used to assess whether the marginal distributions of two subgroups, when projected into PCA/logit space, are the same [16]. The null hypothesis assumes that both distributions are identical. A low p-value from the KS test indicates a significant difference between the distributions, leading to the rejection of the null hypothesis. Overall, statistical tests are most effective when combined with other analytical methods and sole reliance on them might not be too comprehensive [16].

Pixel Attributions

In image classification, saliency maps are commonly used to identify pixels influencing model predictions. These maps, often derived from the gradient of predictions concerning the input image [73], highlight pixel importance, which can help in detecting potential bias. A related approach in [10] hid high-contribution and subgroup-specific patches after training and retrained the model. By comparing the performance of both models, it can reveal subgroup disparities. However, recent studies [74, 55] have raised concerns about the reliability of saliency maps and caution their use.

2.2.6 Bias Inspection in Medical Imaging Models

The methods explained in subsection 2.2.5 were utilised in recent literature to establish potential biases and subgroup disparities in medical imaging models.

For example, in the study by [10], researchers employed transfer learning techniques and found that DL models could identify a patient's race from medical scans with remarkably high accuracy, achieving AUC scores exceeding 0.80. This finding underscores the ability of AI models to implicitly learn to identify sensitive attributes. It also raises concerns about the implications of using such information in their predictions, especially when medical professionals lack this ability.

The work presented in [16] builds upon the results from [10] and introduces a new framework for subgroup analysis in medical imaging disease detection models. The researchers utilise transfer learning, multitask learning, unsupervised methods, and statistical analysis. They discovered that subgroup disparities exist in terms of shifted TPR (True Positive Rate) and FPR (False Positive Rate) in chest X-ray disease detection models. In their subsequent study [75], they employed unsupervised methods alongside statistical tests to detect distribution shifts across sex and race in the recently published chest radiography foundation model [76]. Furthermore, they compared this to the baseline model trained on CheXpert data [77]. Their findings revealed that the foundation model exhibits statistically significant differences in ten out of twelve pairwise comparisons across protected attributes, in contrast to only four in the baseline model. This suggests that the published model might amplify subgroup disparities, making it potentially unsafe for clinical applications.

In [78], researchers used SmoothGrad saliency maps [79] to analyse models predicting biological sex from brain MRI scans. They found that the saliency maps varied across subgroups, with brain areas related to pubertal development being key predictors. This might explain performance differences among subgroups. Similarly, [80] noted that face recognition models emphasised different facial areas based on gender and skin tone.

2.2.7 Subgroup Separability

An important aspect which is often overlooked when looking at subgroup disparities and fairness is *subgroup separability*. It has been defined in [81] as "the ease with which individuals can be identified as subgroup members". The authors shed light on how different sensitive attributes and medical imaging modalities influence the ability to separate the subgroups in DL models.

They discovered that subgroup separability varies vastly based on these aspects. They tested eleven medical dataset-attribute combinations and found that all protected characteristics can be predicted from chest X-ray scans with an AUC score of 0.9+. In contrast, Skin Dermatology and Fundus datasets were considerably lower, hovering at around 0.75 AUC. This indicates that chest X-ray scans convey more attribute-dependent information than skin or fundus images. The authors emphasise that when subgroup separability is high, models tend to leverage sensitive data, leading to pronounced disparities among subgroups. Conversely, with low separability, models perform uniformly across groups, suggesting that traditional group fairness metrics might not always capture bias effectively.

2.3 Fairness of Knowledge Distillation

In the context of KD, it is intriguing how a simpler student model can match or even surpass the performance of its bigger teacher model. The question arises: Is this achievement purely based on the guidance of the teacher, or does the student model resort to shortcuts and unwanted bias? In what ways do the two models differ?

Some degree of attention has been given by previous research in attempting to tackle these questions. For instance, work in [82] presents evidence indicating that improvement in overall accuracy in KD comes at the cost of reduced subgroup accuracy, especially when trained with imbalanced data. Additionally, their results suggest that student networks amplify teachers' mistakes and perform worse on classes already found difficult by the original model. To mitigate it, they suggest reducing the reliance on the teacher for such classes by introducing per class mixing weights $\alpha_1, \alpha_2, ...\alpha_c$ that replace a static α present in Equation 2.2.

The exploration of fairness within KD methods has garnered attention in the NLP domain. [18] showed that the constrained capacity of the student model, coupled with its cross-entropy loss function, can lead to amplified gender bias in DistilBert [83]. A parallel sentiment was echoed in [84], highlighting the harmful effects of KD on the fairness of language models across various scenarios. Contrary, recent studies such as [85] position KD as a mechanism to mitigate detrimental biases. While the insights from the NLP domain may not be directly applicable to this project, they underscore the potential for bias amplification in KD and highlight the nuanced and sometimes conflicting nature of fairness outcomes.

A more empirical analysis was done in [19], where authors discovered that attributes such as adversarial vulnerability, data invariance, and saliency maps are transferred from the teacher model to the student model. Notably, they explored a case where a teacher model showed significant subgroup biases, and a student model trained independently on balanced data was fair. When the same student model was trained through KD from an unfair teacher (but still with balanced data), it inherited the teacher model's biases. While their discoveries offer interesting insights for our study, the limitations of their experimental design compel us to pursue a more exhaustive examination.

Chapter 3

Methodology and Setup

This chapter outlines the project's methodologies and design decisions. Given its experimental nature, we also describe the process of finding a robust and consistent setup used when attempting to answer our main research questions in Chapter 4.

3.1 Fairness and Metrics

In our work, we adopt the notion of a **single-attribute fairness** rather than group-fairness described in subsection 2.2.1. We believe that intersectional fairness presents multiple issues, including small group sizes, data sparsity and challenges in interpretation. It is important to note that our research investigates primarily how fairness behaviour evolves post-KD and does not aim to pinpoint specific biases in medical models. To achieve this, we assess performance disparities across subgroups for each sensitive attribute individually.

We focus on binary classification, such as disease detection, and use binary categorisation for sensitive attributes, like biological sex distinctions. This aligns with the main trends in AI fairness research. For instance, many widely accepted fairness metrics, like equalized odds, are designed specifically for binary setup. Moreover, adopting this method simplifies group comparisons and makes it easier to evaluate impacts on specific subgroups.

Regarding **performance metrics**, we mainly use AUC because of its binary nature and its prevalence in prior medical imaging studies [16, 81, 10]. AUC is especially effective for imbalanced datasets as it is not tied to a specific threshold, unlike metrics like F1 or accuracy. While we have occasionally used Youden's J statistic (TPR - FPR) and Equalised Odds, we focus on AUC in this report for consistency.

Accordingly, for **fairness metrics**, we mainly report AUC^{GAP} for each sensitive group. Specifically, we measure the relative AUC^{GAP}, which is normalised and scaled accordingly:

$$AUC_A^{GAP} = \frac{|AUC_{a_1} - AUC_{a_2}|}{AUC} \times 100 \tag{3.1}$$

with normalisation allowing for consistent fairness comparisons across different models.

3.2 Bias Inspection

For a deeper understanding of the model's behaviour and potential biases, we use inspection methods outlined in subsection 2.2.5. Our main goal is to compare the representations of the teacher model and its students. Therefore, multitask and transfer learning are not ideal due to their multiple training phases, making model comparisons complex. Their outcomes are also challenging to measure in the KD context. We also avoid saliency maps, seeing them more as tools for explainable AI (XAI) that might not reveal bias effectively. Recent critiques further support our decision to skip them [74, 55].

Similarly to [75], we primarily focus on the unsupervised exploration of internal feature representations and statistical analysis for model comparison. Their nature enables us to compare features both visually and quantitatively using marginal distributions.

Precisely, we begin by extracting the relevant features from the entire test set, processing each scan using the backbone of our trained model of interest. Subsequently, we employ both PCA and t-SNE techniques to reduce the dimensionality of these features, which we visualise on 2D scatter plots. In cases where the test set is too big for clarity, we display only a random subset. We aim to determine if PCA modes that differentiate samples based on disease also distinguish them based on sensitive attributes. A similar approach is applied to t-SNE. Should we observe disparities within subgroups in either PCA or t-SNE projections, it could suggest underlying bias.

To numerically confirm our visual data, we check if the distributions of specific subgroups align across the initial four PCA modes. We use KS-tests and adjust p-values for multiple tests with the Benjamini-Yekutieli method, setting a significance level at 95%. Additionally, to measure the distance between these distributions, we compute the kernel density estimation (KDE) for selected groups, normalise them, and then determine the Jensen-Shannon (JS) distance. For two probability distributions P and Q, JS distance is defined as the square root of JS divergence:

$$\begin{aligned} \mathrm{JSD}(P, Q) &= \sqrt{\mathrm{JS}(P, Q)} \\ &= \sqrt{\frac{D_{KL}(P, \mu) + D_{KL}(Q, \mu)}{2}} \end{aligned} \tag{3.2}$$

where μ is the pointwise mean of P and Q and D_{KL} is the KL divergence that can be expanded as in Equation 2.2. A score of 0 shows perfect overlap, while 1 means they're entirely distinct.

3.3 Data

We use public medical datasets within the MEDFAIR medical imaging fairness benchmark [71], namely CheXpert [77] and Ham10000 [86]. CheXpert is a large dataset that contains chest X-rays annotated for 14 conditions, while Ham10000 is moderately sized and features skin lesions categorised into seven types. Both datasets provide protective characteristics of patients like age, biological sex, and self-reported race. We mainly experiment with CheXpert, building on prior research (see subsection 2.2.6). Ham10000, with its different modality, size, and subgroup separability (refer to subsection 2.2.7), offers a valuable counterpoint. Due to its smaller size,

Dataset Name	Modality	Sensitive Attributes	Labels
CheXpert	Chest X-ray	Age, Sex, Race	No Finding, Other
Ham10000	Skin Dermatology	Age, Sex	Benign, Malignant

Table 3.1: Datasets and their overall characteristics used in this study.

we also used Ham10000 for initial training framework tests. A concise overview of datasets is present in Table 3.1.

3.3.1 Data Processing

In preparation for model training, we partition data for training, validation, and testing based on the "Reading Race" study [10] for CheXpert and "Subgroup Separability" study [81] for Ham10000. Our preprocessing follows MEDFAIR's approach [71] of removing data without sensitive attributes and binarising labels and sensitive attributes. Images from the same individual or lesion are not duplicated across various subsets. The validation set is used to select the model during training, while the test set assesses its final performance.

For CheXpert, we categorise the labels into 'no finding' (healthy lungs) and 'other' (medical condition). Analogously, for Ham10000, we classify lesions as 'benign' (non-cancerous) or 'malignant' (cancerous). Sensitive attributes include age, biological sex, and race for CheXpert and age and sex for Ham10000. Based on [81], we classify race as 'white' or 'non-white', sex as 'male' or 'female', and age as 'under 60' or 'over 60'.

During training, scans undergo a series of standard transformations for ResNet models. Images are resized to 224x224 pixels, with data augmentation to enhance generalisation abilities. Additionally, given our use of pre-trained weights, we normalise images using ImageNet's mean and standard deviation for each RGB channel.

3.3.2 Study Population

We break down the population characteristics for both CheXpert and Ham10000 in Table A.1 and Table A.2 respectively.

The CheXpert dataset contains 42,884 individuals with 127,118 chest X-ray images, split into training (76,205), validation (12,673), and testing (38,240) sets. Conversely, the Ham10000 dataset includes 7,418 lesion images with a total count of 9,958 scans distributed among training (7,967), validation (989), and testing (1,002) sets. Both datasets show notable imbalances among their subcategories. In CheXpert, only 9% of scans are labelled as 'no finding', while just 14% of Ham10000 are 'malignant' lesions. Additionally, 78% of CheXpert scans are from White individuals, and 72% of Ham10000 come from younger people. Patients in CheXpert are also on average 11 years older than those in Ham10000. Notably, in both datasets, disease labels are more prevalent among older age groups. Specifically, in CheXpert, these labels are 4 percentage points higher for the elderly compared to the overall dataset. For Ham10000, this difference rises to 13 percentage points.

3.4 ResNet Training Setup

In the course of our research, we exclusively train various Residual Neural Networks (ResNets) [1]. ResNet is a prominent DL imaging architecture that introduced the concept of residual connections, facilitating the network's ability to learn identity mappings across multiple layers. We employ several ResNets of varying scales (e.g., ResNet18, ResNet34), where the number indicates the amount layers in the network. Their details are shown in Figure A.2. This approach enables us to simulate the KD context effectively, wherein a larger-scale model (e.g. ResNet101) typically serves as the teacher for a smaller student (e.g. ResNet18). The deliberate choice to remain consistent with a single model type was to ensure that the architecture had minimal influence on the experimental outcomes.

3.4.1 Hyperparameters

ResNet model training hyperparameters are detailed in Table 3.2. Our main goal was not hyperparameter optimisation for peak performance. Instead, most values were set empirically or from experience to achieve satisfactory performance.

One notable observation was the importance of a learning rate scheduler in preventing overfitting. It reduces the learning rate by a factor of 10 if the monitored metric (validation loss) does not improve within its patience factor. Without this, smaller ResNets performed best, while larger models struggled to learn effectively. When reducing the learning rate, we revert to the best weights seen during training. Otherwise, rate reductions often came too late, leading the model down a less optimal path. We stop training after 9 epochs of non-improving validation loss. While the epoch ceiling was set at 50, instances of reaching it appeared rarely. We also used random weighted sampling with replacement in all tests. We elaborate on this design decision in subsection 3.5.4.

3.4.2 Training Technique

When training ResNet models, it is common to use one of the three following setups:

1. **Training From Scratch**: The model is initialised with random weights, and the entire network is trained from scratch.

2. **Fine-Tuning**: The model begins with pre-trained weights from a ResNet trained on the ImageNet [87] dataset. The old classification layer is replaced, and the entire network is trained end-to-end.

3. **Transfer-Learning**: Like fine-tuning, it uses pre-trained weights and a new prediction layer. However, only this new layer is trained, keeping the rest of the network untouched.

In our search for the best training technique, we assess approaches 1 and 2. We argue that method 3, which leaves the majority of the network unchanged, is not logical since the ImageNet [87] dataset lacks specialised medical images and is not representative of our target domain.

Config	Value
Overall	
Architecture	ResNet18/34/50/101
Optimiser	Adam {lr: $1e-4$, β_1: 0.9, β_2: 0.999}
LR Schedule	ReduceLROnPlateau {Monitor: val loss, Patience: 3}
Early Stopping	{Monitor: val loss, Patience: 9}
Max Epochs	50
Batch Size	64
Augementation	RandomResizedCrop, RandomRotation(15°)
Sampling	Random Weighted
Response-Based KD	
Temperature T	8
Alpha α	0.2
Feature-Based KD	
Alpha α	0.2
Beta β	0.2
Gamma γ	0.6
Feature Matching	2nd & 4th main ResNet layer

Table 3.2: Hyperparamters and training components used throughout the project for traditional training and knowledge distillation.

We perform a set of preliminary training experiments over 3 random seeds using all ResNet scales from Table 3.2 for the Ham10000 dataset and show the AUC scores in Figure 3.1. Fine-tuning both with and without random weighted sampling follows a desirable pattern, indicating enhancements with larger models. These approaches were picked for later experiments. As can be seen on Figure 3.1, most likely because of the lack of data, training from scratch turned out to be an ineffective technique, leading us to abandon it for future tests.

3.5 Knowledge Distillation Setup

To evaluate the fairness of different distillation techniques, we have implemented both response-based KD and feature-based KD. Opting for response-based KD was straightforward, given its status as the original and most widely adopted method. Our curiosity lies in understanding how aligning logits affects the fairness of the student model. Conversely, feature-based KD is another recognised method that acts as a contrasting example to standard KD and adds additional supervision. Precisely, we have implemented and evaluated a variation of FitNets [24] and attention-transfer [25]. While both methods are closely related, their subtle differences, unique objectives, and the potential to address the capacity gap issue make them especially

interesting for our experiments. For simplicity we refer to the response-based KD as *Response KD*, FitNets implementation [24] as *Feature KD* and to attention-transfer [25] as *Attention KD*.

Given our aim to compare teacher and student models, we find offline KD most suitable for our study. Exploring online and self-distillation would complicate matters and might not yield as meaningful results. For instance, assessing fairness in online distillation would probably involve comparing models from different training phases. We believe this is less insightful, as the primary differences would stem from the maturity of the model at different epochs rather than its inherent characteristics.

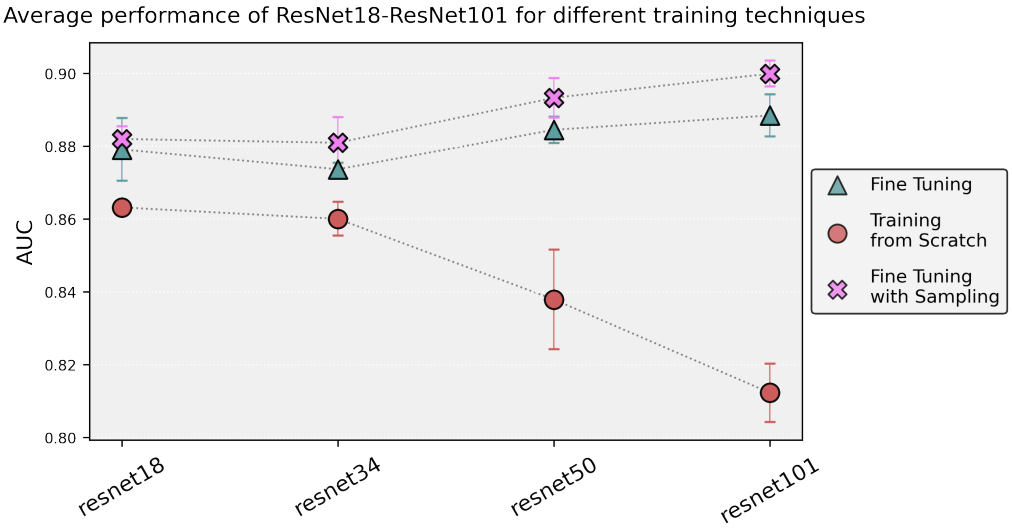

Figure 3.1: Performance comparison of different training techniques for ResNet18-ResNet101 models averaged across 3 seeds. Error bars show standard deviation.

3.5.1 Response KD

We implement the original Response KD method from [13] with the loss function in Equation 2.2. Before our primary experiments, we performed a hyperparameter search for optimal α and T. A ResNet101 served as our teacher model, and we trained ResNets18-34 with different seeds, α, and T values to assess post-distillation results. AUC scores are shown in Figure A.1. While performance was consistent, $\alpha = 0.1$ was notably less effective. In the end, we chose $T = 8$ following best results and $\alpha = 0.2$ to prioritise the teacher model's influence, despite $\alpha = 0.4$ having a peak performance. See Figure 3.2 for a visual representation of the performance boost post Response KD.

3.5.2 Feature KD

For Feature KD, we adapted a version of FitNets [24] to align multiple intermediate feature maps. The ResNet architecture, as illustrated in Figure A.2, consists of four main convolutional layers in every model (excluding the initial simple layer), regardless of its size. Given this structure, it made sense to position feature matching between these core layers.

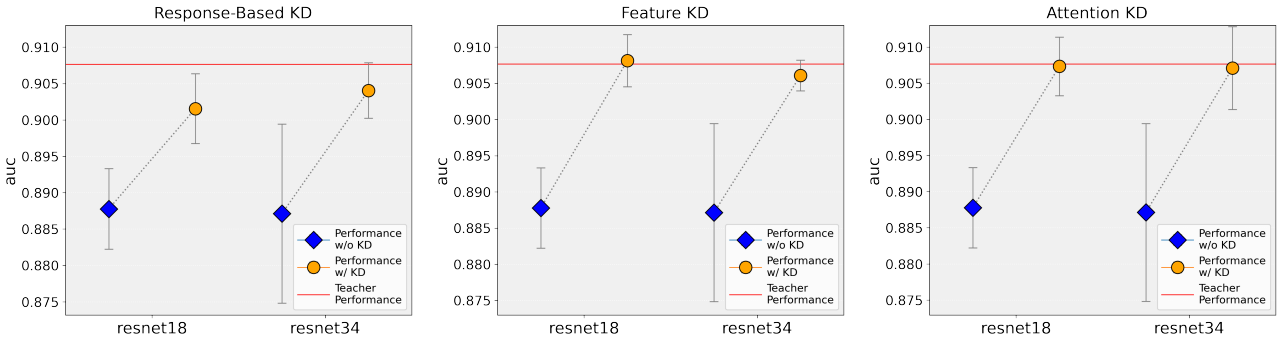

Figure 3.2: Ham10000 performance pre and post Response KD (left), Feature KD (middle) and Attention KD (right) for ResNet18 and ResNet34. Results averaged over 5 random seeds.

Regressors

To achieve accurate alignment between the student's and teacher's feature maps we implement the $\Phi(.)$ transformation function from Equation 2.3 (regressors) through a 1×1 convolutional layer applied to student's feature maps. This method maintains the spatial dimensions of the feature maps and adjusts the channel depth to align with the teacher's depth. The straightforward nature of this approach ensures minimal interference with the main results and leverages the consistent spatial dimensions inherent to ResNet models, regardless of their size. The parameters of regressors are also updated during training.

Training Procedure

In contrast to the 2-stage training procedure outlined in the FitNets paper [24], our approach combines cross-entropy, response-based, and feature distillation losses into a unified loss function. This modification is motivated by recent work [28, 29] and an aspiration to reduce computational training overhead. The respective losses are weighted by coefficients α, β, and γ, producing the following Feature KD loss:

$$L_{FeaKD} = \alpha L_{CE}(y, p(z_s, 1)) + \beta L_{KL}(p(z_s, T), p(z_t, T)) + \gamma \left(\frac{1}{N} \sum_{k \in \mathcal{K}}^{N} L_{MSE}(\Phi(F_s^k), F_t^k) \right) \tag{3.3}$$

where L_{MSE} is a mean squared error loss averaged out over the indices of all matched feature maps $\mathcal{K}$ with N being their total number and M being the number of elements in each map flattened across its dimensions:

$$L_{MSE}(F_s^k, F_t^k) = \frac{1}{M} \sum_{j=1}^{M} (\Phi(F_s^k)[j] - F_t^k[j])^2 \tag{3.4}$$

To determine the best coefficient combination and feature-matching placement, we ran experiments using a proficient teacher model and smaller-scale students. We assessed the alignment of two feature map sets: one after the 2nd and 4th layers and another across all layers. The results with various loss coefficients averaged over 10 random seeds are in Table 3.4.

While we considered excluding the response-based loss, models had learning challenges without it. A linear combination of the three losses showed promising results, with the AUC difference relative to the teacher being minimal. We chose to use all three loss functions, giving weights of 0.2 to both cross-entropy and KL divergence and 0.6 to the MSE between feature maps 2 and 4, as shown in Table 3.4. While equal weights for response and feature losses might seem logical, we emphasised the feature-based aspect for a better comparison with traditional KD. Moreover, aligning all four features gave similar outcomes, but this might introduce excessive guidance. The configuration we used is summarised in Table 3.2, and Figure 3.2 visually displays the performance improvement from Feature KD.

3.5.3 Attention KD

Beyond direct feature alignment, we also use the Attention KD. This method is used primarily in our analysis of the capacity gap detailed in section 4.4, given its beneficial nature for this task. A key difference from Feature KD is that attention transfer does not need regressors. This is because the attention mechanism focuses only on the spatial dimensions of the feature maps, which, due to the ResNet architecture, remain consistent across all ResNet scales. Specifically for a feature map $F \in R^{C\times H\times W}$, with C feature planes of $H \times W$, activation-based mapping function $\mathcal{A}$ flattens it over the spatial dimensions [25]:

$$\mathcal{A} : R^{C\times H\times W} \rightarrow R^{H\times W}.$$

We utilise sum of absolute values of tensor F, raised to the power of 2 as a function $\mathcal{A}$ - a method most commonly used in the original paper:

$$\mathcal{A}^2_{sum}(F) = \sum_{i=1}^{C} |F_i|^2 \tag{3.5}$$

As a final loss function, we again employ a linear combination of all losses:

$$L_{AttKD} = \alpha L_{CE}(y, p(z_s, 1)) + \beta L_{KL}(p(z_s, T), p(z_t, T)) + \gamma \left(\frac{1}{N} \sum_{k\in\mathcal{K}}^{N} L_{MSE}\Big(\frac{Q^k_s}{\|Q^k_s\|_2}, \frac{Q^k_t}{\|Q^k_t\|_2} \Big) \right) \tag{3.6}$$

where $Q^k_s = \mathcal{A}^2_{sum}(F^k_s)$ and $Q^k_t = \mathcal{A}^2_{sum}(F^k_t)$ are their k-th pair of matched attention maps between student and teacher. Additionally, as suggested in the original work, we perform l_2 normalisation of the attention maps before computing their difference. For a better comparison and because of promising performance improvements (as shown in Figure 3.2), we employ the same loss function weighting as Feature KD (see Table 3.4).

3.5.4 Sampling Techniques

In our quest for an optimal training configuration, it became evident that executing KD without sampling resulted in only modest performance enhancements. This observation prompted us to design a series of experiments to explore various sampling combinations, which include:

1. **No Sampling**: The data loader does not employ any sampling technique and data points are simply shuffled randomly.

2. **Random Weighted (RW) Sampling**: Weights are assigned to each data point based on its label y_i. For a label y_i with a count $C(y_i)$ in the dataset, the weight $W_{RW}(i)$ for the i^{th} sample is given by:

$$W_{RW}(i) = \frac{1}{C(y_i)}$$

3. **Subgroup Sampling**: For the i^{th} sample, weights based on its j^{th} sensitive attribute a are calculated as the inverse of their count:

$$W_s(i,j) = \frac{1}{C(a_{ij})}$$

Then, considering the labels and sensitive attributes (where the product runs over all L attributes A_i), the final weight for the i^{th} sample is:

$$W_S(i) = W_{RW}(i) \times \prod_{j=0}^{L} W_s(i,j)$$

We employed three highly performing ResNet101 teacher models, each fine-tuned using a distinct sampling technique. Subsequently, we conducted Response KD on ResNets 18 and 34, experimenting with various combinations of sampling techniques and teacher models. The outcomes, along with baseline performances (trained without KD), averaged over three random seeds, are presented in Table 3.3.

Student \Teacher (with their performance)			**ResNet101 No Sampling** 0.8949	**ResNet101 RW Sampling** 0.9045	**ResNet101 Sub. Sampling** 0.8902
ResNet18	**No Sampling**	0.8791	0.8713(↓0.9)	0.8895(↑1.18)	0.8760(↓0.35)
	RW Sampling	0.8821	0.8768(↓0.6)	0.9031(↑2.38)	0.8965(↑1.63)
	Sub. Sampling	0.8923	0.8752(↓1.91)	0.9001(↑0.87)	0.9006(↑0.93)
ResNet34	**No Sampling**	0.8737	0.8828(↑1.04)	0.8957(↑2.51)	0.8915(↑2.03)
	RW Sampling	0.8810	0.8808(↓0.02)	0.9000(↑2.15)	0.8995(↑2.1)
	Sub. Sampling	0.8860	0.8777(↓0.92)	0.9035(↑1.99)	0.9004(↑1.63)

Table 3.3: AUC performance across various teacher and student models using Response KD and different sampling methods. Numbers in brackets show performance change relative to pre-KD students with arrows indicating the direction. 'RW' means Random Weighted; 'Sub.' is short for Subgroup. Results are averaged over 3 random seeds.

Based on these results, the teacher model without sampling provided suboptimal guidance to student models, even if no sampling was used during distillation. In contrast, the model trained with random weighted sampling consistently excelled, regardless of the sampling method used in KD. While subgroup sampling is a notable alternative, we ultimately chose random weighted sampling for both model training and KD. This was based on its favourable results and our dedication to ensuring consistency across all experiments.

Loss Functions & Their Weights			**Layer Matched**				**Performance Metrics**		
Cross Entropy	**Response (KL-div)**	**Feature (MSE)**	**1st**	**2nd**	**3rd**	**4th**	**AUC**	**AUC Change**	**AUC Teacher Diff.**
Feature KD									
0.2	-	0.8	✗	✓	✗	✓	0.888 ± 0.016	-0.001	0.020
			✓	✓	✓	✓	0.885 ± 0.005	-0.003	0.023
0.5	-	0.5	✗	✓	✗	✓	0.890 ± 0.010	0.001	0.018
			✓	✓	✓	✓	0.892 ± 0.008	0.004	0.015
0.2	0.4	0.4	✗	✓	✗	✓	0.907 ± 0.003	0.019	0.001
			✓	✓	✓	✓	0.906 ± 0.004	0.017	0.002
0.2	0.2	0.6	✗	✓	✗	✓	0.908 ± 0.006	0.019	0.000
			✓	✓	✓	✓	0.906 ± 0.006	0.018	0.001
Attention KD									
0.2	0.2	0.6	✗	✓	✗	✓	0.907 ± 0.005	0.019	0.000
Response KD									
0.2	0.8	-	-	-	-	-	0.903 ± 0.004	0.014	0.005

Table 3.4: Feature KD preliminary results on Ham10000 with varied loss functions and layer matching. 'Layer Matched' refers to the index of main ResNet layers from Figure A.2. 'AUC Change' is the difference pre-KD to students and 'AUC Teacher Diff.' is the difference to the teacher. The green combination is our final setup, with Attention and Response KD compared.

3.6 Teacher-Student Setup

This section provides insight into our methodology for choosing specific teacher and student models for upcoming experiments. We delve into the reasoning behind these choices and how they relate to the broader research landscape.

3.6.1 Motivation

Consider a reputable healthcare institution that publishes a groundbreaking disease prediction foundation model trained on a costly, diverse, and balanced dataset. This model showcases impressive predictive accuracy and consistently performs well across different demographic groups.

Compare this to a smaller clinic facing limited computational resources and data constraints. For them, deploying such a resource-heavy model in real time is unfeasible. Furthermore, their limited access to a comprehensive and balanced dataset hinders their ability to develop a proficient model internally. However, recognizing the potential of KD, they aspire to craft a more compact model, using their data but guided by the published foundational model.

This hypothetical scenario underpins our research's experimental design. We utilise a teacher model that demonstrates high performance coupled with reduced subgroup disparities, which we name as the 'fair teacher'. It is worth noting that although we use the term fair, it does not mean complete fairness. Instead, in our context, it denotes minimal biases compared to other models we have evaluated. Conversely, we introduce 'unfair students' models trained to emphasise subgroup disparities. With the rising significance of foundational models in healthcare [88, 76] and growing discussions on AI fairness, we believe such a scenario is increasingly relevant.

Moreover, choosing models this way lets us simulate a situation with distinct model differences beyond overall performance. In this context, because of their discrepancy, subsequent comparisons between teacher and students, before and after distillation, emerge as an interesting avenue of exploration. Additionally, the inverse setup, where an unfair teacher instructs a fair student, has been preliminary explored in [19]. We argue that our scenario more accurately reflects practical real-world considerations.

3.6.2 Picking Fair Teachers & Unfair Students

The following subsection explains experiments that determined training techniques for picking fair and unfair models for future analysis.

Experimental Set Up

In our experiments, we evaluate training methods using either constraint optimization or unique data compositions. We train 20 models for each technique, distributed across ResNet with 18, 34, 50 and 101 layers. This procedure is carried out separately for Ham10000 and CheXpert. Training methods are grouped by fairness: blue hues represent fair approaches, while orange indicates unfair paradigms.

1. **Lagrange Optimisation with Equal Loss constraint** (Lagrange): Using the original data composition, we apply Lagrange optimization constrained with equal loss (Equation 2.15), bounded by ϵ. Initial hyperparameter tuning determined the optimal ϵ and starting Lagrange multiplier λ.
2. **Equal Data Composition** (Equal Data): In this method, we ensure equal sample distribution across all values of all sensitive attributes in training and validation sets. Given HAM10000's limited size, we also use an upsampled variation (Equal Upsample).
3. **Subgroup-Reversed Data Composition** (Subgroup Reversed): This method inverts subgroup distributions. For example, if the original dataset had a male majority, after additional sampling, the majority becomes female, and vice versa.
4. **Original Data Composition** (Original): Include training models with the original data splits described in subsection 3.3.2, representative of the study population.

5. **Unfair Data Composiion** (No Subgroup): One subgroup is omitted from the development set, expecting potential subgroup disparities in the test set. We evaluate all possible exclusions and refer to them as 'No *Subgroup*', e.g. No Female.

Across all techniques, the test set remains unchanged, with modifications limited to training and validation sets. We assess each training approach using both the overall AUC and the relative average AUC^{GAP}. Ideally, in contrast to unfair students, a fair teacher should display high overall performance coupled with smaller subgroup disparities.

Results

In Figure 3.3, we display training results for all methods on the Ham10000 and CheXpert datasets. Overall performance varies more than subgroup disparities, with most techniques yielding a lower AUC than the original approach. In contrast, AUC^{GAP} shows limited variation, particularly for Ham10000, regardless of the fairness method. Exceptions include the 'Lagrange' and 'Subgroup Reversed' methods, which increase and decrease subgroup disparities, respectively. However, CheXpert's results differ: 'Lagrange' reduces bias, while 'Subgroup Reversed' intensifies it. In general, CheXpert's results, though subtly different, align closer to our initial expectations. Unfair techniques tend to result in lower performance and a higher AUC^{GAP}, while fair methods show disparities just below the baseline.

Evaluation

The results highlight the intricacies of achieving fairness in model development, as shown by extensive prior research on fairness-constrained training (refer to subsection 2.2.4). Within our specific context, identifying a clear fair and unfair training method is challenging due to similar performance metrics. This challenge is amplified by the differing results between the Ham10000 and CheXpert datasets, especially in the 'Lagrange' and 'Subgroup Reversed' outcomes. We speculate that Ham10000's limited fairness variability might be due to its low subgroup separability [81]. As shown in Figure 3.3 (second top plot), excluding a subgroup does not necessarily increase bias but often reduces overall performance, likely due to a smaller dataset size. Thus, creating universally effective methods for both datasets is demanding.

Our final training choices with specific teacher models are detailed in Table 3.5. To identify unfair students, we chose one unfair data composition for each sensitive attribute: No Female, No Old, and No White. While seemingly arbitrary, our closer analysis showed these methods yielded satisfactory outcomes. For fair methods, we favoured the original data splits due to their high performance, low subgroup disparities, and enhanced clarity.

3.7 Implementation Details

We implement all our code using Python 3.11[1]. The majority of our models and training components are constructed using PyTorch Lightning 2.0[2], a streamlined PyTorch wrapper used by the research community for faster development. We utilise PyTorch implementation of

[1] https://www.python.org
[2] https://www.pytorchlightning.ai

ResNet with pretrained ImageNet weights. All model training was conducted on GPUs housed within the BioMedia[3] research group at Imperial College London, specialising in biomedical image analysis. The GPU suite comprises 11GB Nvidia RTX1080Ti, RTX2080Ti, GTX1080Ti, and 25GB Nvidia TITAN RTX units. The models necessitate a minimum GPU memory of approximately 8 GB. We employed CUDA version 12.2, and to ensure reproducibility, we initialised a random seed via Pytorch Lightning's `seed_everything()` function.

For the computation of our performance metrics, as well as PCA and t-SNE analyses, we leveraged scikit-learn[4], a python-based ML library. The loss functions we employed are sourced from PyTorch, and both loss functions and the metrics are computed batch-wise and subsequently averaged over the entire epoch. In our statistical evaluations, we utilised SciPy 1.8.0[5], specifically employing the two-sample KS test and the JS distance to assess distributional disparities.

Data	**Fairness**	**Model**	**Random Seed(s)**	**AUC**	**AUC^{GAP}**
Teachers					
Original Ham10000	Fair	ResNet34	46	0.9074	0.054
		ResNet101	42	0.9129	0.0504
Original CheXpert	Fair	ResNet34	43	0.8591	0.0192
		ResNet101	46	0.8547	0.0211
Students					
Ham10000 No Female	Unfair	ResNet18 & ResNet34	41-46	0.8789	0.0817
Ham10000 No Old	Unfair	ResNet18 & ResNet34	41-46	0.8752	0.0812
CheXpert No Female	Unfair	ResNet18 & ResNet34	41-46	0.8221	0.0339
CheXpert No Old	Unfair	ResNet18 & ResNet34	41-46	0.8404	0.0239
CheXpert No White	Unfair	ResNet18 & ResNet34	41-46	0.8273	0.0223

Table 3.5: Comparison of various fair teacher and unfair student ResNet models used throughout this project for both CheXpert and Ham10000 datasets. Their performance (AUC) and average subgroup disparities (AUC^{GAP}) are also highlighted.

[3] https://biomedia.doc.ic.ac.uk
[4] https://scikit-learn.org/
[5] https://scipy.org

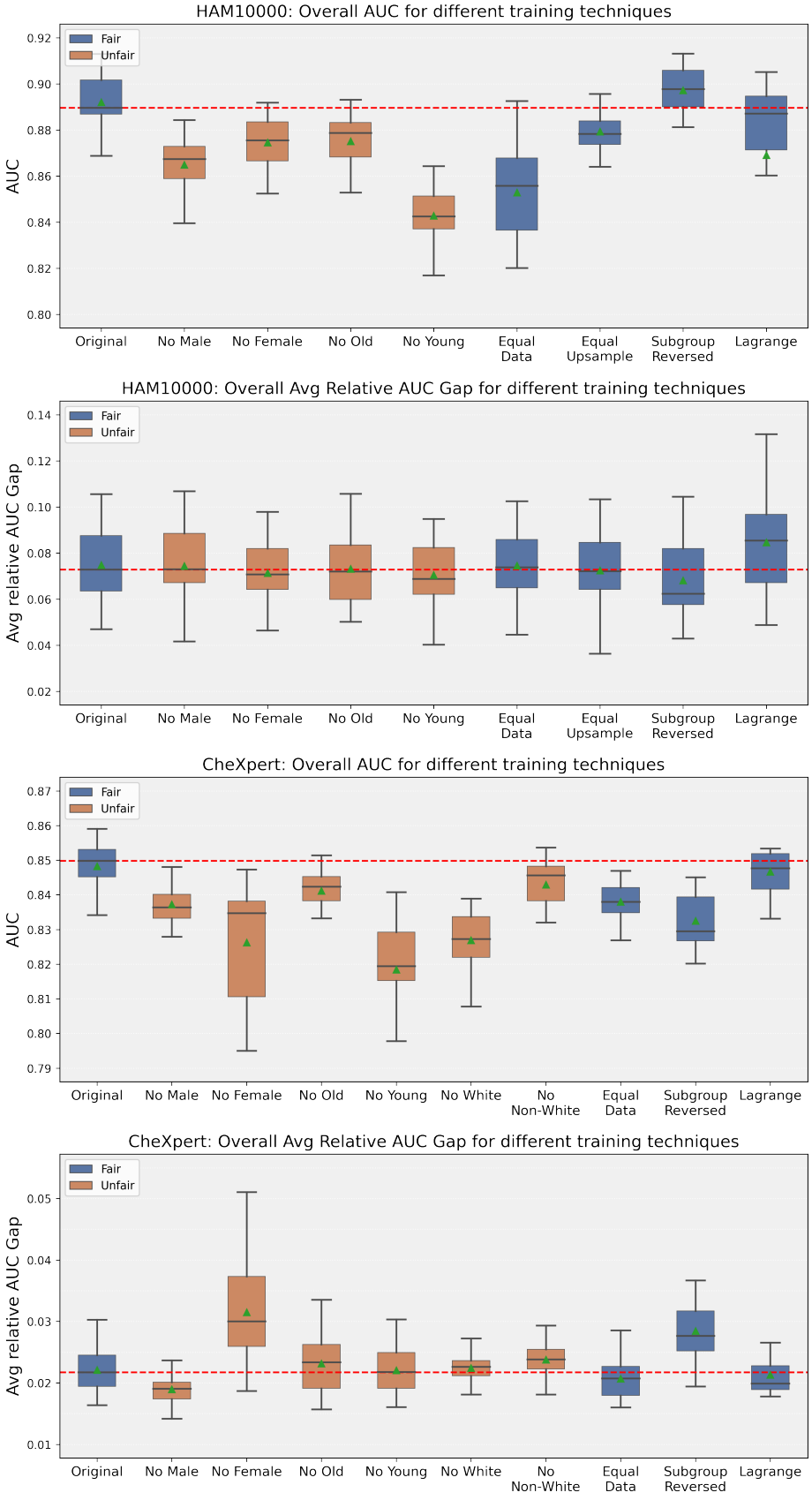

Figure 3.3: Overall AUC and AUC Gap for both fair (blue) and unfair (orange) training methods. The top two subfigures represent Ham10000, while the bottom two depict CheXpert.

Chapter 4

Research Questions

In this chapter, we explore a series of main research questions aimed at understanding how knowledge distillation affects fairness and model behaviour. It is purposefully structured to ensure each study case is presented in its entirety, from its motivation and experimental setup to results and their interpretation. For clarity, it is best to read them sequentially, as the outcome of one often informs and motivates the subsequent experiment.

4.1 Can KD from a Fairer Teacher Help in Training Originally Unfair Students?

In prior literature, student models have been shown to emulate their teacher models during KD [19, 41]. While most studies have emphasised performance outcomes, fairness and subgroup disparities remain overlooked. Additionally, these investigations typically occur in environments tailored for optimal performance, with carefully selected teacher and student models. In this experiment, we explore the effects of discrepancy between teacher and students concerning subgroup disparities. We specifically examine if biased student models can improve their fairness by distilling knowledge from a fairer teacher model, realising the scenario depicted in section 3.6

Experimental Setup

We perform Response KD using a fair ResNet34 teacher from both Ham10000 and CheXpert datasets and ResNet18-34 student models derived from biased data compositions (detailed in Table 3.5). The intentional selection of a smaller-scale teacher aims to avoid capacity gap issues. The distinct training technique of the teacher and students introduces additional discrepancies between them, offering a unique experimental setup where KD models have divergent training backgrounds. Results are averaged over five random seeds, and we evaluate changes in performance (AUC), subgroup disparities (AUC^{GAP}), and subgroup performance (Subgroup AUC) pre and post-KD.

Figure 4.1: Radar plots comparing teacher and student overall (AUC) and subgroup (e.g. Male AUC) performance (with and without KD) trained with unfair data compositions: Ham10000 (top row) and CheXpert (bottom row). The student metrics are averaged over 5 random seeds.

Results

Figure 4.1 illustrates that student models when trained using Response KD on biased data, demonstrate improved performance metrics that closely match the teacher's. This alignment is especially pronounced in the CheXpert datasets, both in overall and subgroup metrics. Here, the teacher model's contour is largely covered by the student models after KD. In contrast, the Ham10000 dataset shows a less pronounced increase in similarity, but the plot with KD still trends towards the teacher across all metrics. Notably, images from older patients present the most significant challenges, with the age 60+ AUC consistently being the lowest in all cases. However, when examining various biased datasets, the differences between them are subtle. While we might expect a performance decrease for absent subgroups, it is not always the case. For example, in Ham10000 the performance of older individuals in the No Old training is closer to the teacher's than in the No Female setup. Yet, the Female AUC is slightly higher when trained with the No Old, rather than the No Female dataset.

The overall trends from Figure 4.1 are consistent with the disparities shown in Figure 4.2, as indicated by attribute-specific AUC^{GAPS}. In the CheXpert models, post-KD results often

align with or even surpass the teacher's performance. For instance, the No Female setup displays the most significant disparities before KD, but these are effectively mitigated post-distillation. Age in Ham10000 datasets remains the only attribute where discrepancies are not entirely rectified. This observation is further supported by detailed Ham10000 results in Table A.3, where, notably, ResNet18 models in the No Old training exhibit increased age-related disparities post-KD. A similar trend is observed for race disparities in ResNet34 models within the CheXpert dataset, as indicated in Table A.4.

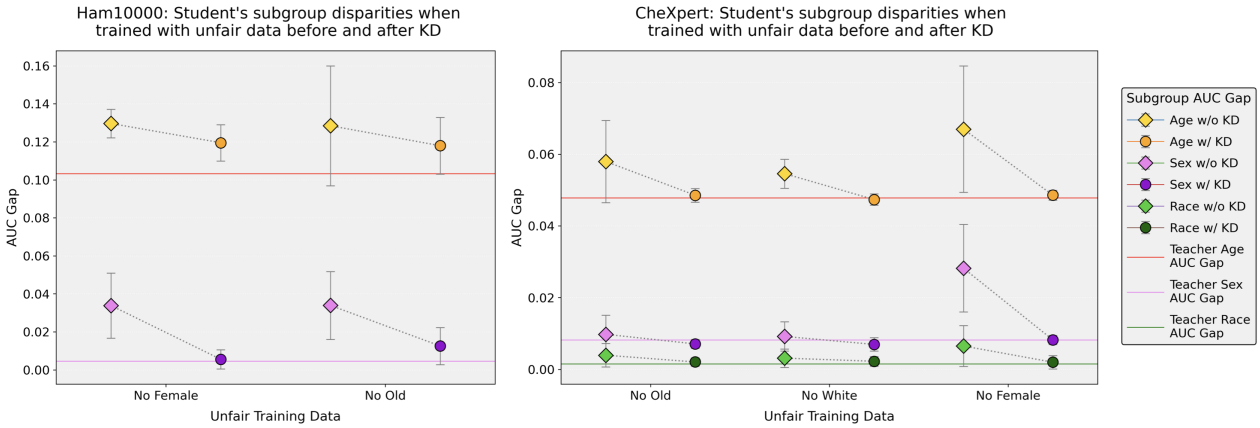

Figure 4.2: Subgroup performance disparities for student models (with and without KD) using unfair data compositions for both Ham10000 (left) and CheXpert (right). Student metrics are averaged over 5 seeds, and error bars show standard deviation.

Evaluation

The overall results suggest that even when trained on biased data, student models can significantly benefit by distilling knowledge from a more proficient and equitable teacher model. This is evident as they closely emulate the teacher in all evaluated scenarios. Hence, to a considerable degree, KD from a balanced model can correct the biases from the original training.

Our presented radar plots in Figure 4.1 offer a compelling visual representation of the model's prediction biases. They simplify the identification of more equitable models through symmetrical performance metrics. They also evidently, show that models face challenges in accurately diagnosing older patients, leading to pronounced age disparities. This observation holds for both the Ham10000 and CheXpert datasets. Such challenges might arise from complex medical cases, overlapping symptoms, and data limitations.

Furthermore, the superior effectiveness of KD on CheXpert compared to Ham10000 might be due to its larger dataset and, crucially, higher subgroup separability as outlined in [81]. For Ham10000, there seems to be minimal difference in the model's ability to learn from either biased or balanced datasets. The slight improvements seen are probably due to the diminished training data quantity. Notably, the occasional boost in the Race AUC^{GAP} post-KD in CheXpert, as shown in Table A.4, is most likely due to the baseline student's already elevated fairness for this attribute.

Answering our initial research question, the results reveal that student models trained on biased data exhibited improved fairness after Response KD from an equitable teacher model. Particularly in the CheXpert datasets, the students closely aligned with the teacher's metrics. This highlights a valuable attribute of KD and shows its potential as a corrective tool in ML.

4.2 What Patterns Emerge in the Unsupervised Feature Space for Response-Based KD?

Building on our earlier discoveries from section 4.1, which indicate that through KD unfair models can mirror the performance of a fair teacher, we delve further into the unsupervised feature space. Utilising model inspection methodologies on teachers and students both with and without KD, we aim to uncover underlying patterns that may not be directly evident from mere performance metrics.

Experimental Setup

We use PCA and t-SNE for high-dimensional data visualization, while logits provide a glimpse into our model's decision process. Given the vast outputs from these techniques, we concentrate our visual analysis on selected seeds from each data composition and student model size, as referenced in Table A.3 and Table A.4. These student models were chosen due to their pronounced unfairness. For a detailed analysis, we compute the JS distance between attribute-specific marginal distributions for the first four PCA modes of CheXpert test data, averaging the outcomes across all student models. Consistent with our prior experiment (section 4.1), we draw comparisons with the fair ResNet34 teacher model, focusing on Response KD.

Results

Considering the high number of plots generated for each model-dataset pairing, both with and without KD, encompassing PCA, t-SNE, and logits, Figure 4.3 and Figure 4.4 serve as illustrative summaries of our main observations. We showcase scatter plots from the teacher for PCA modes 1-2 and 3-4, complemented by selected student model examples post-KD. To see all model inspection plots, we refer the reader to the Appendix. It's worth noting that the CheXpert plots display clearer patterns, and thus, our discussion will largely centre on them.

Generally, feature representations projected to PCA space are not typically transferred from teacher to student during Response KD, resulting in diverse distributions. Yet, there are clear instances where students somewhat reflect their teachers' dimensionally-reduced feature embeddings. For example, ResNet18 distilled on the No Female CheXpert data for PCA modes 1 & 2 displays a U-shaped representation similar to its teacher (Figure 4.4 second row). This similarity is especially evident in the logits space, where distilled models consistently adopt the same scatter plot shape. Additionally, students after KD exhibit fewer outliers than their non-KD counterparts. Both observations can be seen in subsection A.5.2 and subsection A.5.1.

While PCA mode 1 in both teacher and students separated binary labels similarly, the direction was often inverted. For instance, negative PCA values for 'no finding' in the CheXpert teacher matched positive values in students, and vice versa. Differentiation by age and occasionally

sex is also apparent, where especially in higher modes, sex-based separation becomes more pronounced (see Figure 4.4 fifth row). In Ham10000, a similar trend is observed, with disparities primarily associated with age subgroups. Overall, identifying significant distribution differences and post-KD corrections proves difficult. Table 4.1 shows that only in PCA mode 1 does the gap between marginal distributions consistently narrow towards the teacher after Response KD (indicated by green or yellow colouring). In later modes, this happens just four times out of twelve, with non-KD students often aligning more with the teacher. Notably, unfair students without KD often show lower separation between sensitive groups than fairer teacher (indicated by a negative sign). For example, the teacher separates patients based on age more distinctly than, or at least as pronounced as, the no KD students in PCA modes 1, 2, and 4, as shown in both Table 4.1 and Figure 4.4. Similar patterns are seen for race in modes 3 and 4 and sex in mode 3, supported by the KS-tests in section A.6. Finally, the t-SNE results in subsection A.5.2 echo the PCA insights, though their arbitrary dimensions are harder to interpret.

PCA Mode	Model(s)	Label Dist	Sex Dist	Age Dist	Race Dist
1	Teacher	0.56	0.04	0.212	0.033
	No KD	-0.048	0.023	-0.02	0.005
	Response KD	-0.005	-0.003	-0.001	-0.001
	Feature KD	-0.012	0.002	-0.009	-0.004
2	Teacher	0.332	0.111	0.187	0.073
	No KD	-0.078	0.01	-0.054	0.003
	Response KD	-0.084	-0.046	-0.038	-0.03
	Feature KD	0.024	0.005	0.004	0.012
3	Teacher	0.185	0.216	0.062	0.164
	No KD	0.004	-0.088	0.053	-0.078
	Response KD	0.036	-0.047	0.063	-0.05
	Feature KD	0.028	-0.054	0.018	-0.041
4	Teacher	0.286	0.058	0.138	0.093
	No KD	-0.104	0.077	-0.047	-0.025
	Response KD	-0.089	0.099	-0.048	0.02
	Feature KD	0.051	0.069	0.003	0.029

Table 4.1: JS distance between attribute-specific marginal distributions (e.g., sex: male vs. female) for the first 4 PCA modes on the **CheXpert** test set. The values represent the fair ResNet34 teacher's distance and the average difference to it for students across Response and Feature KD, averaged over all data compositions and seeds. Green and orange signify distances that are more aligned with the teacher than no-KD, with green being the closer of the two. Red indicates that no-KD distances are better aligned. Positive and negative signs denote higher and lower separation, respectively, compared to the teacher.

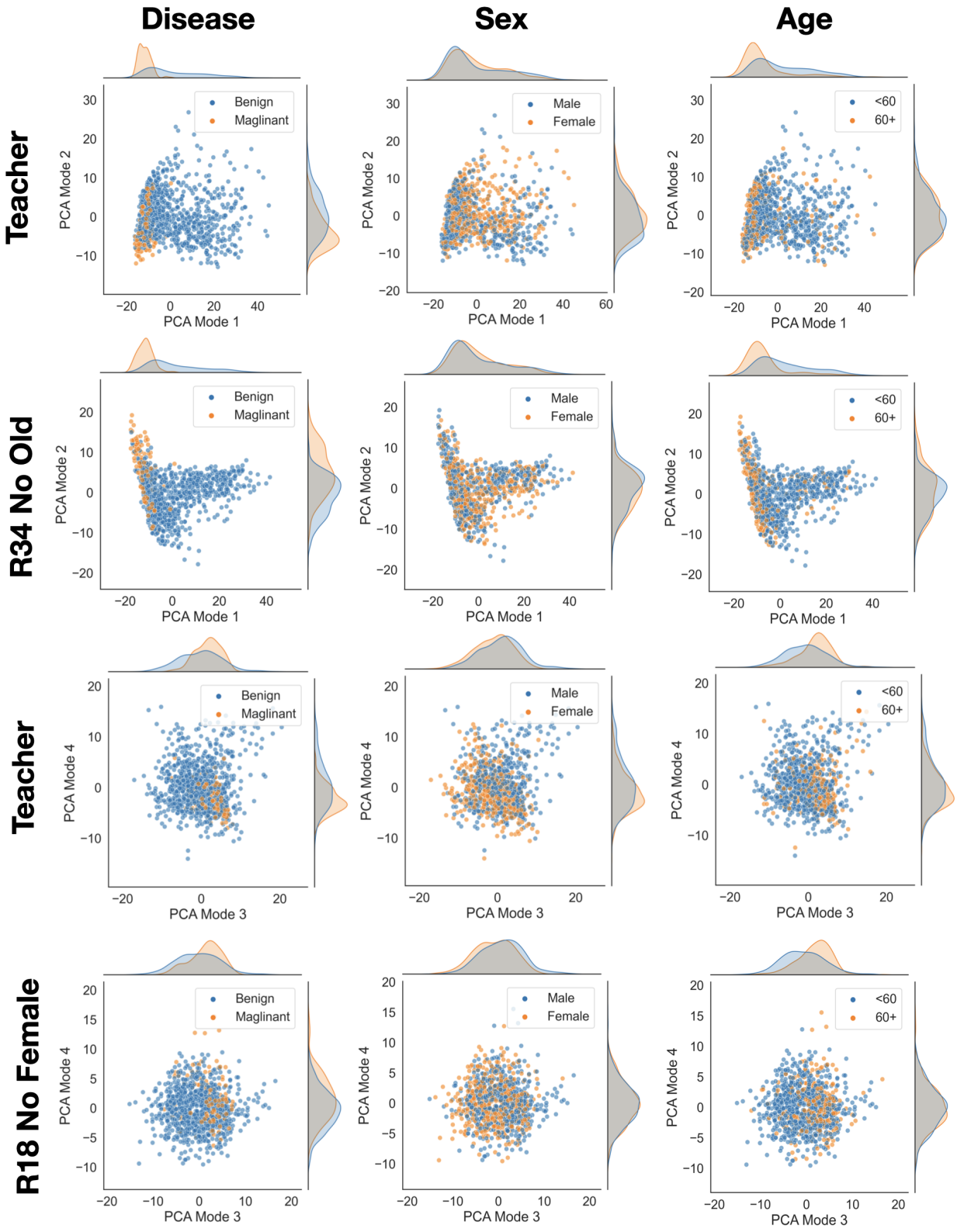

Figure 4.3: Example scatter plots with marginal distributions of the first 4 PCA modes for **Ham10000** test data feature representations, using the fair teacher and ResNet student models (e.g., R18 for ResNet18) trained on unfair data with Response KD. Displayed samples include disease, sex, and age details. See subsection A.5.2 for full results.

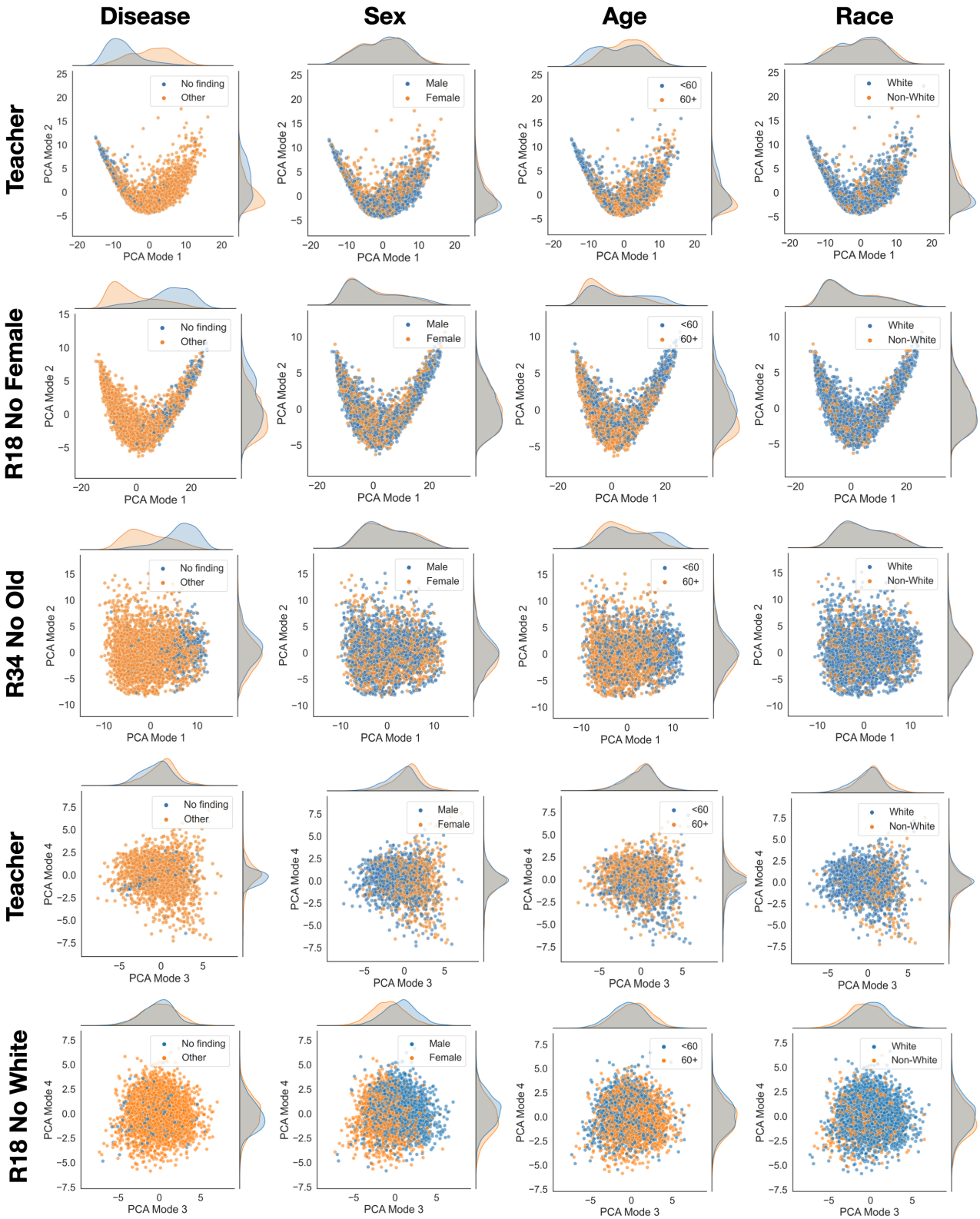

Figure 4.4: Example scatter plots with marginal distributions of the first 4 PCA modes for **CheXpert** test data feature representations, using the fair teacher and ResNet student models (e.g., R18 for ResNet18) trained on unfair data with Response KD. Displayed samples include disease, sex, and age details from a random patient set. See subsection A.5.2 for full results.

Evaluation

In examining Response KD, we find that while logits align, embeddings are not directly passed from teacher to student models. This indicates a subtle difference: even if models exhibit similar performance and behaviour, their internal information encodings and prediction processes differ. This variation is further underscored by the inversions seen in PCA modes. The U-shaped embeddings observed in some post-KD student models, primarily in ResNet18, suggest that these patterns may arise from the architecture's inherent characteristics and randomness rather than the KD process.

Overall, the results from the unsupervised feature space align with the findings presented in the previous study. The subgroup performance disparities, as referenced in Figure 4.2, manifest visibly in the embedding space. Consistent with the subgroup AUC^{GAP} metrics, age emerges as the most prominent factor in distribution separations, trailed by gender and subsequently race. A significant challenge, however, emerges when assessing the degree to which KD either amplifies or alleviates bias. Although our study indicates that student models mimic their teachers, it is only in the first PCA mode that the distance between the marginal distribution of student models closely matches those of the teacher. This is somewhat expected since later PCA modes account for less data variation. However, the interpretation becomes difficult when observing that the fairer teacher often displays more distinct separations than the unfair students. Moreover, the noticeable disparity in subgroup performance does not always match the differences in distribution, making our analysis more complicated.

We also recognise that more findings correlate with CheXpert than with Ham10000. This again could be attributed to Ham10000's smaller size and limited subgroup separability, leading to less distinct patterns. However, a consistent trend across both datasets is the reduced number of outliers in the embedding space for post-KD models compared to those trained without KD, as showcased in subsection A.5.2. This tendency might relate to the decreased performance variability shown in Figure 4.2, suggesting that training under a specific model's guidance offers more consistency than training networks autonomously.

In summary, while Response KD enables student models to mimic the performance of their teacher models, the internal feature encodings remain unique. Notably, performance in relation to age is identified as the most inequitable. However, variability in results across different datasets and conflicting findings in higher PCA modes highlight the inherent challenges of conducting a precise evaluation.

4.3 Does using Feature KD Lead to Different Distillation Outcomes?

After exploring Response KD, we now turn our attention to Feature KD. In this study, we investigate the effects of aligning not just the logits but also the intermediate feature maps between teacher and student models, aiming to discover the differences between the two KD techniques.

Experimental Setup

In this experiment, we focus on the CheXpert dataset due to its clearer findings relative to Ham10000 and the already extensive volume of results we have gathered. Our analysis covers performance metrics and subgroup disparities. When showing subgroup performance, we no longer use radar plots like Figure 4.1, because the patterns become unclear when adding the Feature KD contour. Instead, we introduce the *Similarity Boost* metric, which measures the change in Euclidean distance between the student and teacher subgroup performance metrics post-KD. We also present the *Performance Boost*, reflecting post-KD improvement for the same metrics relative to the baseline. Further, we explore the unsupervised feature space, mainly using PCA to evaluate high-dimensional data and subgroup marginal distributions' similarity via JS distance. We opted to exclude KS-tests and t-SNE as they rarely offer new insights. For Feature KD, we align the 2nd and 4th feature maps of the models through a loss function referenced in Equation 3.3. Consistent with our previous approach, we employ identical fair ResNet34 teacher and unfair students. The two sets of models are trained using the same five random seeds, and the same set of examples is chosen for a detailed unsupervised analysis.

Results

Both Response KD and Feature KD exhibit similar overall performance and subgroup disparities, as illustrated in Figure 4.5 and Figure 4.6, respectively. While Response KD largely mirrors the teacher's performance, Feature KD sometimes surpasses it. These distinctions, though subtle, are further shown in Table 4.2. In five out of six instances, Response KD exhibited a more pronounced similarity to the teacher (indicated by green colour), whereas it achieved this in just one out of six cases for performance. Notably, ResNet18 consistently had higher similarity gains than ResNet34 with Response KD, but this pattern was not observed with Feature KD. A closer look at PCA modes visualisation in Figure 4.7 shows a supporting example. For instance, ResNet18 trained on the No White dataset, displays an inverse marginal distribution for PCA mode 3 (4th row) in comparison to the teacher model (similarly as in section 4.2). However, this is absent for ResNet34, which closely mimics the teacher visually (5th row).

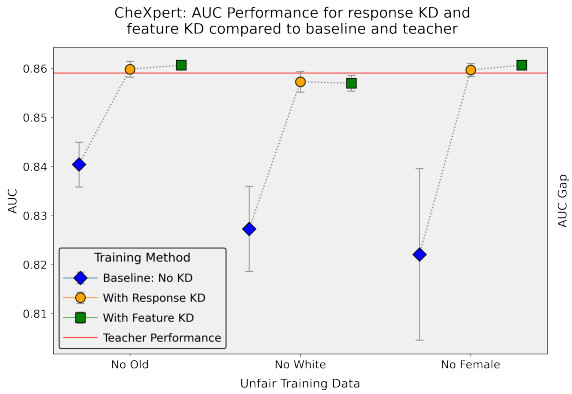

Figure 4.5: Performance for Response and Feature KD compared to no-KD models and ResNet34 teacher, across 3 unfair **CheXpert** data splits.

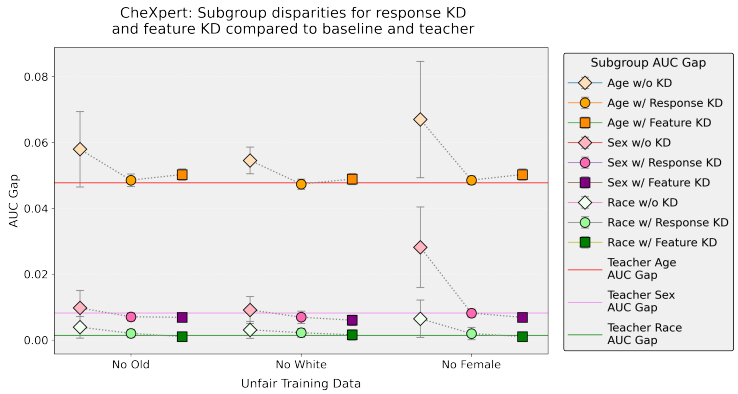

Figure 4.6: Subgroup disparities for Response and Feature KD compared to baseline no-KD models and ResNet34 teacher, split across 3 unfair **CheXpert** data splits.

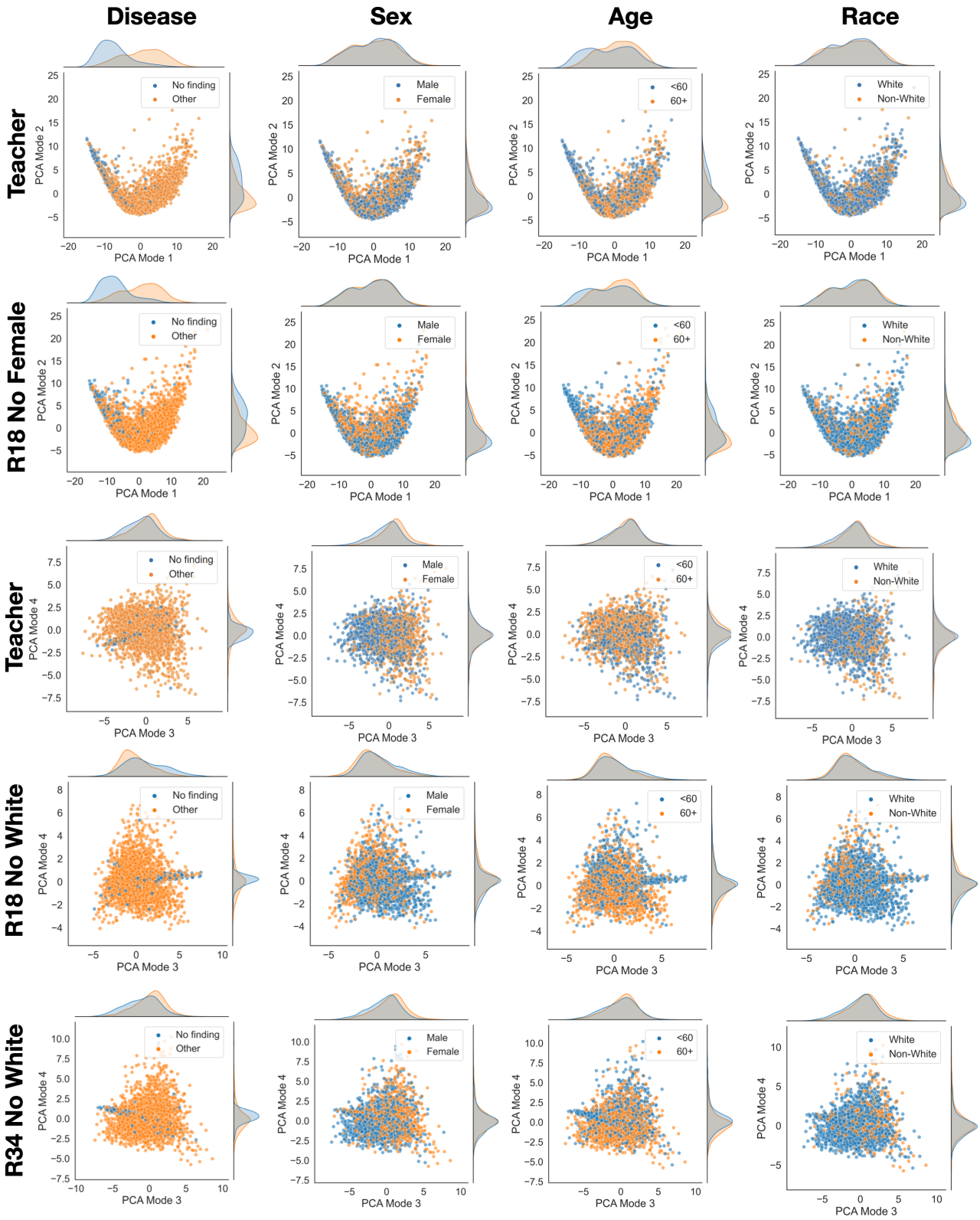

Figure 4.7: Example scatter plots with marginal distributions of the first 4 PCA modes for **CheXpert** test data feature representations, using the fair teacher and ResNet student models (e.g., R18 for ResNet18) trained on unfair data with Feature KD. Displayed samples include disease, sex, and age details from a random patient set. See subsection A.5.2 for full results.

Data	KD	Similarity Boost		Performance Boost	
		ResNet18	ResNet34	ResNet18	ResNet34
No Female	Response	97.12%	96.12%	3.73%	5.29%
	Feature	94.60%	94.68%	3.85%	5.39%
No Old	Response	94.12%	89.79%	2.19%	2.49%
	Feature	90.77%	87.99%	2.31%	2.54%
No White	Response	94.49%	93.04%	4.19%	3.12%
	Feature	92.47%	93.48%	4.07%	3.14%

Table 4.2: CheXpert subgroup similarity and performance boost for Response and Feature KD with varied unfair data compositions. Similarity boost measures the change in Euclidean distance between student and teacher subgroup metrics (e.g., Male AUC, Female AUC) post-KD. Performance boost denotes the student model's subgroup metric enhancement after KD. Green signifies a superior boost between Response and Feature KD.

Despite that, results in subsection A.5.2 and example visualisations for PCA modes 1 & 2 in Figure 4.7 suggest that Feature KD embeddings align closely with the teacher's, regardless of the student scale. Additionally, we see distinct clusters of subgroups and labels appearing on the same sides of the distributions (see Figure 4.7 second row). This feature map similarity is further validated and quantified in Table 4.1. The data there reveals that the distance between marginal distributions after Feature KD more closely mirrors the teacher's configuration compared to baseline no-KD models. Specifically, this is observed in 13 out of 16 possible instances, compared to a mere 8 for the Response KD.

Evaluation

Our findings confirm that while Response KD effectively mirrors teachers' abilities, Feature KD, by distilling logits and feature maps, provides a richer avenue for knowledge transfer. This dual alignment potentially allows student models to not only emulate but occasionally outperform the teacher. On the other hand, Response KD appears to be more limited by the teacher. However, the closeness of Feature KD and Response KD across metrics hints that our current experimental setup might be approaching a performance plateau.

Notably, while Response KD provides a superior similarity boost to the teacher based on subgroup performance, Feature KD stands out in its ability to reduce subgroup disparities and more closely match marginal distributions with the teacher. This result was unexpected and warrants further investigation. Additionally, the closer alignment of ResNet34 with the teacher is expected due to their common architecture, simplifying the distillation of feature maps. In contrast, ResNet18, perhaps because of its limited capacity and distinct architecture, might present anomalies like the noted PCA inversion.

In summary, both KD methods show similar effectiveness, but Feature KD excels in replicating the teacher's internal representations. If the goal is to closely emulate the teacher's decision-making process, Feature KD may be a better choice. Notably, it often surpasses the teacher in both performance and subgroup disparities, a phenomenon that warrants further investigation.

4.4 How does Capacity Gap Influence the Effectiveness and Behaviour of KD Methods?

In KD, the capacity gap (see subsection 2.1.3) between teacher and student often emerges as a significant concern. Our previous experiments, leveraging the architectural similarity between ResNet34 teacher and ResNet18/34 students, have conveniently avoided this issue. However, when ResNet101 was used as the teacher, we observed a decline in performance across all evaluation metrics for both datasets, as shown in Table A.12 and Table A.13. In this experiment, we delve into the dynamics of KD methods when faced with the capacity gap issue. We aim to uncover any subgroup patterns that might emerge and examine what happens in the unsupervised feature space. To further improve our analysis, we introduce Attention-KD, designed specifically to address the capacity difference of models.

Experimental Setup

We employ proficient and fairer ResNet101 teacher models with similar characteristics to smaller-scale teachers as detailed in Table 3.5. We reintroduce the Ham10000 dataset along with CheXpert to our analysis, especially given the notable differences in Ham10000 when faced with the capacity gap issue (see Table A.13). Our evaluation covers three distinct KD methods: Response, Feature and Attention KD. The latter operates on a training setup analogous to the Feature KD, as explained in subsection 3.5.3. Our goal is to compare these methods using performance metrics and model inspection techniques computed in a manner consistent with previous experiments. For unsupervised methods, given the large volume of data and plots, we summarise them with the JS distance between marginal distributions for each technique in tables, similar to previous experiments.

Results

Our results frequently reveal inconsistent outcomes within and between the two datasets. In CheXpert, feature-based methods often deviated from the teacher's performance metrics, as seen in Figure 4.8. Yet, they displayed subgroup disparities more in line with the teacher (see Figure 4.9). Specifically, Attention KD, despite its lower performance, exhibited disparities most similar to the teacher's. In contrast, for the Ham10000 dataset, performance metrics were more consistent with the teacher (refer to Figure 4.10), but post-distillation disparities were often amplified, as highlighted in Figure 4.11. Notably, while Attention KD exhibited top performance, it also displayed the most pronounced disparities, indicating increased unfairness. The pattern of amplified disparities for Attention KD over Response KD was consistent in all instances for Ham10000 and in 4 out of 6 for CheXpert. Meanwhile, Feature KD presented a more balanced approach between performance and fairness.

For similarity and performance boosts, CheXpert favoured Response KD (Table 4.5), while Ham10000 leaned towards Attention KD (Table 4.6). Notably, no significant advantages were observed for ResNet18 over ResNet34 for feature-based methods. In both pre and post-capacity gap introduction, ResNet34 often outperformed ResNet18 in terms of similarity and performance boosts.

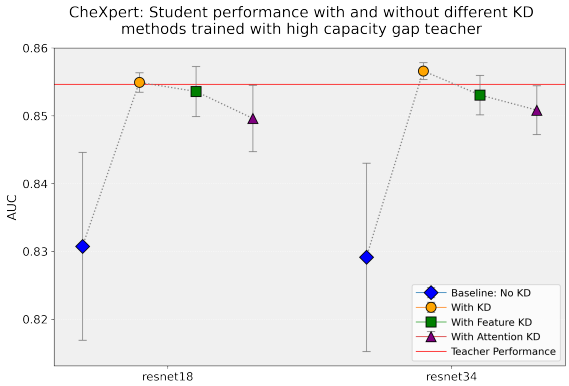

Figure 4.8: Performance for KD methods across students trained with a high-capacity teacher and averaged for all unfair **CheXpert** splits.

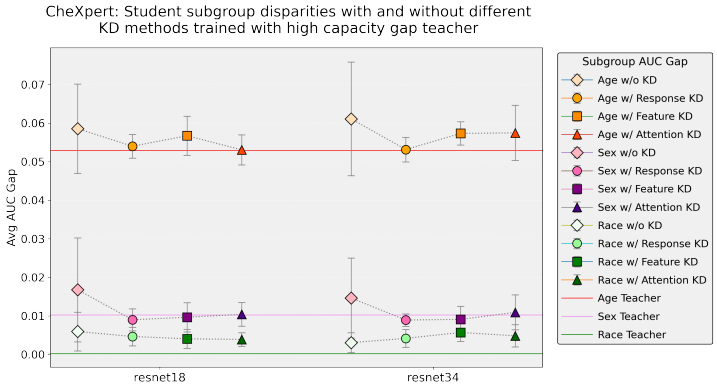

Figure 4.9: Subgroup Disparities for different KD methods across student models trained with a high-capacity teacher and averaged for all unfair **CheXpert** data splits.

PCA Mode	Model(s)	Label Dist	Sex Dist	Age Dist	Race Dist
1	Teacher	0.557	0.055	0.218	0.044
	No KD	-0.044	0.008	-0.026	-0.006
	Response KD	-0.011	-0.013	-0.002	0.001
	Feature KD	-0.008	-0.006	0	-0.003
	Attention KD	-0.01	-0.005	-0.007	-0.003
2	Teacher	0.345	0.065	0.129	0.028
	No KD	-0.09	0.057	0.004	0.047
	Response KD	-0.156	0.054	-0.051	0.019
	Feature KD	-0.106	-0.004	-0.009	0.027
	Attention KD	-0.132	0.015	-0.016	0.033
3	Teacher	0.293	0.075	0.161	0.091
	No KD	-0.103	0.054	-0.046	-0.005
	Response KD	-0.087	0.068	-0.057	-0.008
	Feature KD	-0.098	0.004	-0.069	-0.032
	Attention KD	-0.101	0.073	-0.075	-0.005
4	Teacher	0.287	0.043	0.188	0.054
	No KD	-0.105	0.092	-0.096	0.014
	Response KD	-0.096	0.07	-0.093	0.013
	Feature KD	-0.076	0.047	-0.071	0.004
	Attention KD	-0.105	0.081	-0.099	0.013

Table 4.3: JS distance between attribute-specific marginal distributions (e.g., sex: male vs. female) for the first 4 PCA modes on the **CheXpert** test set. The values represent the fair ResNet101 teacher's distance and the average difference to it for students across different KD methods, averaged over all data compositions and seeds. Green and orange signify distances that are more aligned with the teacher than no-KD, with green being the closest of the three. Red indicates that no-KD distances are better aligned. Positive and negative signs denote higher and lower separation, respectively, compared to the teacher.

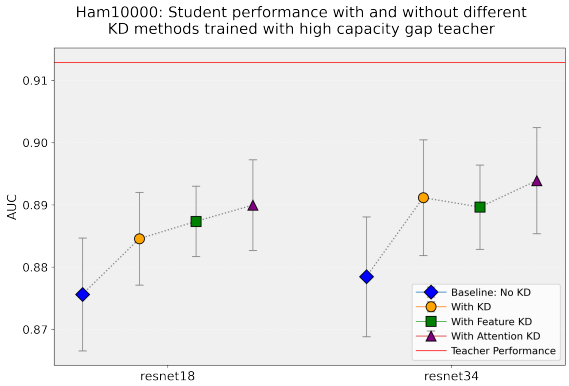

Figure 4.10: Performance for KD methods across students trained with a high-capacity teacher and averaged for all unfair **Ham10000** splits.

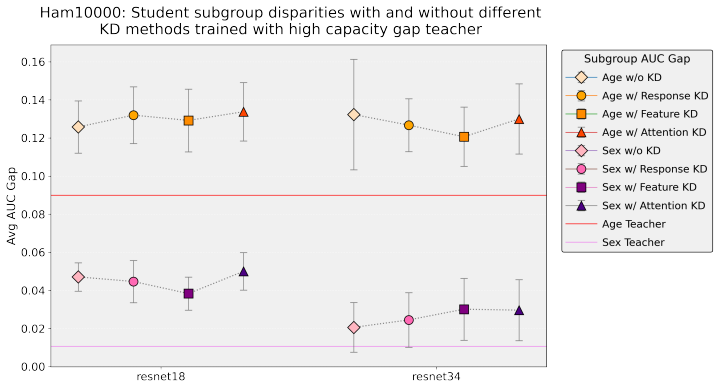

Figure 4.11: Subgroup Disparities for different KD methods across student models trained with a high-capacity teacher and averaged for all unfair **Ham10000** data splits.

PCA Mode	Model(s)	Label Dist	Sex Dist	Age Dist
1	Teacher	0.636	0.063	0.35
	No KD	-0.026	0.026	0.001
	Response KD	-0.003	0.039	0.015
	Feature KD	-0.006	0.037	-0.021
	Attention KD	-0.008	0.033	0.02
2	Teacher	0.339	0.215	0.159
	No KD	-0.047	-0.056	-0.013
	Response KD	-0.012	-0.072	-0.019
	Feature KD	0.035	-0.043	0.016
	Attention KD	-0.02	-0.048	-0.024
3	Teacher	0.309	0.145	0.121
	No KD	-0.086	0.024	0.03
	Response KD	-0.107	-0.005	0.059
	Feature KD	-0.009	0.045	0.071
	Attention KD	-0.08	-0.011	0.007
4	Teacher	0.206	0.107	0.17
	No KD	-0.028	0.015	-0.021
	Response KD	-0.008	0.038	-0.01
	Feature KD	-0.034	-0.018	0.01
	Attention KD	-0.044	-0.008	-0.032

Table 4.4: JS distance between attribute-specific marginal distributions (e.g., sex: male vs. female) for the first 4 PCA modes on the **Ham10000** test set. The values represent the fair ResNet101 teacher's distance and the average difference to it for students across different KD methods, averaged over all data compositions and seeds. Green and orange signify distances that are more aligned with the teacher than no-KD, with green being the closest of the three. Red indicates that no-KD distances are better aligned. Positive and negative signs denote higher and lower separation, respectively, compared to the teacher.

Data	KD	Similarity Boost		Performance Boost	
		ResNet18	ResNet34	ResNet18	ResNet34
No Female	Response	94.29%	93.17%	3.30%	4.88%
	Feature	89.53%	92.65%	2.89%	4.34%
	Attention	92.22%	84.44%	2.97%	3.98%
No Old	Response	79.75%	73.07%	1.81%	2.13%
	Feature	72.00%	79.88%	1.97%	1.95%
	Attention	67.17%	86.78%	1.09%	1.57%
No White	Response	94.34%	92.54%	3.72%	2.90%
	Feature	91.18%	87.06%	3.54%	2.47%
	Attention	74.68%	84.70%	2.90%	2.38%

Table 4.5: CheXpert capacity gap subgroup similarity and performance boost for different KD methods with varied unfair data compositions. Similarity boost measures the change in Euclidean distance between student and teacher subgroup metrics (e.g., Male AUC, Female AUC) post-KD. Performance boost denotes the student model's subgroup metric enhancement after KD. Green signifies a superior boost between Response and Feature KD.

Regarding PCA, feature-based methods, particularly Feature KD, aligned more with the teacher across datasets, as depicted by green and orange colour coding in Table 4.3 and Table 4.4. In CheXpert, Feature KD outperformed the no-KD method in 12 out of 16 cases and was the top method 8 times (indicated in green). In a parallel comparison, Attention KD exceeded in 11 out of the 16 scenarios, yet was the top performer only twice. In Ham10000, Response KD excelled in label alignment (3 out of 4 cases) but struggled with certain subgroups. Specifically, age was a challenge, where KD training often separated patients based on age to a greater degree than baseline models. Furthermore, our PCA results indicate that all KD methods, especially the feature-based ones, often reduced subgroup distribution distances compared to the teacher (indicated by a negative difference). This is also consistent for labels in 7 out of 8 cases, however; in their case, a higher separation would have been preferable and would most likely translate to better performance. This contrasts with the no-capacity gap findings in Table 4.1, where Feature KD typically exhibited greater subgroup and label separation.

Evaluation

The conflicting results across the two datasets underscore the complexity of the capacity gap issue. It appears that the behaviour of KD methods is not universal and might be influenced by dataset-specific characteristics such as subgroup separability or modality. One potential interpretation is that student models need to match the overall performance of the teacher to mimic its disparities. In the CheXpert dataset, where there's a high capacity gap but performance parity is achieved, other metrics seem to follow. In contrast, the Ham10000 dataset shows reduced similarity when performance does not align.

The consistent observation that Attention KD often introduces higher subgroup disparities than Response KD is noteworthy. A potential hypothesis is that Attention KD, having two

Data	KD	Similarity Boost		Performance Boost	
		ResNet18	ResNet34	ResNet18	ResNet34
No Female	Response	8.44%	15.14%	0.57%	0.62%
	Feature	21.07%	18.61%	1.29%	0.70%
	Attention	25.69%	10.87%	1.84%	0.62%
No Old	Response	21.96%	38.50%	1.19%	2.16%
	Feature	27.10%	36.20%	1.23%	1.98%
	Attention	24.24%	48.37%	1.29%	2.75%

Table 4.6: Ham10000 capacity gap subgroup similarity and performance boost for different KD methods with varied unfair data compositions. Similarity boost measures the change in Euclidean distance between student and teacher subgroup metrics (e.g., Male AUC, Female AUC) post-KD. Performance boost denotes the student model's subgroup metric enhancement after KD. Green signifies a superior boost between Response and Feature KD.

sources of possible bias (logits and spatial attention learned through the teacher), might amplify disparities over Response KD. However, this remains speculative. The lack of additional gains for ResNet18 over ResNet34 in feature-based KD might suggest that while addressing the capacity gap, the techniques still favour bigger-sized models.

All KD methods often made subgroup and label distributions closer to each other compared to the teacher. This indicates that in capacity gap setting models might not separate targets as effectively, which may be the reason for less pronounced subgroup separations. Alternatively, feature-based methods might force models to attend similarly to all data points, potentially reducing bias. Nevertheless, there's a noticeable discrepancy between subgroup disparities and PCA similarity, and consistent with our previous findings from section 4.2 the relationship between those aspects remains questionable.

In summary, we confirm that the capacity gap lowers the effectiveness of KD methods. However, the analysis is clearly influenced by various entangled factors, including dataset characteristics, model architectures, and the specific method employed and its implementation. These make it challenging to pinpoint specific reasons for observed performance tendencies. While certain patterns emerge, drawing definitive conclusions is complex. It's also worth noting that the Attention KD was not extensively fine-tuned, potentially impacting its effectiveness.

4.5 Does Distance Between Marginal Distributions in PCA Correlate with Subgroup Disparities?

In previous experiments, we noticed that fair models sometimes displayed unexpected differences in PCA modes. For instance, fairer teacher models often showed higher attribute separability than unfair students. This leads us to a question: Does the distance between marginal distributions after PCA correlate with subgroup performance disparities? In this study, we aim to measure this relationship.

Experimental Set Up

We train models ranging from ResNet18 to ResNet101 using the original CheXpert data splits across five distinct seeds. Our preference for CheXpert over Ham10000 is driven by more consistent results in previous experiments. For each of these models, we extract embeddings from the penultimate layer and apply PCA. Leveraging Kernel Density Estimation, we then compute the marginal distributions of specific subgroups for each PCA mode. To quantify the divergence between pairs of these distributions, we utilise the JS distance. For simplicity, we call this 'PCA Distance'. With these, we create a correlation heatmap that illustrates the relationship between the differences in marginal distributions for attributes, like age, and related subgroup disparities, such as the age AUC^{GAP}. Lastly, we determined the average difference in marginal distributions across various PCA modes to see if any additional patterns emerged.

Results

The correlation heatmap for AUC^{GAP} is shown in Figure 4.12, with additional F1-score and Youden's J Index heatmaps located in section A.8. Generally, a positive correlation implies that as the PCA distance between marginal distributions increases, so does the performance disparity. For the AUC metric, this correlation is notably weak, showing a mild positive trend. However, it is not uniform across all attributes. For instance, a larger PCA distance for biological sex underlines its disparities, but this correlation weakens with each subsequent PCA mode. In contrast, the race attribute mainly shows a negative correlation, albeit weak. The age subgroups stand out with distinct correlations in two primary PCA modes: negative for mode 1 and positive for mode 2. The Youden Index heatmap appears to mirror the general AUC trend. However, the F1-score heatmap shows a reversal in the strong age correlation and a notable positive correlation for race. Lastly, the cumulative PCA distances do not yield substantial new insights. Figure 4.13 provides a graphical illustration of the correlations, positioning each model's metric combination on a scatter plot. We can see that age in PCA modes 1 and 2 is the only visible relationship between data points, confirming the overall weak trend.

Evaluation

Generally, results highlight a weak correlation for the AUC metric, suggesting that it may not reliably indicate performance disparities in relation to PCA distances for various subgroups. This underscores the complexity of fairness in machine learning models: shifts in feature space might not always equate to performance disparities, and significant AUC gaps do not necessarily reflect shifts in the PCA space. Furthermore, the correlation is inconsistent across different attributes, which underlines the importance of a detailed, attribute-specific analysis when evaluating performance disparities. For example, the observed results for race subgroups might explain its overall lower disparities and insignificant distribution shifts observed in previous experiments. On the other hand, the age attribute, with its volatile and inconsistent changes, aligns with our earlier findings where age consistently posed challenges in achieving fairness.

The fact that the Youden Index largely mirrors the AUC's trends is not surprising given that both metrics are derived from the ROC curve, capturing the trade-offs between true and false positive rates. However, the differences seen in the F1-score heatmap, particularly for age, suggest that various metrics may capture unique aspects of performance disparities.

In summary, the distance between marginal distributions in PCA does not consistently correlate with subgroup performance disparities. Furthermore, the varied correlations across attributes and metrics highlight the importance of a comprehensive approach when evaluating performance disparities. This underscores the multifaceted nature of fairness in ML.

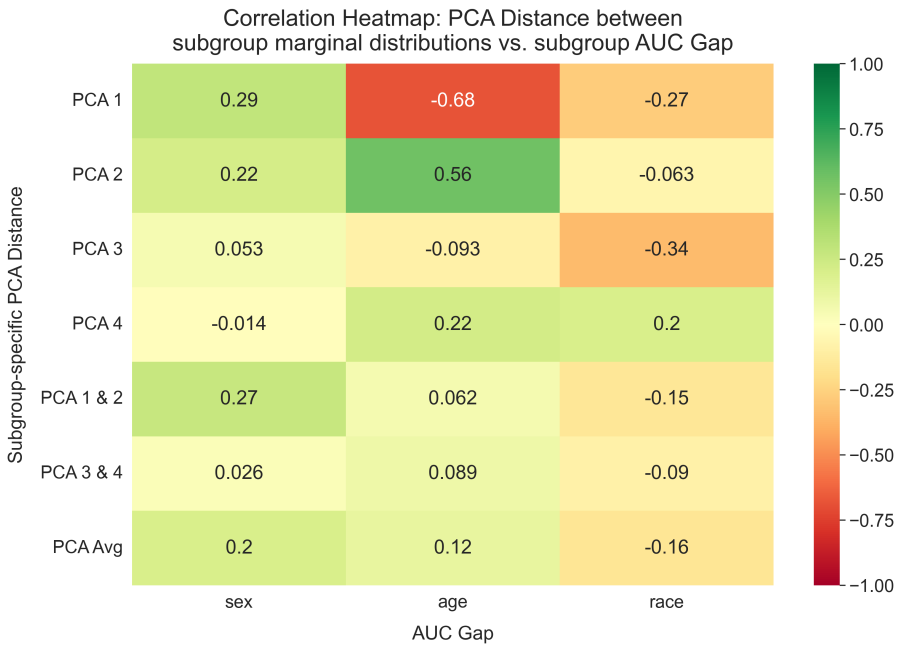

Figure 4.12: Heatmap visualising the correlation between the PCA distributional differences in subgroups, captured by JS distance, and their associated performance disparities, represented by the AUC gap.

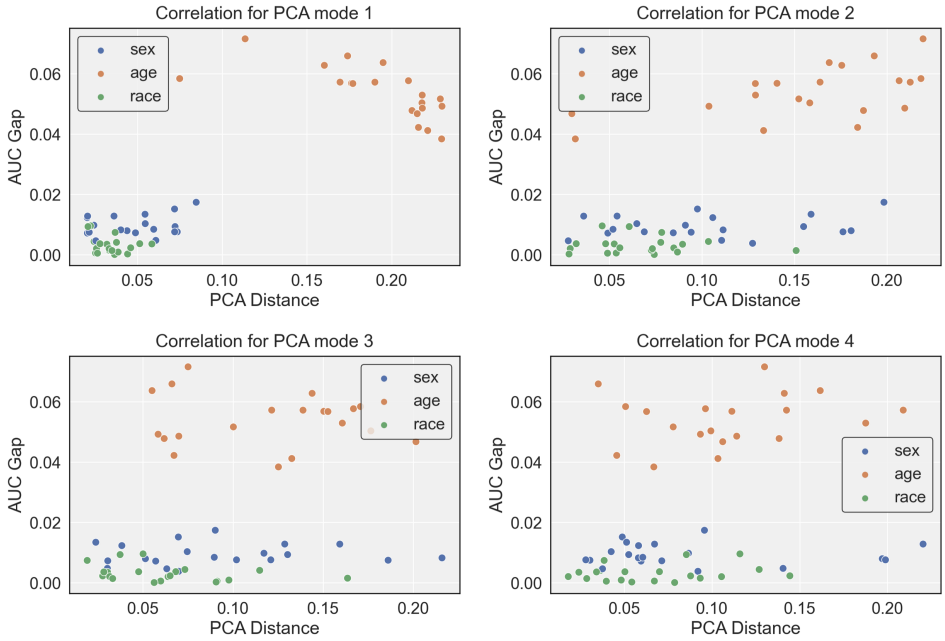

Figure 4.13: Relationship between PCA distributional differences in subgroups and AUC Gap for different PCA modes, with each subgroup attribute differentiated by colour.

4.6 What is the Influence of the Alpha Parameter on the Effectiveness of Fair Distillation?

In ML, hyperparameter tuning is often geared towards optimizing performance metrics. Yet, the pursuit of optimal performance can sometimes overshadow the importance of group fairness. Our initial hyperparameter search showed consistent accuracy across various alpha values. This observation prompted us to investigate the influence of alpha on subgroup disparities.

Experimental Setup

For this experiment, we employed previously used fair ResNet34 teacher models from both the Ham10000 and CheXpert datasets and one ResNet18 student trained on all unfair data compositions. Our focus was on Response KD, and we tested nine different values of alpha, ranging from 0.1 to 0.9. The alpha parameter represents the weighting factor on the cross-entropy loss, while $1-\alpha$ determines the weight of the distillation loss. We examine both changes in overall performance through AUC and average subgroup disparities through AUC^{GAP}.

Results

Our analysis of how the alpha parameter affects KD produced varying results between the two datasets. For CheXpert in Figure 4.14 a consistent pattern emerged across all tested unfair compositions. As the reliance on the teacher decreased (i.e., as alpha values increased), the student's performance and subgroup disparities gradually diverged from the teacher's. Specifically, for alpha values below 0.5, there was minimal variance in performance, with metrics remaining relatively stable. Conversely, the Ham10000 dataset presented a more nuanced picture, as illustrated in Figure 4.15. While the performance for $\alpha < 0.5$ remained consistent across the dataset compositions, the subgroup disparities did not follow suit. Notably, when $\alpha > 0.5$, the No Old training displayed an elevated performance, alongside high fluctuations in metrics. It shows a closer alignment to the teacher's behavior, despite a reduced reliance on it.

Evaluation

The experiment underscores the alpha parameter's pivotal role in KD. For CheXpert, an $\alpha = 0.5$ was identified as a threshold, beyond which, the teacher's impact on the student's performance and fairness diminished. This suggests that while the precise alpha value might be flexible (we see similar performance for $\alpha < 0.5$), it should ideally remain below this threshold if our goal is to maximise the teacher's guidance. For the Ham10000 dataset, the results were more complex. The teacher's influence on the student's metrics decreased with rising alpha values, but the effect varied depending on data compositions. This variability, marked by sudden changes in performance metrics, indicates that higher alpha might introduce unpredictability, which could be beneficial when the objective is to derive diverse student models from a single teacher.

In summary, these varied results are consistent with our earlier experiments, emphasizing the need to account for dataset-specific nuances when training models. We stress the significance of careful hyperparameter tuning to ensure both optimal performance and fairness.

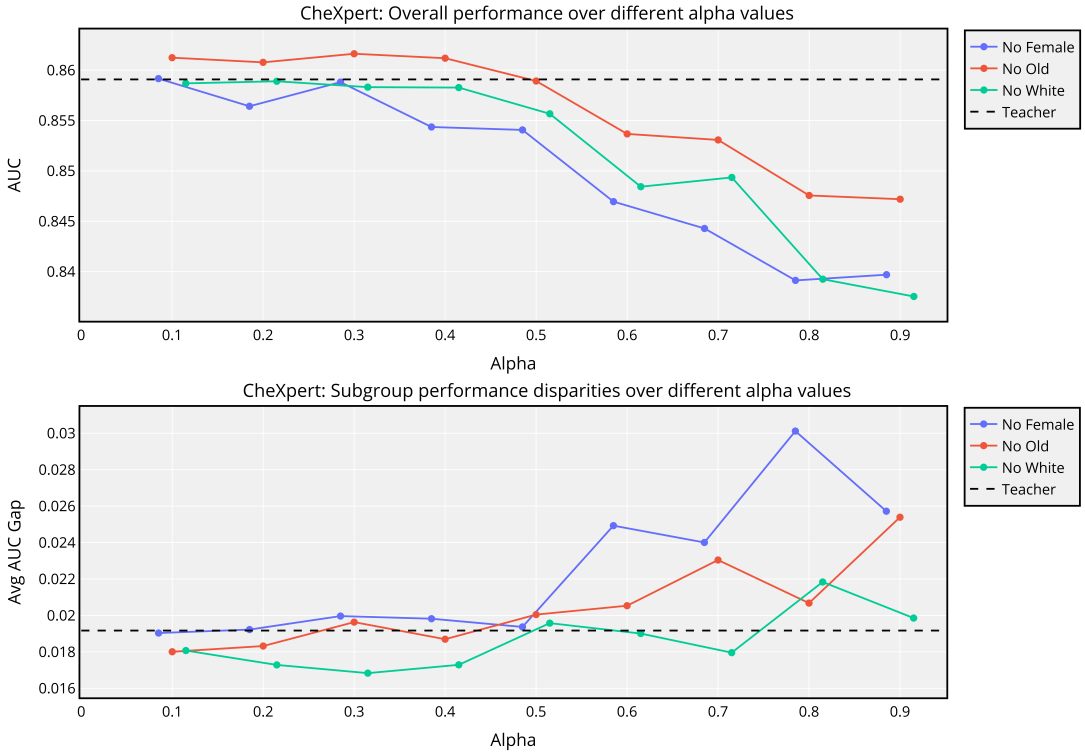

Figure 4.14: ResNet18 student performance (top) and subgroup disparities (bottom) for alpha values, trained with Response KD from a ResNet34 teacher on unfair **CheXpert** sets.

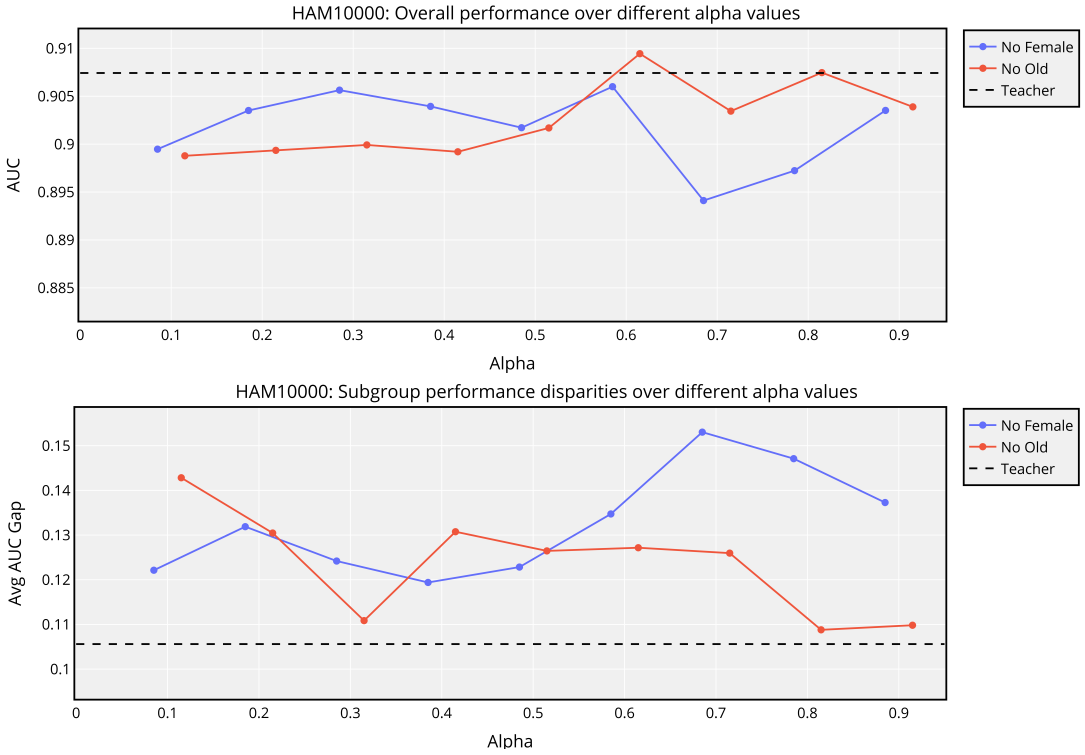

Figure 4.15: ResNet18 student performance (top) and subgroup disparities (bottom) for alpha values, trained with Response KD from a ResNet34 teacher on unfair **Ham10000** sets.

Chapter 5

Conclusions & Outlook

In this study, we investigated knowledge distillation techniques in medical imaging, focusing on how student models mimic their teacher models in an offline training setting. Throughout it, we formulated six research questions, the conclusions of which are summarised below:

Q1: Can KD from a fairer teacher help in training originally unfair students?

Answer: Yes, during KD, students trained on biased data can closely emulate fairer teacher and improve both their performance and fairness.

Q2: What patterns emerge in the unsupervised feature space for response-based KD?

Answer: After response-based KD, student models exhibit varied distributions compared to teacher, with fairer models often separating subgroups more distinctly than unfair ones (predominantly by age), making the analysis challenging.

Q3: Does using Feature KD lead to different distillation outcomes?

Answer: While very similar to Response KD in terms of metrics, Feature KD enables student models to sometimes improve over the teacher and align with its internal representations, making it better for replicating its inner structure.

Q4: How does capacity gap influence the effectiveness and behaviour of KD methods?

Answer: In a capacity-gap setting, KD's effectiveness diminishes, but outcomes and patterns are heavily influenced by specific datasets and methods, making them hard to untangle.

Q5: Does distance between marginal distributions in PCA correlate with subgroup disparities?

Answer: The overall correlation is weak, with the relationship differing by attributes and metrics, suggesting that shifts in PCA space may not reflect performance disparities.

Q6: What is the influence of the alpha parameter on the effectiveness of fair distillation?

Answer: Alpha's value proportionally affects the student-teacher resemblance; yet, dataset-specific outcomes underscore the importance of rigorous medical dataset analysis.

Overall, our findings indicate that the alignment between teacher and students in most cases is not limited to just performance metrics but also includes fairness and unsupervised embedding similarity. While we were originally motivated by [18] that suggested bias amplification through KD, we argue that student models, when appropriately trained, are not inherently predisposed to shortcut learning. Instead, they tend to acquire the characteristics of their teacher models. This observation is particularly notable in our experiments with a fair teacher and biased student models, where KD, to some extent, managed to overcome the bias of data. Given the growing interest in foundation models and the challenges of sourcing quality data, our findings show a promising avenue for KD's applications. However, we also confirmed that a significant difference in model capacity can negatively impact the distillation process, even when using strategies specifically designed to address this issue. Therefore, careful distillation technique and model selection are essential when deploying KD.

In our results, we observed patterns and differences between tested KD methods. For instance, there was a reduction in subgroup disparities with Feature KD (section 4.3), the highest subgroup performance similarity boost by Response KD (section 4.3), and bias amplification of Attention KD (section 4.4). Moreover, we have seen that all KD methods exhibited closer marginal distributions in PCA space than their teacher (section 4.4). However, identifying the precise causes for such specific behaviours was complex due to multiple variables at play, including model architecture, metrics, training techniques, sampling methods, weighting on the teacher model, or label noise. Hence, our initial hypotheses remain rather speculative and further experiments would have to be done to pinpoint the exact reasons for such behaviour.

One of the consistent patterns observed throughout our studies was varying results for the CheXpert and Ham10000 datasets. Echoing recent literature [81], we believe that factors such as subgroup separability and dataset modality might explain the discrepancies between them. We advocate for a careful understanding of the dataset's domain both during model evaluation and their subsequent deployment.

In summary, ensuring and evaluating fairness in AI is a complex problem. Relying solely on metrics can provide a limited perspective, potentially overlooking the broader context and introducing biases associated with specific measurements. Moreover, in light of the observed divergence between subgroup disparities and their PCA marginal distributions, it is important to understand the underlying relationship between common fairness metrics and model inspection methods. With the widespread of AI models in local devices [89] and the promise of KD in the healthcare domain [20, 21] we aimed to bring awareness and guide future research on how knowledge distillation and fairness work together.

5.1 Limitations

In this project, we focused on the fair teacher and unfair student KD setup. Although we identified the potential of KD to address biases in training techniques, both the student and teacher models were trained using overlapping data subsets. For a more comprehensive evaluation, the teacher should have been trained on a distinct dataset from the student. The teacher would then be exposed to the so-called domain adaptation, which is an open challenge in the ML community. We also recognize that our designation of fair and unfair models might be somewhat

arbitrary, and their naming conventions are not fully realised. However, the project's objective was to compare models with significant differences in fairness metrics after applying KD, not to achieve perfect group fairness.

Moreover, in our experiments, we utilise AUC as both the performance and fairness metric in the form of AUC^{GAP}. While keeping the same measurement offers a consistent view and better relative comparisons, we acknowledge that introducing additional metrics for performance and fairness would have given a broader understanding of the system's behaviour in various contexts.

Our study's binary approach for labels and sensitive attributes is another limitation. While binary classification is common in disease prediction models, it may not fully exploit the potential of soft logits in KD. Moreover, many studies adopt a binary setup for sensitive attributes, but exploring beyond this might offer more nuanced insights. Lastly, we used random weighted sampling in our training. Although relatively standard, this could influence the models' outcomes and subsequently, our final results.

5.2 Future Work

Firstly, future work can address the limitations of our approach. Evidently, a transition from a binary setting to a non-binary one for both labels and sensitive attributes would provide a more comprehensive evaluation of KD. Additionally, incorporating a broader range of metrics would enhance the robustness of our results and provide deeper insights.

An interesting direction for subsequent research is to stress-test KD training under suboptimal conditions. Exploring different model architectures would reveal how the structural nuances of both teacher and student models impact the distillation process. Investigating domain shifts, where input data distribution changes, will show KD's adaptability and better realise our promise of an unfair student and fair teacher dynamic. Training the model on historical data and the student on current data would simulate data distribution changes over time. By probing these scenarios, we can establish KD's boundaries and identify methods to boost its performance under challenging conditions.

Furthermore, several open questions could benefit from additional experimentation. The visual examination of Attention KD in an unsupervised space could shed light on its behaviour. Similarly, a deeper dive into how different sampling techniques influence the overall results in terms of fairness and performance would be valuable. While we excluded the unsupervised prediction layer information test (SPLIT) from our analysis, we are curious if its results would get indirectly distilled through KD.

Moreover, the concept of fairness optimisation presents a promising avenue. By employing more advanced techniques to develop fair and unfair networks, we can establish a scale of models of different fairness. One could investigate how the behaviour of KD changes with varying degrees of fairness. Determining the threshold at which KD's effectiveness fades away would be also insightful. Additionally, this would pave the way for examining the dynamics of a fair student and an unfair teacher scenario, building on initial studies such as [19].

5.3 Ethical Considerations

In our pursuit to understand fairness in KD within medical models, we have primarily approached the challenge from a technical perspective. However, it is crucial to recognise that any attempt to address fairness from this point only is doomed to fail. Achieving genuine fairness in AI is an interdisciplinary endeavour that needs collaboration from developers, ethicists, sociologists, domain experts, and more. It is not enough to optimise algorithms; we must also engage with a diverse set of stakeholders, especially marginalised communities, to ensure a holistic and inclusive perspective on fairness. By doing so, we can better understand the multifaceted nature of fairness and work towards solutions that are both technically sound and ethically responsible.

Furthermore, while our study underscores the potential of KD, it is important to recognise the risks associated with it. Our research predominantly focused on a fair teacher guiding an unfair student. However, the inverse scenario, where an unfair teacher passes on biases to the student model, is equally if not more likely. Additionally, if the AI community heavily relies on a few popular teacher models to distil numerous students, it can lead to a lack of diversity in DL models and amplification of biases and errors. It is also worth noting that the very nature of KD, which involves compressing large networks, might strip away vital information, leading to more opaque AI models.

Lastly, all data utilised in this study is publicly accessible. The Ham10000 dataset, which contains patient information, can be accessed at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T. The CheXpert dataset is available at https://stanfordmlgroup.github.io/competitions/chexpert/.

Bibliography

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[3] Jian Huang, Junyi Chai, and Stella Cho. Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1):1–24, 2020.

[4] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[5] Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging-a mini review. In *Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021*, pages 33–38. Springer, 2022.

[6] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[7] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[8] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Chen, and Marzyeh Ghassemi. Medical imaging algorithms exacerbate biases in underdiagnosis. 2021.

[9] Paul H Yi, Jinchi Wei, Tae Kyung Kim, Jiwon Shin, Haris I Sair, Ferdinand K Hui, Gregory D Hager, and Cheng Ting Lin. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emergency Radiology*, 28:949–954, 2021.

[10] Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, et al. Reading race: Ai recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*, 2021.

[11] Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7370–7379, 2017.

[12] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4820–4828, 2016.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

[15] Jason Ramapuram, Dan Busbridge, and Russ Webb. Evaluating the fairness of fine-tuning strategies in self-supervised learning. *arXiv preprint arXiv:2110.00538*, 2021.

[16] Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *Ebiomedicine*, 89, 2023.

[17] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in Neural Information Processing Systems*, 34:6906–6919, 2021.

[18] Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of distilbert. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, 2022.

[19] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? *arXiv preprint arXiv:2205.16004*, 2022.

[20] Hefeng Meng, Zhiqiang Lin, Fan Yang, Yonghui Xu, and Lizhen Cui. Knowledge distillation in medical data mining: a survey. In *5th International Conference on Crowd Science and Engineering*, pages 175–182, 2021.

[21] Alankar Mahajan and Aruna Bhat. A survey on application of knowledge distillation in healthcare domain. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 762–768. IEEE, 2023.

[22] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[23] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[24] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[25] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[26] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019.

[27] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

[28] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019.

[29] Emanuel Ben-Baruch, Matan Karklinsky, Yossi Biton, Avi Ben-Cohen, Hussam Lawen, and Nadav Zamir. It's all in the head: Representation knowledge distillation through classifier sharing. *arXiv preprint arXiv:2201.06945*, 2022.

[30] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

[31] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Learning student networks via feature embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1): 25–35, 2020.

[32] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701, 2017.

[33] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[34] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[35] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3430–3437, 2020.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[37] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[38] Minh Pham, Minsu Cho, Ameya Joshi, and Chinmay Hegde. Revisiting self-distillation. *arXiv preprint arXiv:2206.08491*, 2022.

[39] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019.

[40] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.

[41] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[42] Zengyu Qiu, Xinzhu Ma, Kunlin Yang, Chunya Liu, Jun Hou, Shuai Yi, and Wanli Ouyang. Better teacher better student: Dynamic prior knowledge for knowledge distillation. *arXiv preprint arXiv:2206.06067*, 2022.

[43] Huan Tian, Tianqing Zhu, Wei Liu, and Wanlei Zhou. Image fairness in deep learning: problems, models, and challenges. *Neural Computing and Applications*, 34(15):12875–12893, 2022.

[44] Mark Sendak, Michael Gao, Marshall Nichols, Anthony Lin, and Suresh Balu. Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMs*, 7(1), 2019.

[45] Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P Georgiou, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5):584–595, 2020.

[46] Richard J Chen, Tiffany Y Chen, Jana Lipkova, Judy J Wang, Drew FK Williamson, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603*, 2021.

[47] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[48] Zikang Xu, Jun Li, Qingsong Yao, Han Li, Xin Shi, and S Kevin Zhou. A survey of fairness in medical image analysis: Concepts, algorithms, evaluations, and challenges. *arXiv preprint arXiv:2209.13177*, 2022.

[49] Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *arXiv preprint arXiv:2305.06969*, 2023.

[50] KBRK Ramesha, KB Raja, KR Venugopal, and LM Patnaik. Feature extraction based face recognition, gender and age classification. *International Journal on Computer Science and Engineering,*, 2:14–23, 2010.

[51] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[52] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[53] Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1):40, 2021.

[54] Jessica Zosa Forde, A Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman. Model selection's disparate impact in real-world deep learning applications. *arXiv preprint arXiv:2104.00606*, 2021.

[55] Schrasing Tong and Lalana Kagal. Investigating bias in image classification using model explanations. *arXiv preprint arXiv:2012.05463*, 2020.

[56] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3662–3666. IEEE, 2020.

[57] Julien Rouzot, Julien Ferry, and Marie-José Huguet. Learning optimal fair scoring systems for multi-class classification. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 197–204. IEEE, 2022.

[58] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[59] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[60] Silvio Amir, Jan-Willem van de Meent, and Byron C Wallace. On the impact of random seeds on the fairness of clinical classifiers. *arXiv preprint arXiv:2104.06338*, 2021.

[61] Jennifer L Skeem and Christopher T Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

[62] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on Health, Inference, and Learning*, pages 204–233. PMLR, 2022.

[63] Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

[64] Youjin Kong. Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 485–494, 2022.

[65] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[66] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020, 2019.

[67] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

[68] Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization*, 2020.

[69] Valeriia Cherepanova, Vedant Nanda, Micah Goldblum, John P Dickerson, and Tom Goldstein. Technical challenges for training fair neural networks. *arXiv preprint arXiv:2102.06764*, 2021.

[70] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.

[71] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.

[72] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[73] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[74] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

[75] Ben Glocker, Charles Jones, Melanie Bernhardt, and Stefan Winzeck. Risk of bias in chest x-ray foundation models. *arXiv preprint arXiv:2209.02965*, 2022.

[76] Andrew B Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022.

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

[78] Emma AM Stanley, Matthias Wilms, Pauline Mouches, and Nils D Forkert. Fairness-related performance and explainability effects in deep learning models for brain image analysis. *Journal of Medical Imaging*, 9(6):061102, 2022.

[79] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[80] Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos D Castillo, P Jonathon Phillips, and Rama Chellappa. Distill and de-bias: Mitigating bias in face recognition using knowledge distillation. *arXiv preprint arXiv:2112.09786*, 2021.

[81] Charles Jones, Mélanie Roschewitz, and Ben Glocker. The role of subgroup separability in group-fair medical image classification. *arXiv preprint arXiv:2307.02791*, 2023.

[82] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *arXiv preprint arXiv:2106.10494*, 2021.

[83] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[84] Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, 2023.

[85] Guangxuan Xu and Qingyuan Hu. Can model compression improve nlp fairness. *arXiv preprint arXiv:2201.08542*, 2022.

[86] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9, 2018.

[87] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[88] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[89] Abdolmaged Alkhulaifi, Fahad Alsahli, and Irfan Ahmad. Knowledge distillation in deep learning and its applications. *PeerJ Computer Science*, 7:e474, 2021.

Appendix A

Supplementary Material

A.1 Response KD Alpha - Temperature Results

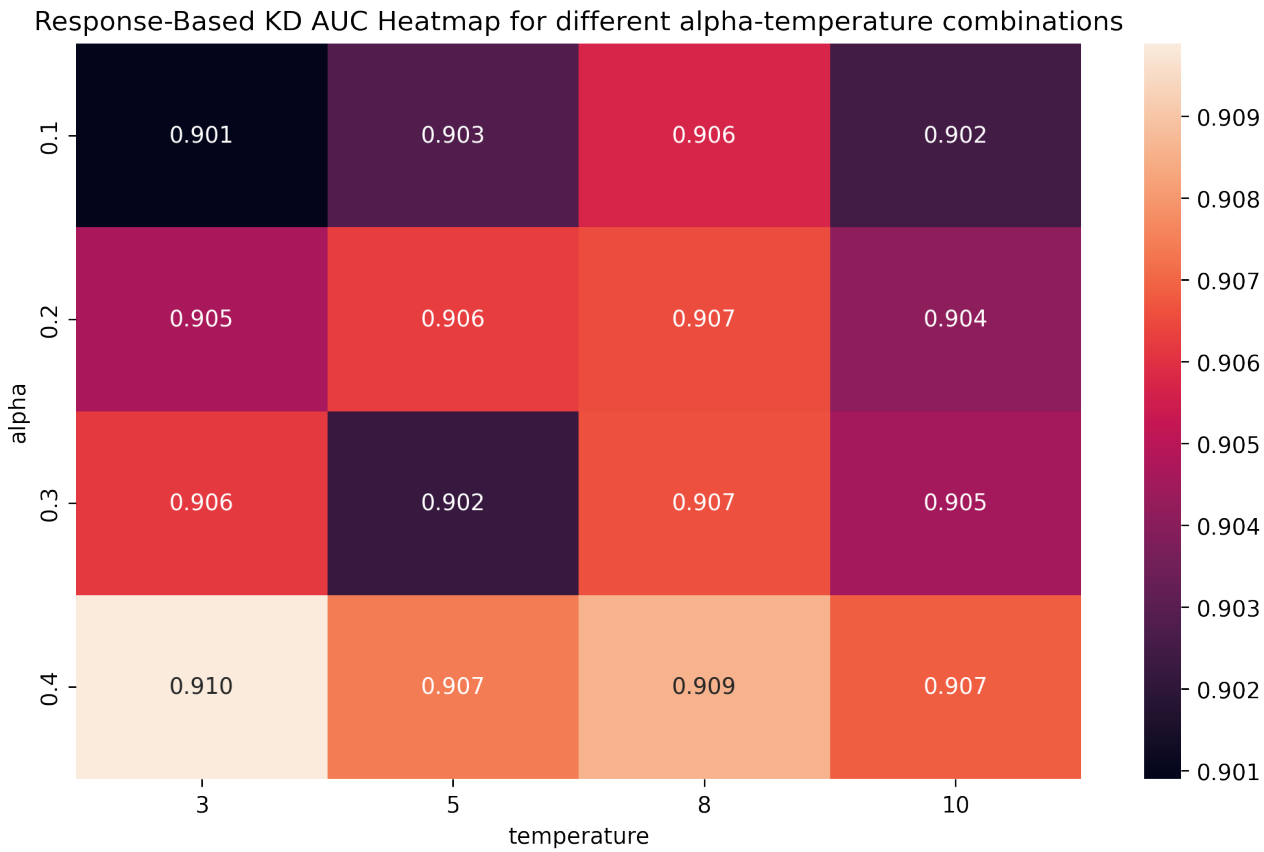

Figure A.1: Heatmap representing AUC scores for different α-T combinations, averaged over 5 random seeds for both ResNet18 and ResNet34 when trained with better performing ResNet101.

A.2 Study Population

CheXpert							
	All	White	Non-White	Male	Female	Young	Old
Attribute	All Data						
Patients	42,884	33,338	9,546	23,623	19,261	21,835	21,525
Scans	127,118	99,027 (78)	28,091 (22)	74,682 (59)	52,436 (41)	64,065 (50)	63,053 (50)
Age (years)	63 ± 17	64 ± 17	60 ± 17		64 ± 18	49 ± 12	77 ± 8
Female	52,436 (41)	39,735 (40)	12,701 (45)	-	-	25,359 (40)	27,077 (43)
White	99,027 (78)	-	-	59,292 (79)	39,735 (76)	47,921 (75)	51,106 (81)
No Finding	10,916 (9)	8,236 (8)	2,680 (10)	6,280 (8)	4,636 (9)	7,595 (12)	3,321 (5)
Other	116,202 (91)	90,791 (92)	25,411 (90)	68,402 (92)	47,800 (91)	56,470 (88)	59,732 (95)
Trainig Data							
Patients	25,730	20,034	5,696	14,164	11,566	13,087	12,931
Scans	76,205	59,238 (78)	16,967 (22)	44,773 (59)	31,432 (41)	38,294 (50)	37,911 (50)
Age (years)	63 ± 17	64 ± 17	60 ± 17	62 ± 17	64 ± 18	49 ± 12	77 ± 8
Female	31,432 (41)	23,715 (40)	7,717 (45)	-	-	15,198 (40)	16,234 (43)
White	59,238 (78)	-	-	35,523 (79)	23,715 (75)	28,609 (75)	30,629 (81)
No Finding	6,514 (9)	4,910 (8)	1,604 (9)	3,756 (8)	2,758 (9)	4,541 (12)	1,973 (5)
Other	69,691 (91)	54,328 (92)	15,363 (91)	41,017 (92)	28,674 (91)	33,753 (88)	35,938 (95)
Validation Data							
Patients	4,288	3,348	940	2,367	1,921	2,240	2,091
Scans	12,673	9,945 (78)	2,728 (22)	7,643 (60)	5,030 (40)	6,561 (52)	6,112 (48)
Age (years)	62 ± 17	64 ± 17	60 ± 17	62 ± 17	64 ± 18	49 ± 12	77 ± 8
Female	52,436 (41)	39,735 (40)	12,701 (45)	-	-	25,359 (40)	27,077 (43)
White	99,027 (78)	-	-	59,292 (79)	39,735 (76)	47,921 (75)	51,106 (81)
No Finding	10,916 (9)	8,236 (8)	2,680 (10)	6,280 (8)	4,636 (9)	7,595 (12)	3,321 (5)
Other	116,202 (91)	90,791 (92)	25,411 (90)	68,402 (92)	47,800 (91)	56,470 (88)	59,732 (95)
Testing Data							
Patients	12,866	9,956	2,910	7,092	5,774	6,508	6,503
Scans	38,240	29,844 (78)	8,396 (22)	22,266 (57)	15,974 (42)	19,210 (50)	19,030 (50)
Age (years)	63 ± 17	64 ± 17	60 ± 17	63 ± 16	64 ± 18	50 ± 12	77 ± 8
Female	15,974 (42)	12,087 (41)	3,887 (46)	-	-	7,716 (40)	8,258 (43)
White	29,844 (78)	-	-	17,757 (80)	12,087 (76)	14,359 (75)	15,485 (81)
No Finding	3,316 (9)	2,509 (8)	807 (10)	1,922 (9)	1,394 (9)	2,293 (12)	1,023 (5)
Other	34,924 (91)	27,335 (92)	7,589 (90)	20,344 (91)	14,580 (91)	16,917 (88)	18,007 (95)

Table A.1: Breakdown of demographics over the set of patient scans by sensitive subgroups and dataset splits for CheXpert. Numbers in the brackets are percentages with respect to the number of scans. Inspired by [16].

Ham10000					
	All	Male	Female	Young	Old
Attribute	All Data				
Lesions	7,418	3,998	3,416	5,513	1,905
Scans	9,958	5,400 (54)	4,548 (46)	7,159 (72)	2,799 (28)
Age (years)	52 ± 17	55 ± 17	49 ± 16	44 ± 12	73 ± 6
Female	4,548 (46)	-	-	3,624 (51)	924 (33)
Benign	8,520 (86)	4,492 (83)	4,018 (88)	6,475 (90)	2,045 (73)
Malignant	1,438 (14)	908 (17)	530 (12)	684 (10)	754 (27)
	Trainig Data				
Lesions	5,934	3,211	2,722	4,426	1,508
Scans	7,967	4,334 (54)	3,630 (46)	5,763 (72)	2,204 (28)
Age (years)	52 ± 17	54 ± 17	49 ± 16	44 ± 12	73 ± 6
Female	3,630 (46)	-	-	2,900 (50)	730 (33)
Benign	6,836 (86)	3,609 (83)	3,224 (89)	5,218 (91)	1,618 (73)
Malignant	1,131 (14)	725 (17)	406 (11)	545 (9)	586 (27)
	Validation Data				
Lesions	742	385	355	533	209
Scans	989	523 (53)	461 (47)	673 (68)	316 (32)
Age (years)	53 ± 17	57 ± 17	49 ± 16	44 ± 11	73 ± 7
Female	461 (47)	-	-	365 (54)	96 (30)
Benign	823 (83)	419 (80)	399 (87)	604 (90)	219 (69)
Malignant	166 (17)	104 (20)	62 (13)	69 (10)	97 (31)
	Testing Data				
Lesions	742	402	339	554	188
Scans	1,002	543 (54)	457 (46)	723 (72)	279 (28)
Age (years)	51 ± 18	53 ± 18	49 ± 16	43 ± 13	73 ± 7
Female	457 (46)	-	-	359 (50)	98 (35)
Benign	861 (86)	464 (85)	395 (86)	653 (90)	208 (75)
Malignant	141 (14)	79 (15)	62 (14)	70 (10)	71 (25)

Table A.2: Breakdown of demographics over the set of scans by sensitive subgroups and dataset splits for Ham10000. Numbers in the brackets are percentages with respect to the number of scans. Inspired by [16].

A.3 Detailed Response KD Models Breakdown

Ham10000

Data	Model	Seeds(s)	AUC	Age AUC Gap	Sex AUC Gap	Avg AUC Gap
Original	Teacher	46	0.907	0.103	0.005	0.054
No Female	ResNet18	41-46	0.901 (3.16)	0.125 (-4.44)	0.005 (-90.67)	0.065 (-27.18)
		46	0.902 (4.27)	0.121 (-13.18)	0.003 (-90.71)	0.062 (-29.24)
	ResNet34	41-46	0.904 (2.22)	0.114 (-11.24)	0.007 (-67.66)	0.060 (-19.12)
		42	0.900 (2.95)	0.112 (-10.27)	0.005 (-84.55)	0.064 (-24.10)
No Old	ResNet18	41-46	0.899 (2.36)	0.127 (5.51)	0.019 (-60.36)	0.073 (-13.06)
		46	0.897 (4.36)	0.139 (25.28)	0.009 (-82.15)	0.074 (-7.17)
	ResNet34	41-46	0.911 (4.47)	0.108 (-20.31)	0.006 (-68.80)	0.057 (-26.64)
		42	0.910 (3.91)	0.105 (-38.81)	0.004 (-88.88)	0.055 (-48.08)

Table A.3: Breakdown of fair ResNet34 teacher and unfair ResNet18/34 student models after Response KD with **Ham10000** data compositions. The performance in the brackets is the relative (%) increase/decrease for this particular metric compared to no-KD counterparts. Green indicates an improvement for metric. We showcase results for all 5 random seeds and single seed that was later picked for unsupervised model inspection.

CheXpert

Data	Models	Seed(s)	AUC	Age AUC Gap	Sex AUC Gap	Race AUC Gap	Avg AUC Gap
Original	Teacher	43	0.859	0.048	0.008	0.0015	0.019
No Female	ResNet18	41-46	0.859 (3.74)	0.048 (-18.89)	0.009 (-71.45)	0.001 (-86.00)	0.020 (-41.79)
		46	0.857 (7.00)	0.046 (-47.38)	0.010 (-77.67)	0.002 (-89.39)	0.019 (-61.85)
	ResNet34	41-46	0.860 (5.43)	0.049 (-34.43)	0.007 (-70.25)	0.003 (-16.87)	0.020 (-42.66)
		45	0.861 (7.41)	0.050 (-25.23)	0.009 (-79.54)	0.007 (300.30)	0.022 (-41.55)
No Old	ResNet18	41-46	0.859 (2.14)	0.049 (-21.99)	0.007 (-26.80)	0.002 (-42.33)	0.019 (-23.56)
		46	0.861 (2.77)	0.048 (-31.95)	0.006 (-13.44)	0.002 (-24.18)	0.019 (-30.20)
	ResNet34	41-46	0.861 (2.48)	0.048 (-9.49)	0.007 (-28.08)	0.002 (-52.98)	0.019 (-15.01)
		45	0.861 (2.42)	0.050 (-13.76)	0.007 (-58.53)	0.004 (13.43)	0.020 (-22.26)
No White	ResNet18	41-46	0.857 (4.20)	0.047 (-11.78)	0.006 (-35.90)	0.002 (-50.81)	0.018 (-17.75)
		46	0.859 (6.28)	0.045 (-25.90)	0.007 (-52.71)	0.000 (-94.13)	0.017 (-36.47)
	ResNet34	41-46	0.857 (3.06)	0.048 (-14.55)	0.008 (-11.26)	0.002 (25.41)	0.019 (-12.99)
		45	0.854 (2.66)	0.047 (-8.40)	0.010 (-12.03)	0.002 (-45.39)	0.019 (-10.89)

Table A.4: Breakdown of fair ResNet34 teacher and unfair ResNet18/34 student models after Response KD with **CheXpert** data compositions. The performance in the brackets is the relative (%) increase/decrease for this particular metric compared to no-KD counterparts. Green indicates an improvement for metric. We showcase results for all 5 random seeds and single seed that was later picked for unsupervised model inspection.

A.4 ResNet Architecture Details

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure A.2: Details of ResNet architecture as outlined in the original work [1]. Each ResNet scale has five main layers (under the 'layer name'), regardless of its size. When matching the feature maps in feature-based KD we skip the initial and simple 'conv1' layer and look only at 'conv2_x' - 'conv5_x' layers.

A.5 Unsupervised Model Inspection

Here, for completion we provide all unsupervised model inspection plots for the teacher, ResNet18-ResNet34 students with and without KD. We do so for all 4 PCA modes, logits and t-SNE for all unfair CheXpert and Ham10000 dataset compositions that we have trained a selected model with. Besides showing disease separation we overlay information on patient characteristics for each sensitive attribute.

A.5.1 Ham10000

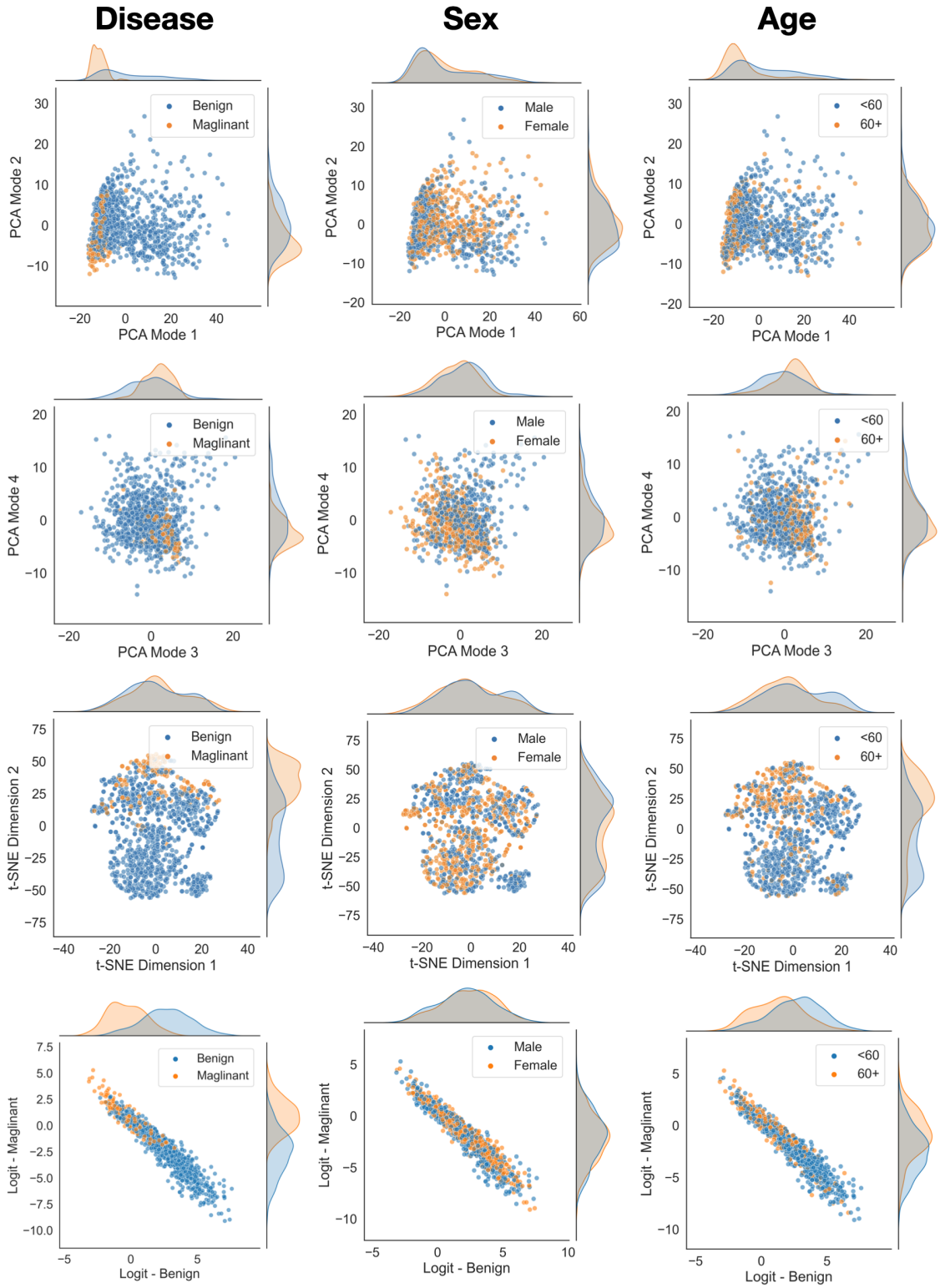

Figure A.3: PCA, t-SNE and Logit scatter plots with marginal distributions for **Ham10000** test data feature representations for fair ResNet34 **Teacher**.

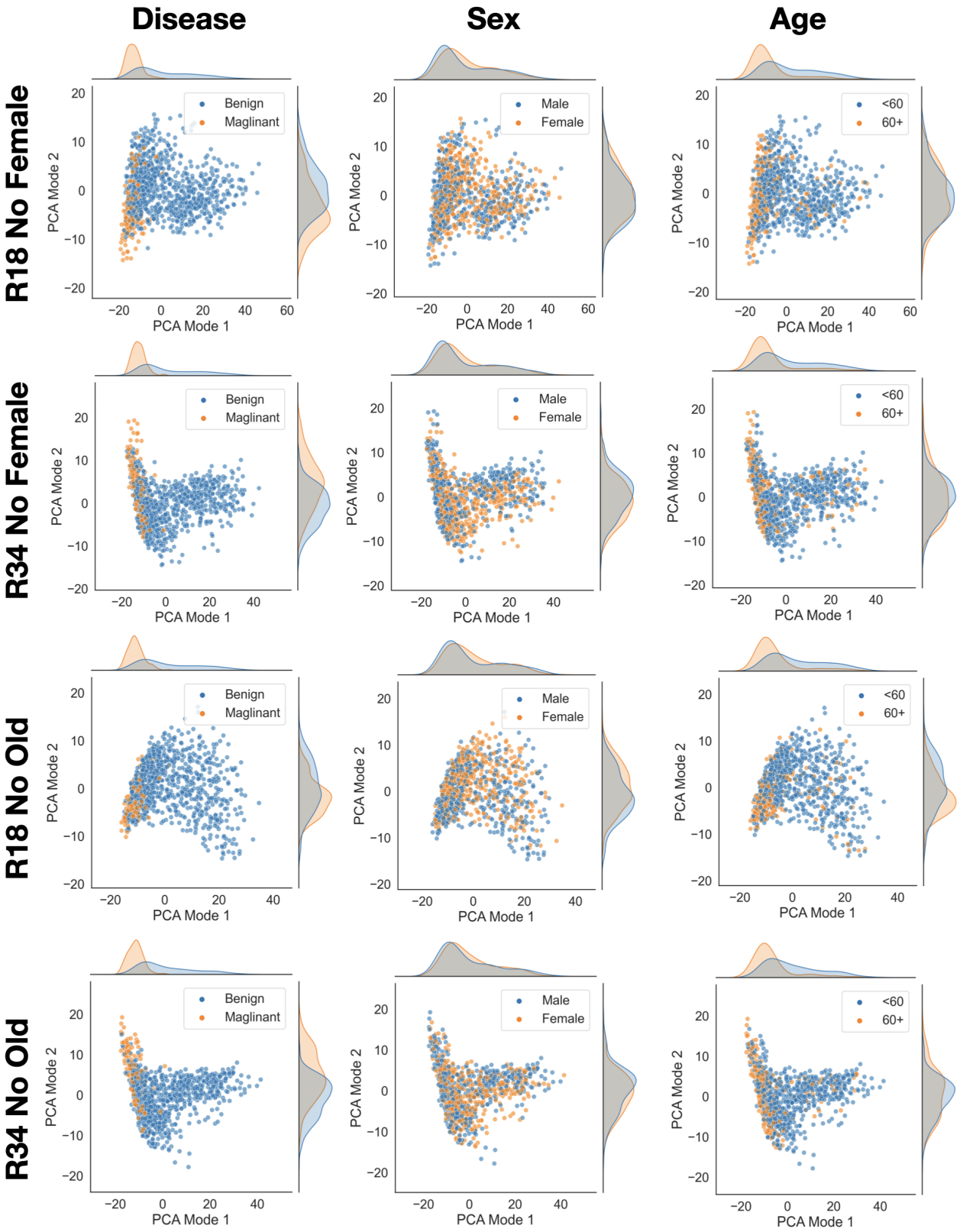

Figure A.4: PCA mode 1 & 2 students scatter plots with marginal distributions for **Ham10000** test data feature representations trained with **Response KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

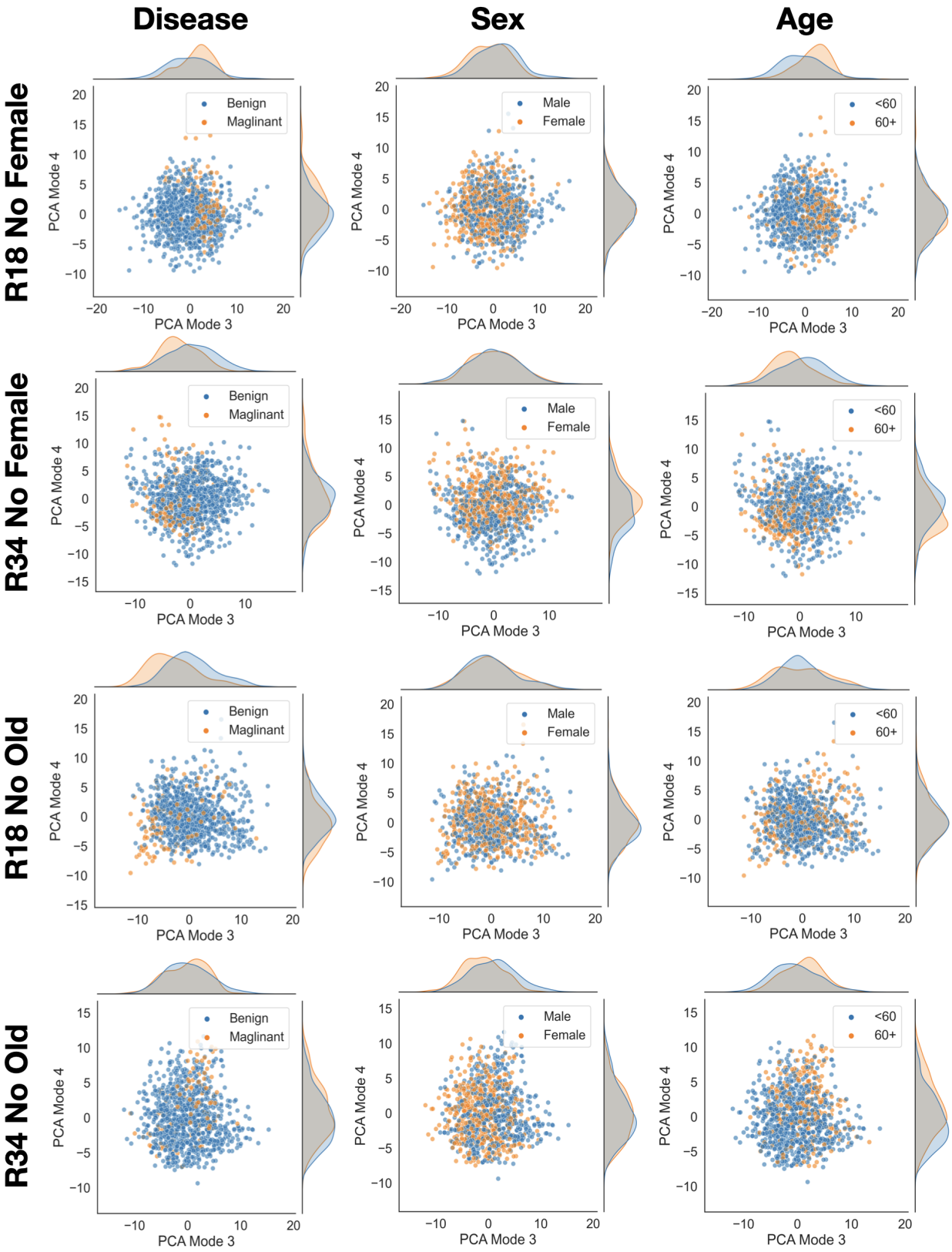

Figure A.5: PCA mode 3 & 4 students scatter plots with marginal distributions for **Ham10000** test data feature representations trained with **Response KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

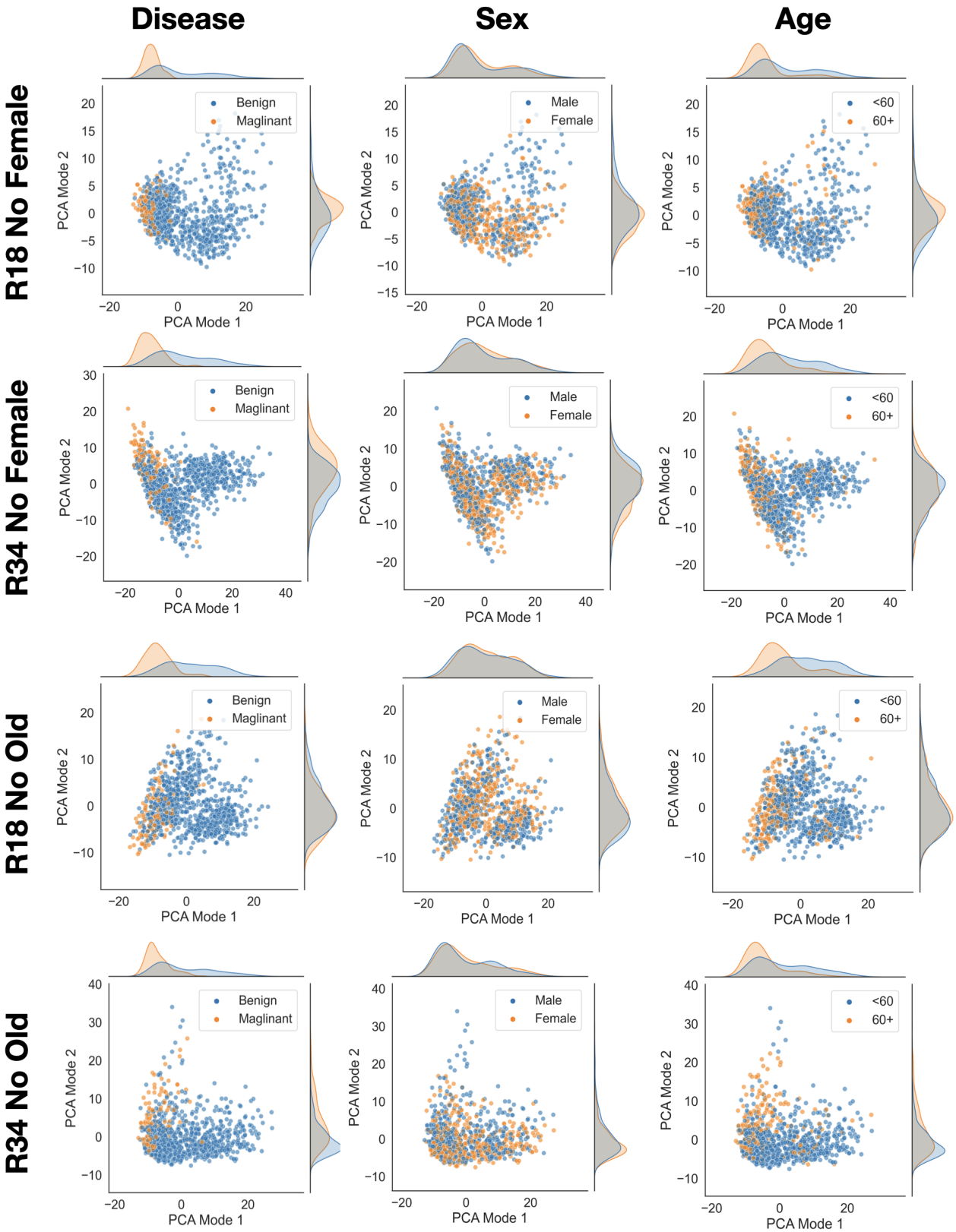

Figure A.6: PCA mode 1 & 2 students scatter plots with marginal distributions for **Ham10000** test data feature representations trained **without KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

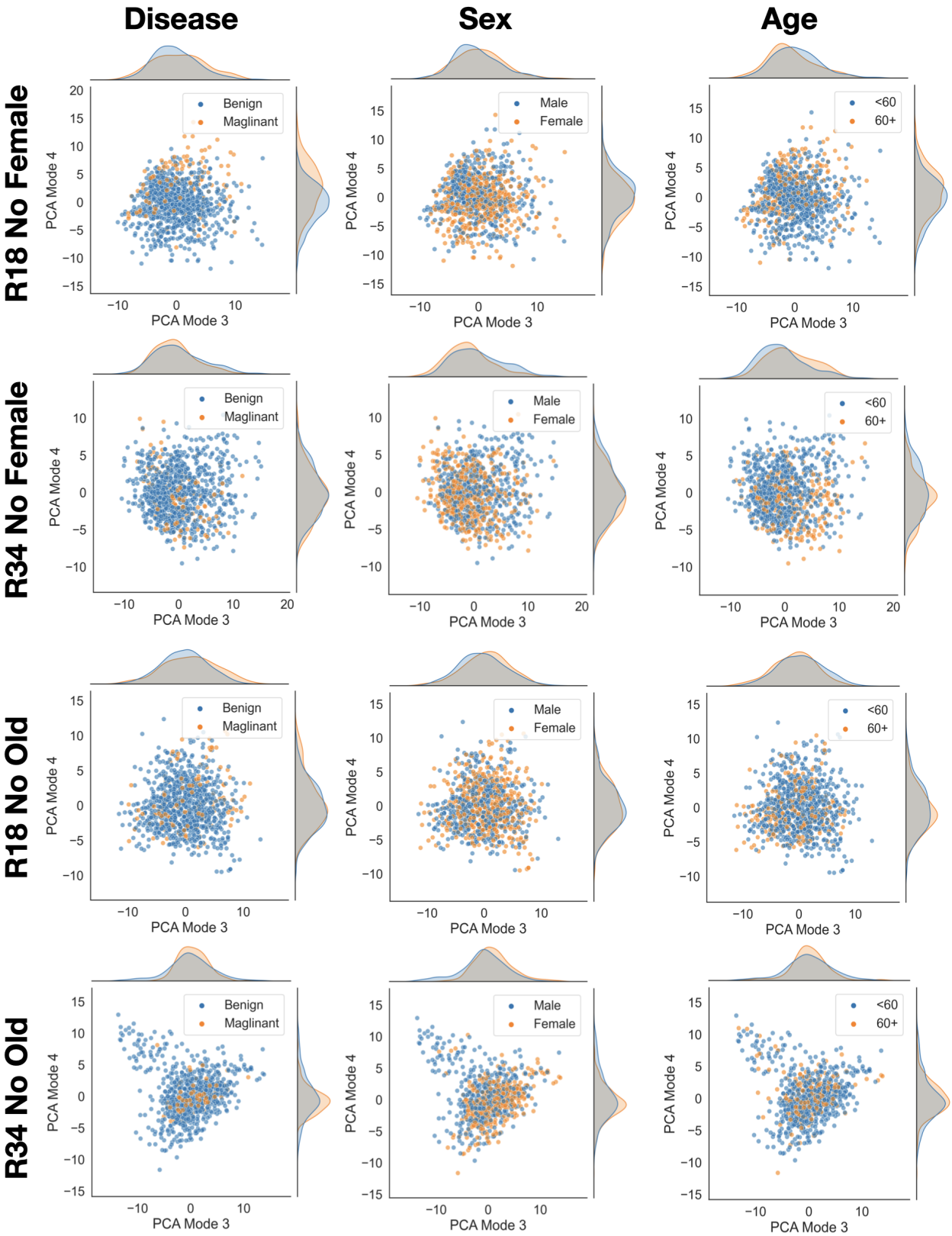

Figure A.7: PCA mode 3 & 4 students scatter plots with marginal distributions for **Ham10000** test data feature representations trained **without KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

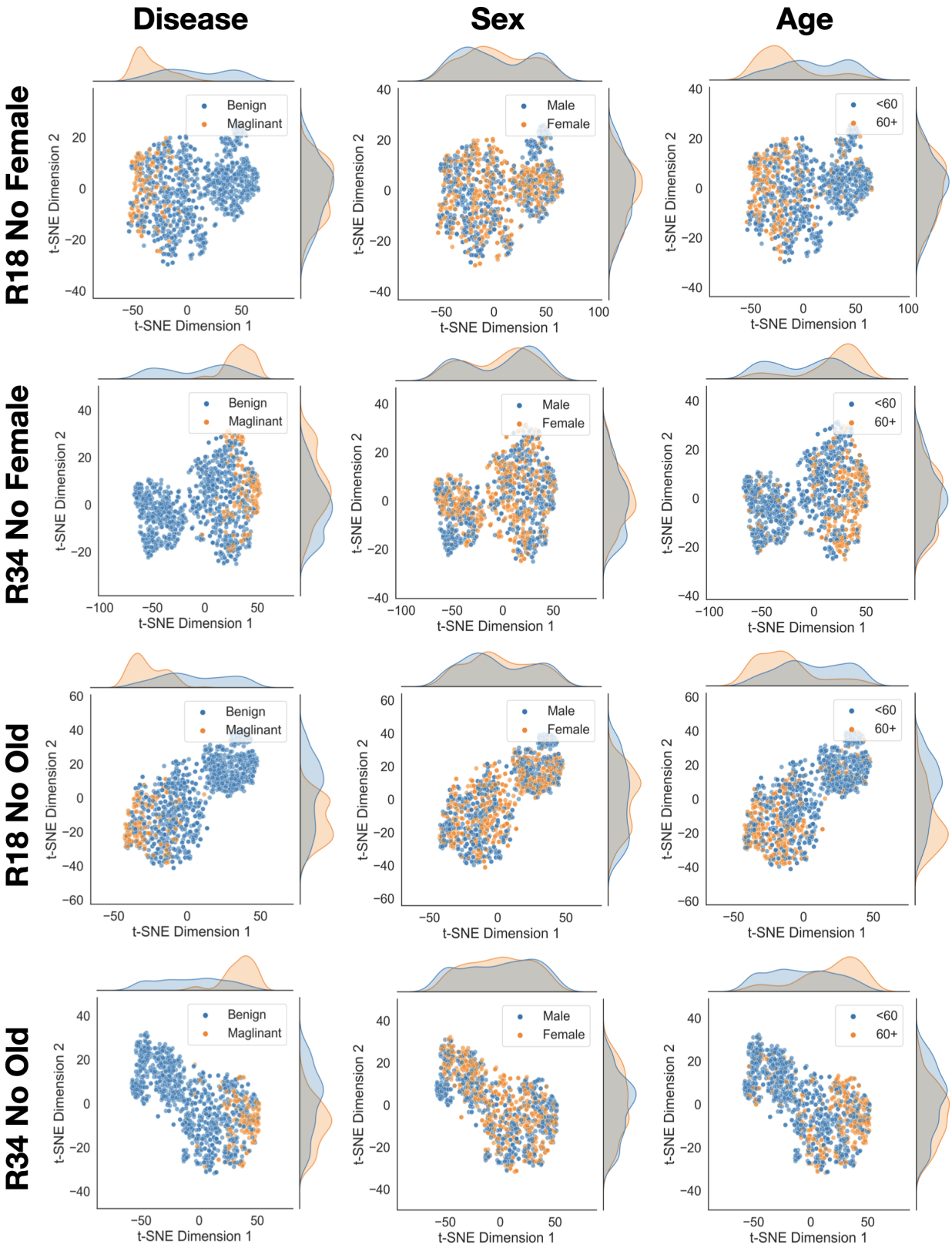

Figure A.8: t-SNE students scatter plots with marginal distributions for **Ham10000** test data feature representations trained with **Response KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

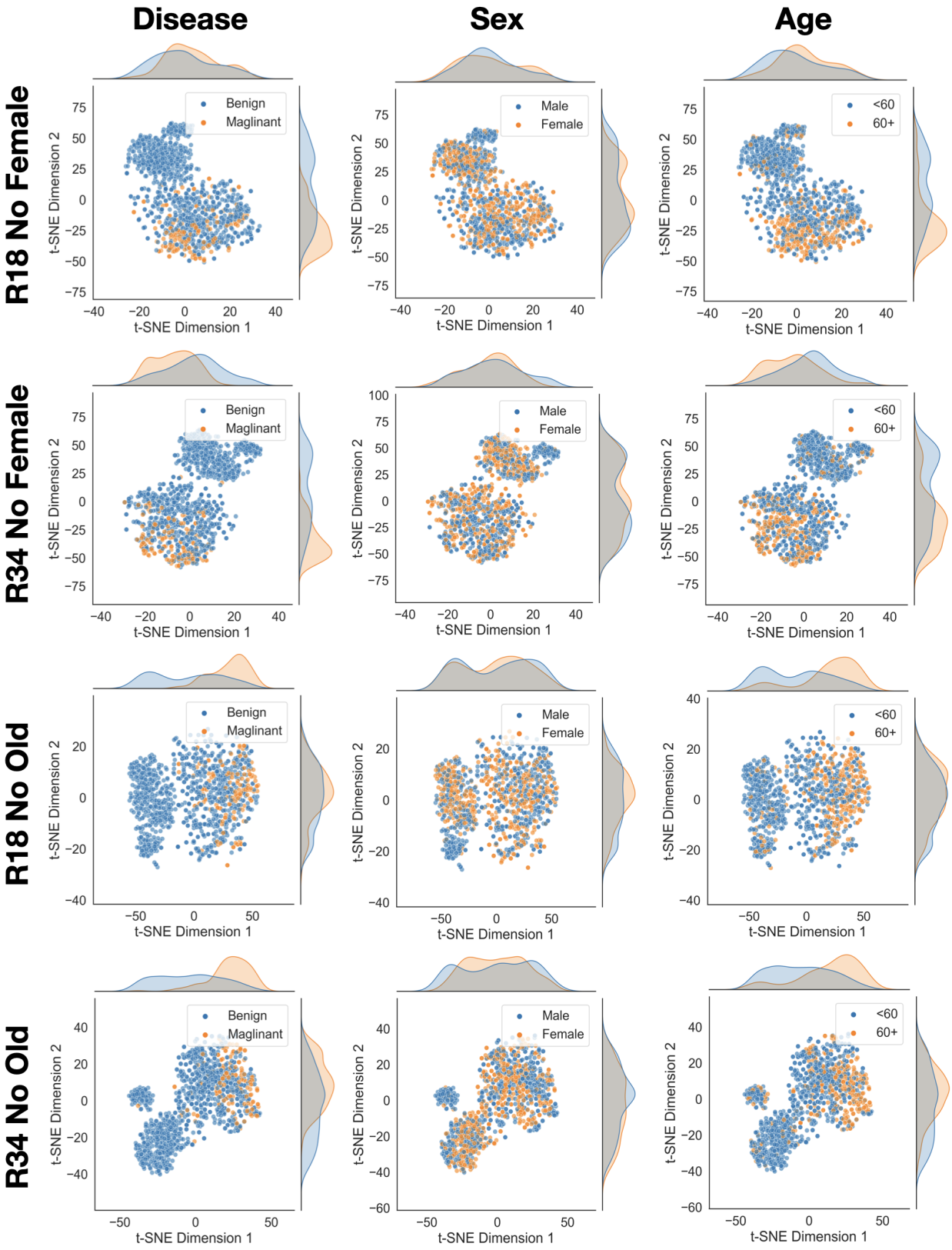

Figure A.9: t-SNE students scatter plots with marginal distributions for **Ham10000** test data feature representations trained **without KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respecitvely.

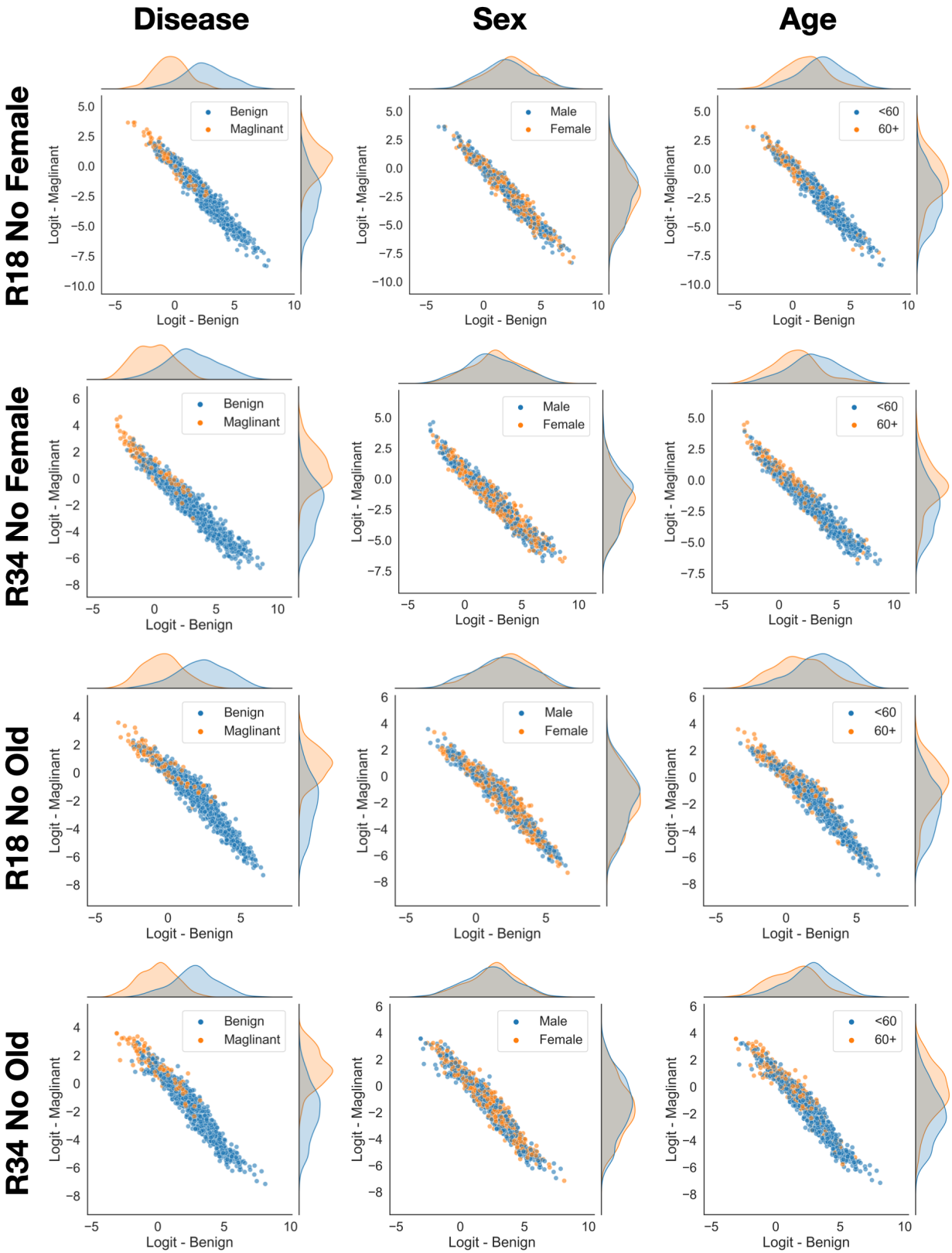

Figure A.10: Logits students scatter plots with marginal distributions for **Ham10000** test data feature representations trained with **Response KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

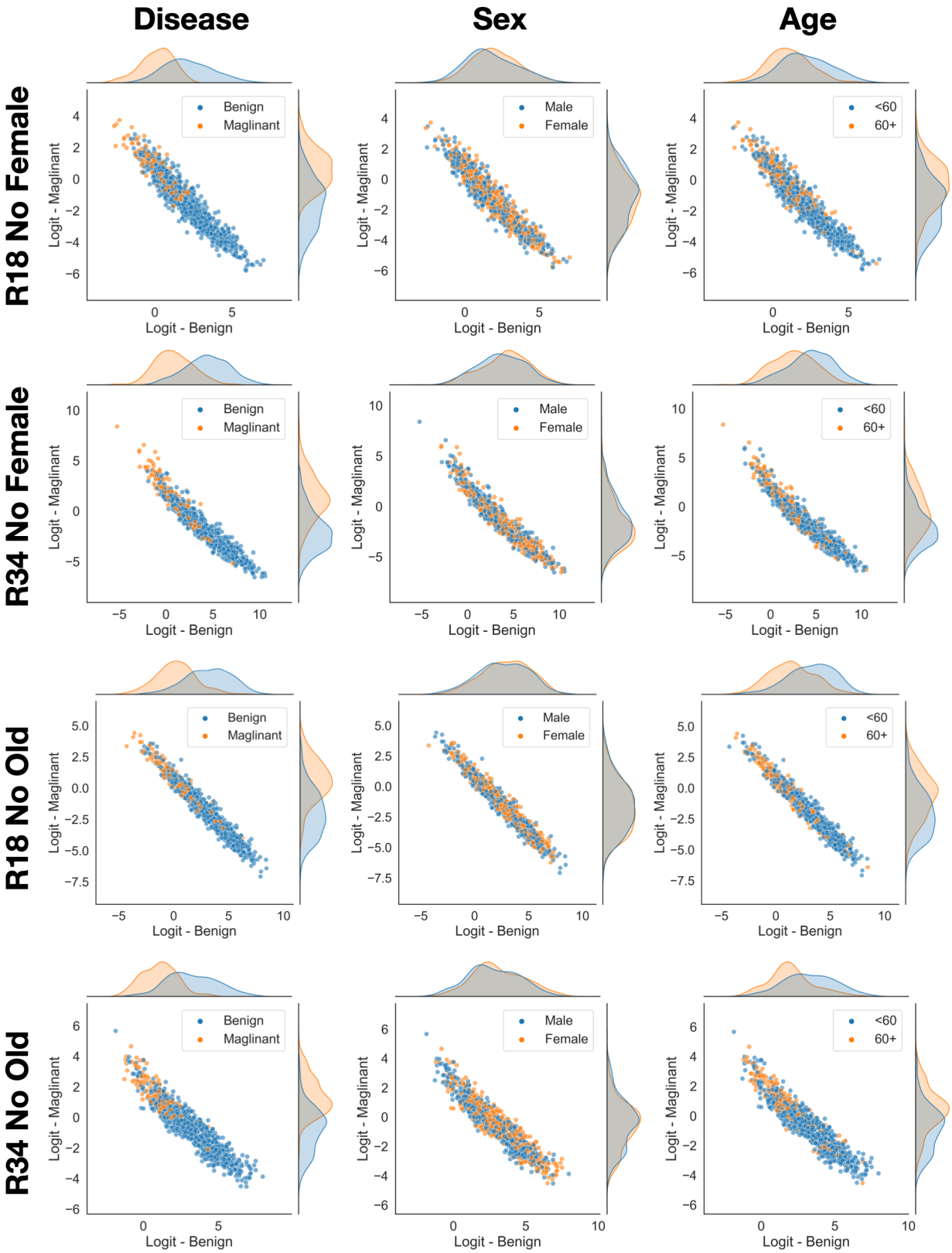

Figure A.11: Logits students scatter plots with marginal distributions for **Ham10000** test data feature representations trained **without KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

A.5.2 CheXpert

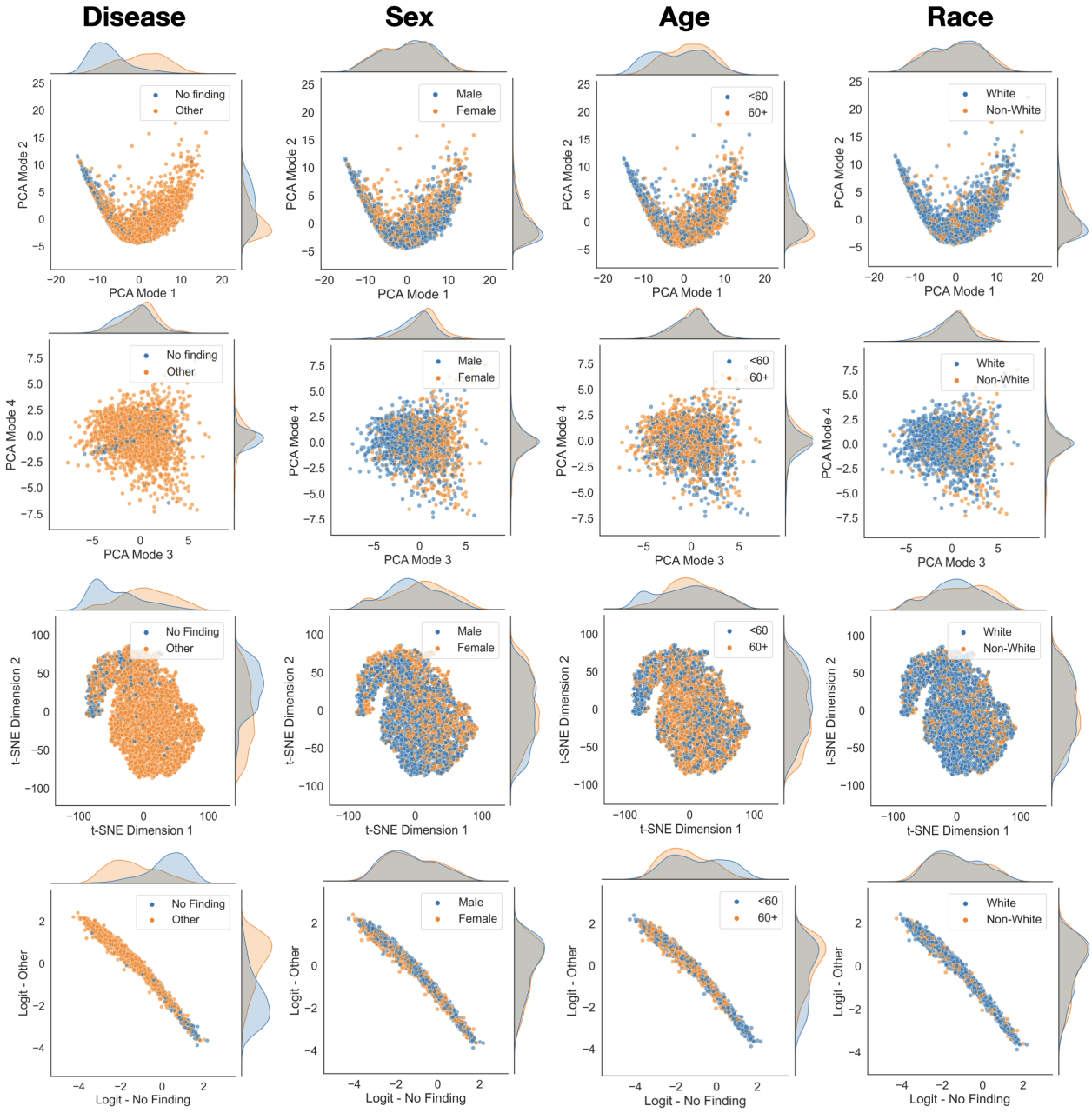

Figure A.12: PCA, t-SNE and Logit scatter plots with marginal distributions for **CheXpert** test data feature representations for fair ResNet34 **Teacher**.

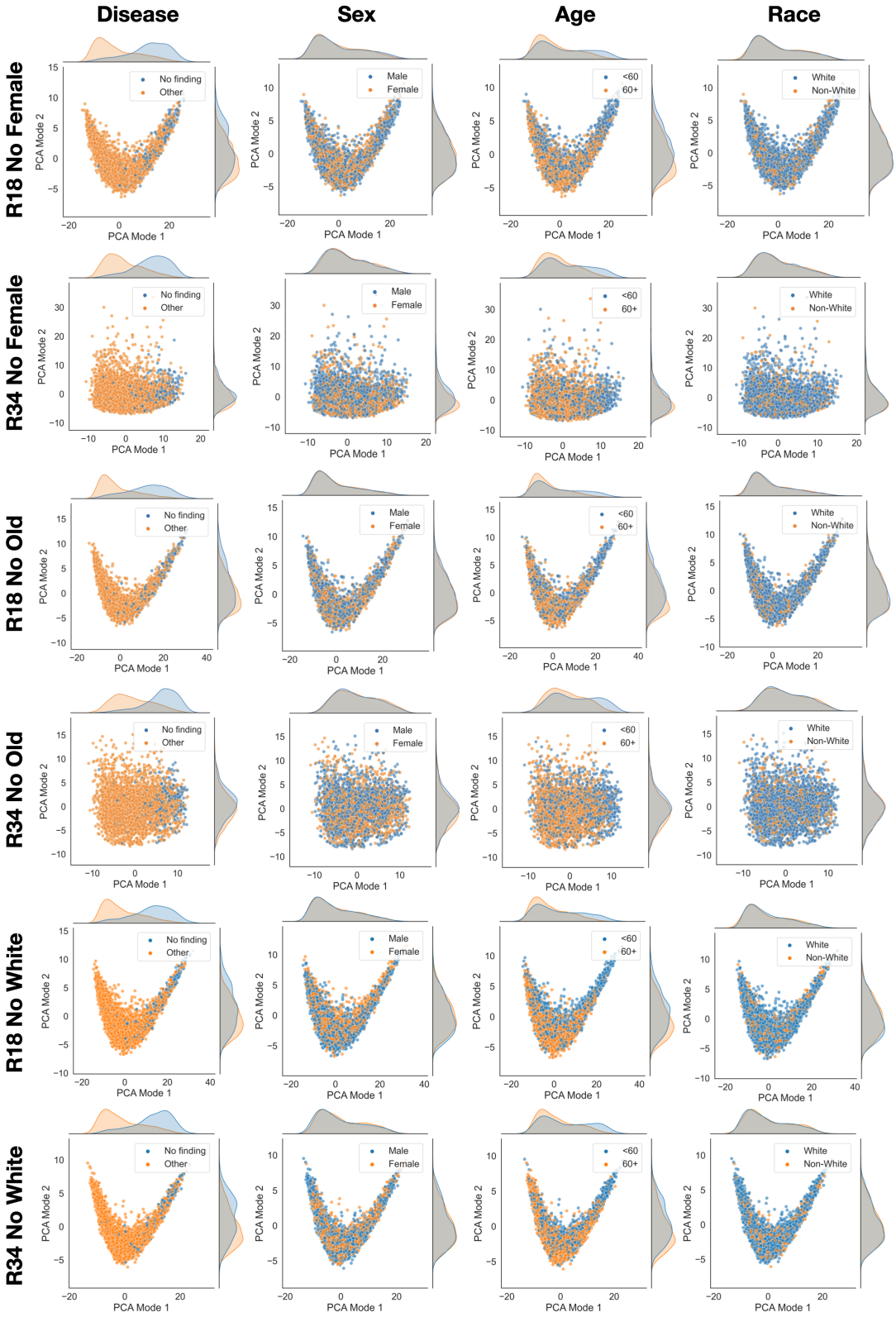

Figure A.13: PCA mode 1 & 2 students scatter plots with marginal distributions for **CheXpert** test data feature representations trained with **Response KD** on No Female/Old/White unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

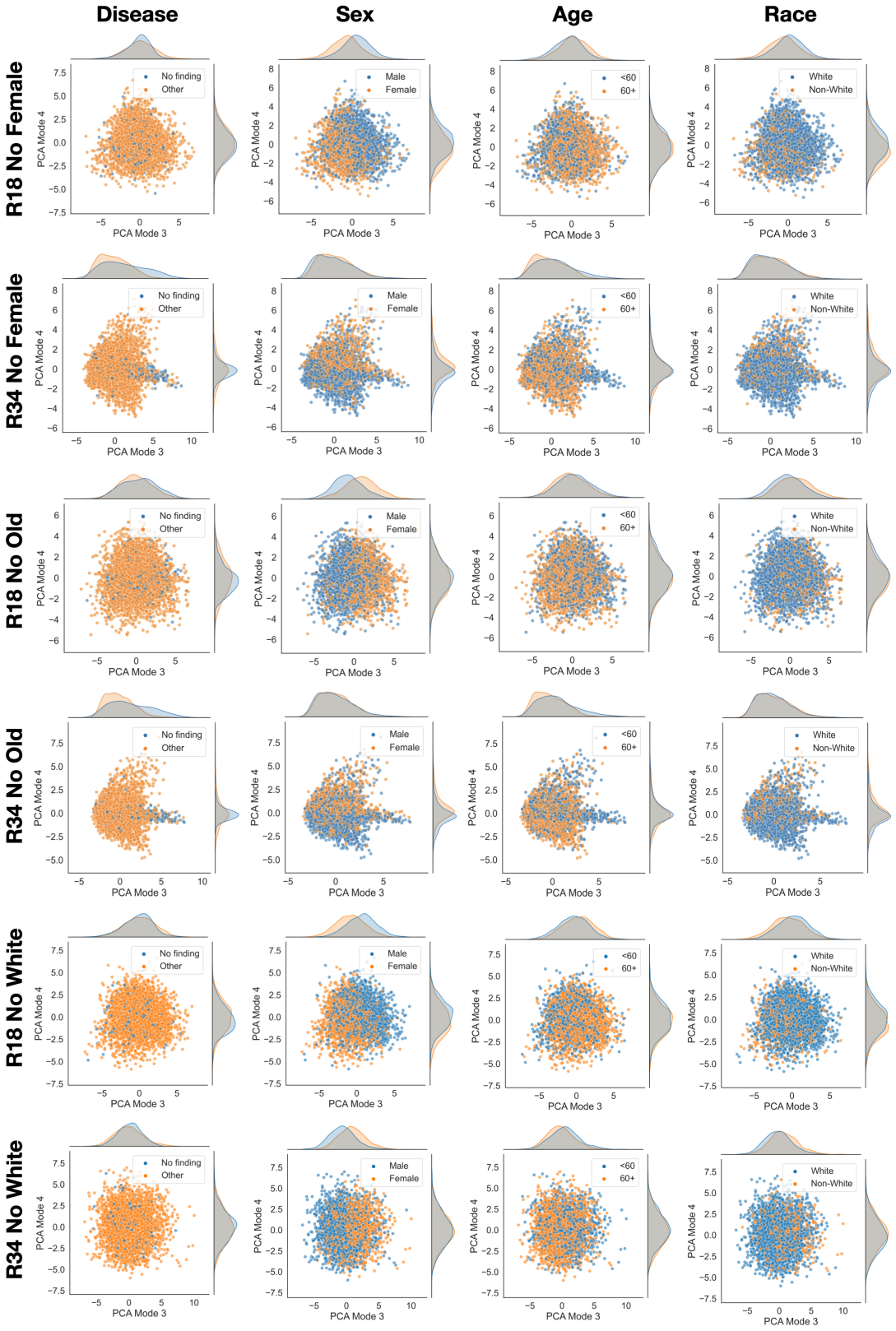

Figure A.14: PCA mode 3 & 4 students scatter plots with marginal distributions for **CheXpert** test data feature representations trained with **Response KD** on No Female/Old/White unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

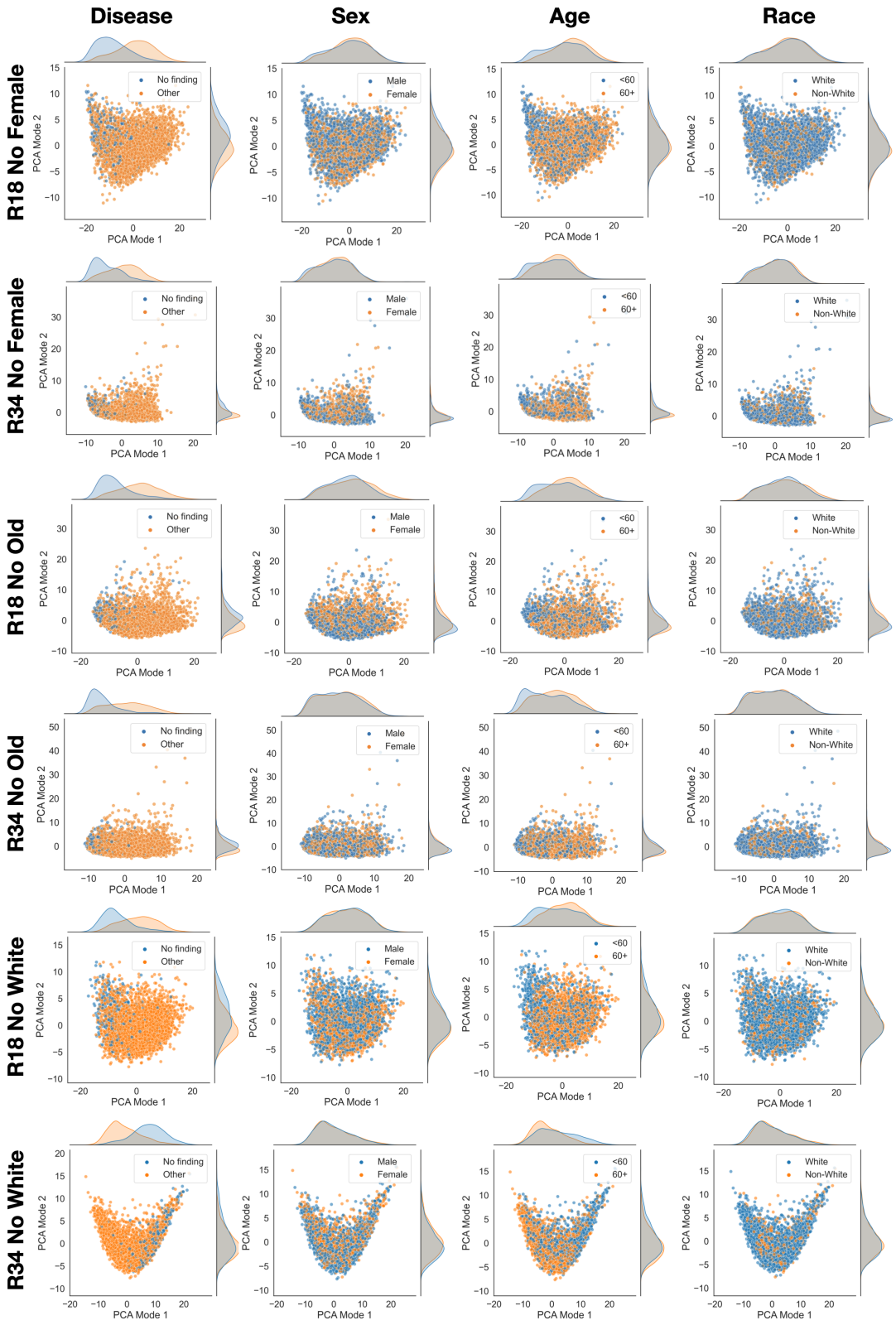

Figure A.15: PCA mode 1 & 2 students scatter plots with marginal distributions for **CheXpert** test data feature representations trained **without KD** on No Female/Old/White unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

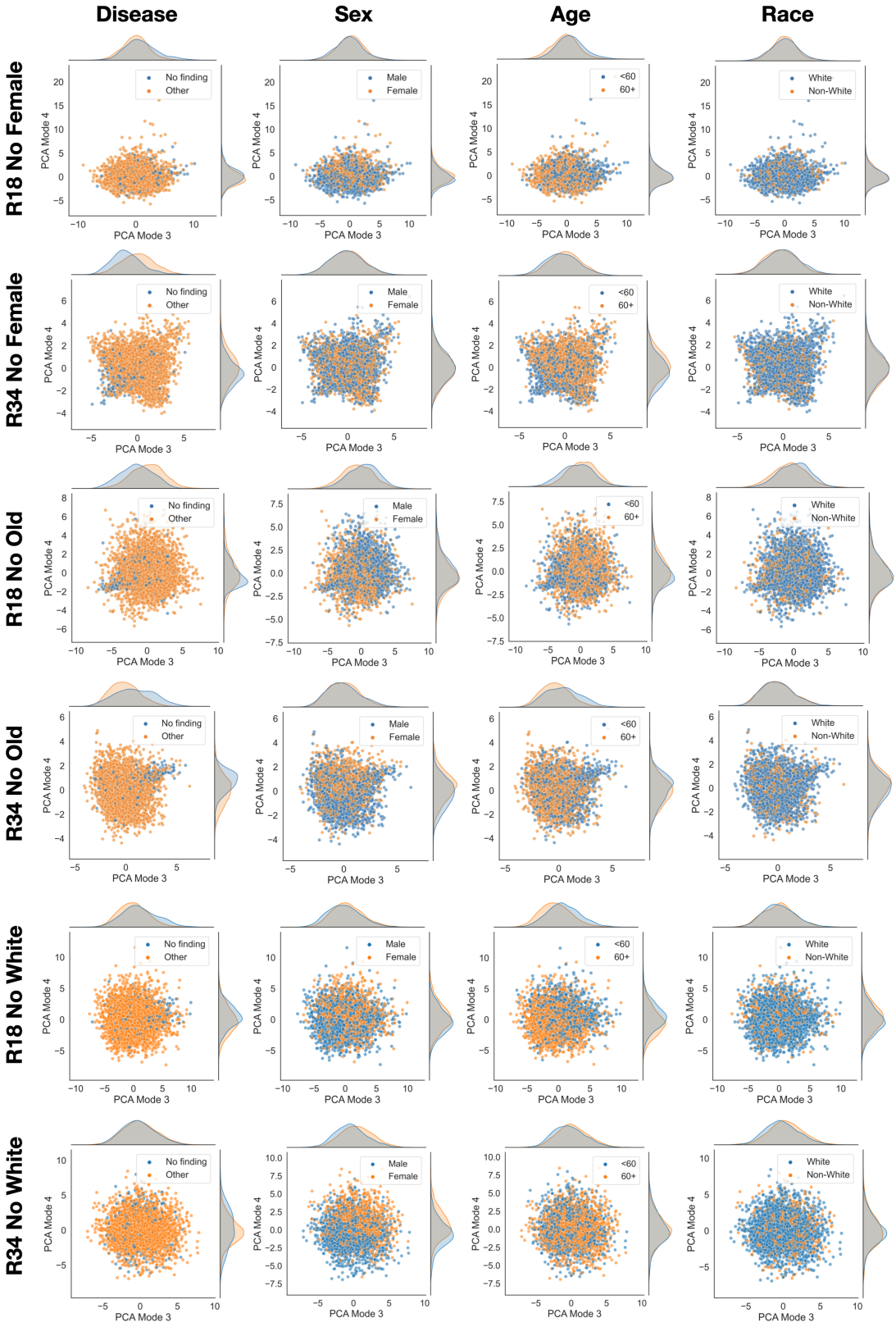

Figure A.16: PCA mode 3 & 4 students scatter plots with marginal distributions for **CheXpert** test data feature representations trained **without KD** on No Female/Old/White unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

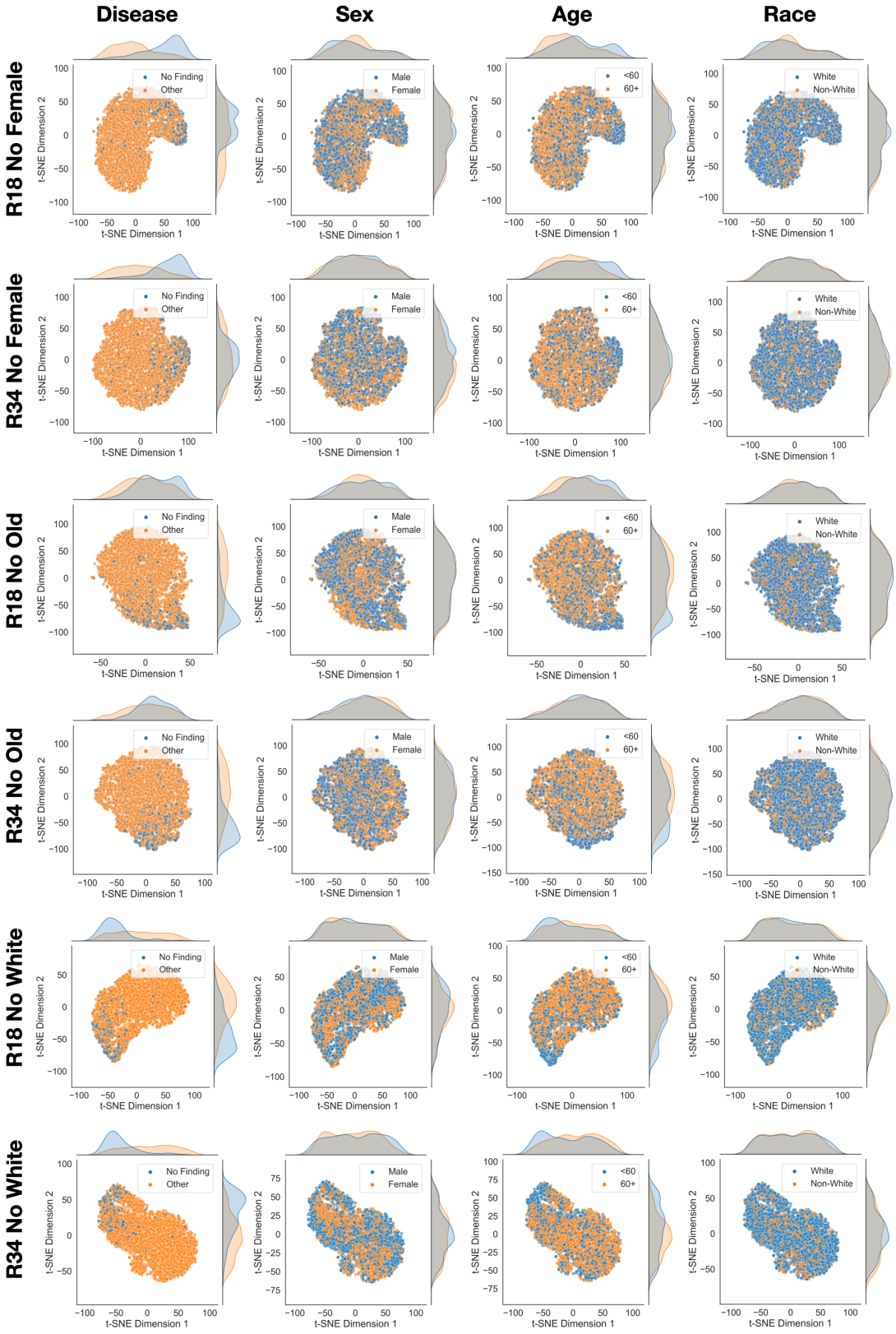

Figure A.17: t-SNE students scatter plots with marginal distributions for **Ham10000** test data feature representations trained with **Response KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

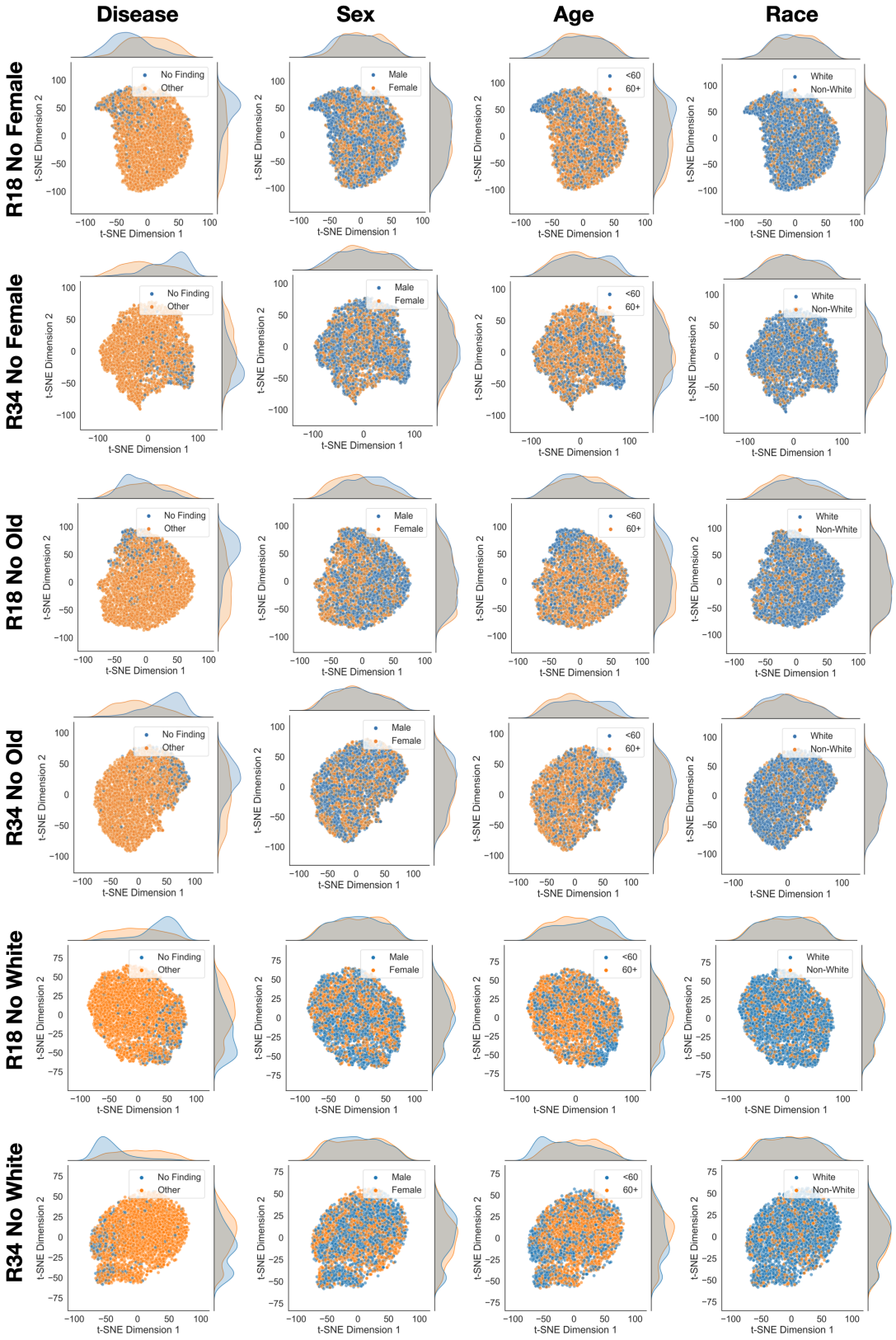

Figure A.18: t-SNE students scatter plots with marginal distributions for **Ham10000** test data feature representations trained **without KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

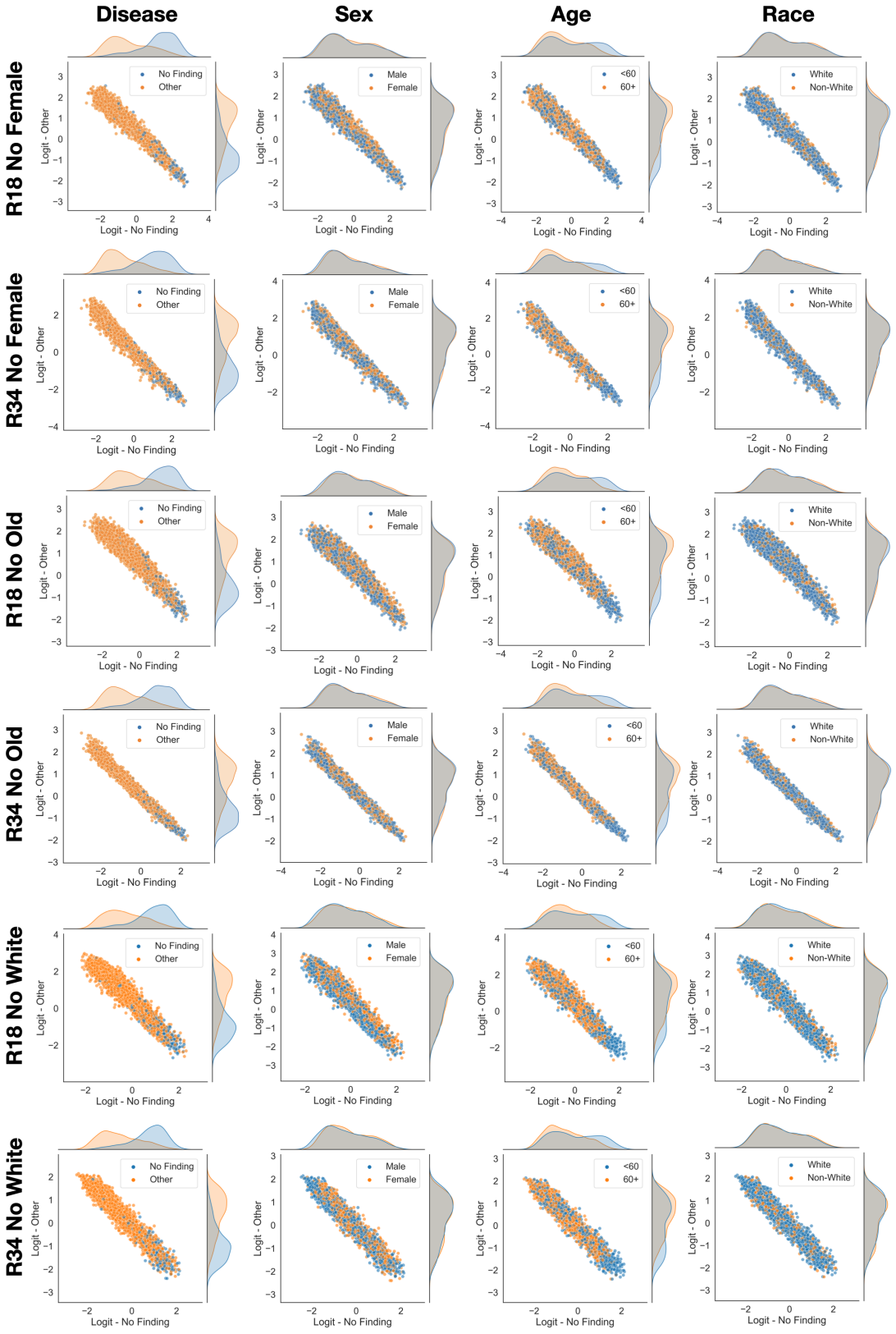

Figure A.19: Logits students scatter plots with marginal distributions for **Ham10000** test data feature representations trained with **Response KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

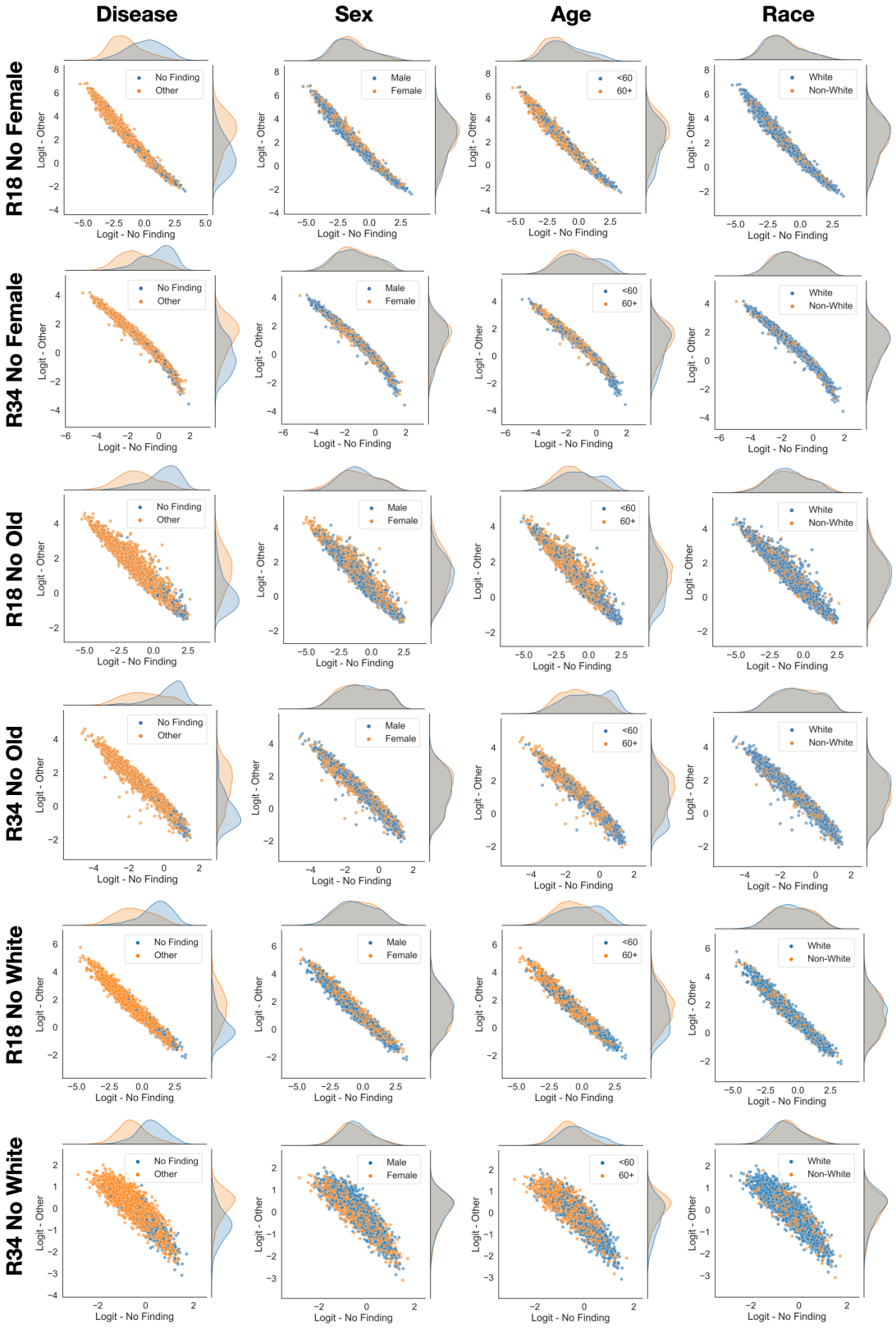

Figure A.20: Logits students scatter plots with marginal distributions for **Ham10000** test data feature representations trained **without KD** on No Female and No Old unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

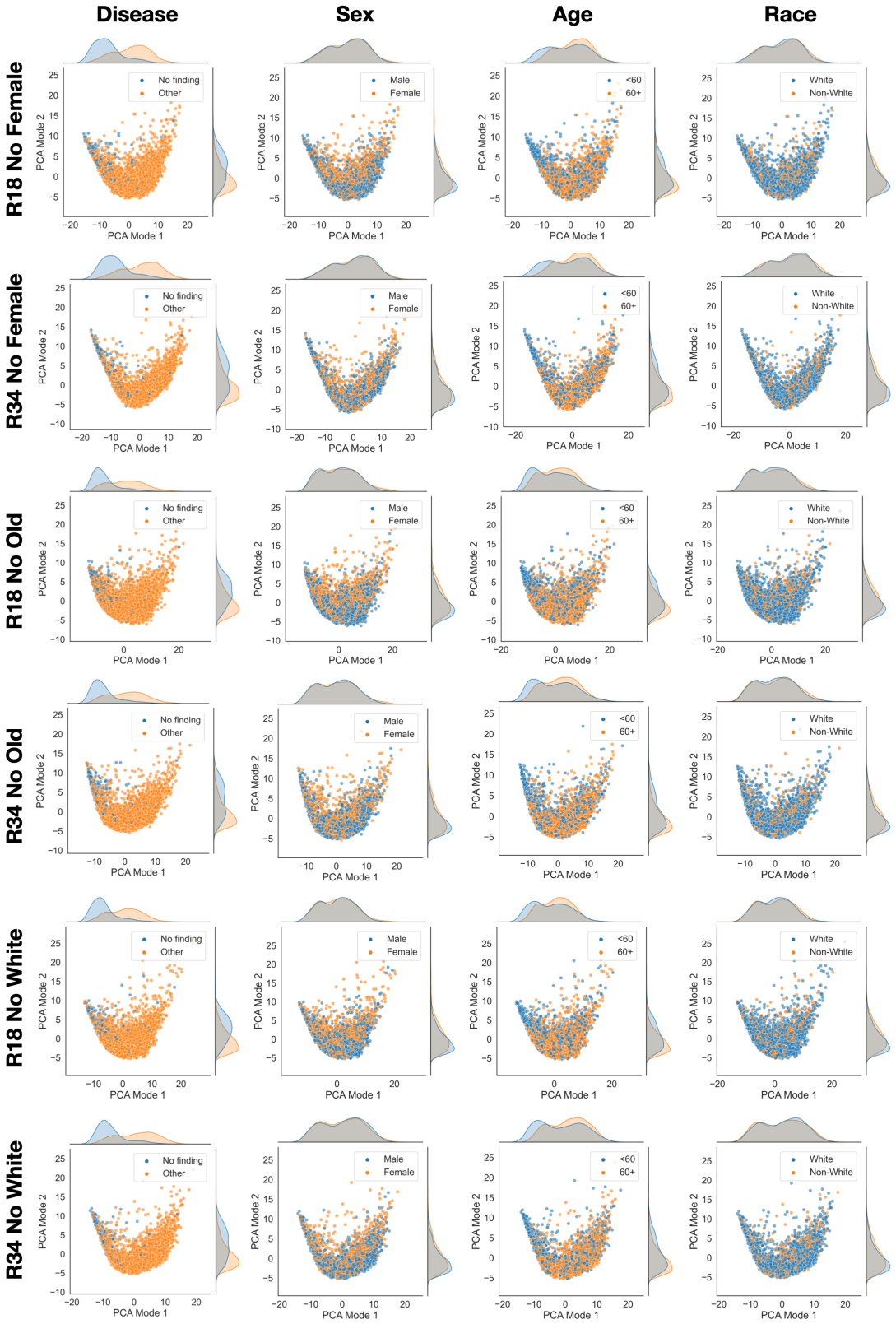

Figure A.21: PCA mode 1 & 2 students scatter plots with marginal distributions for **CheXpert** test data feature representations trained with **Feature KD** on No Female/Old/White unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

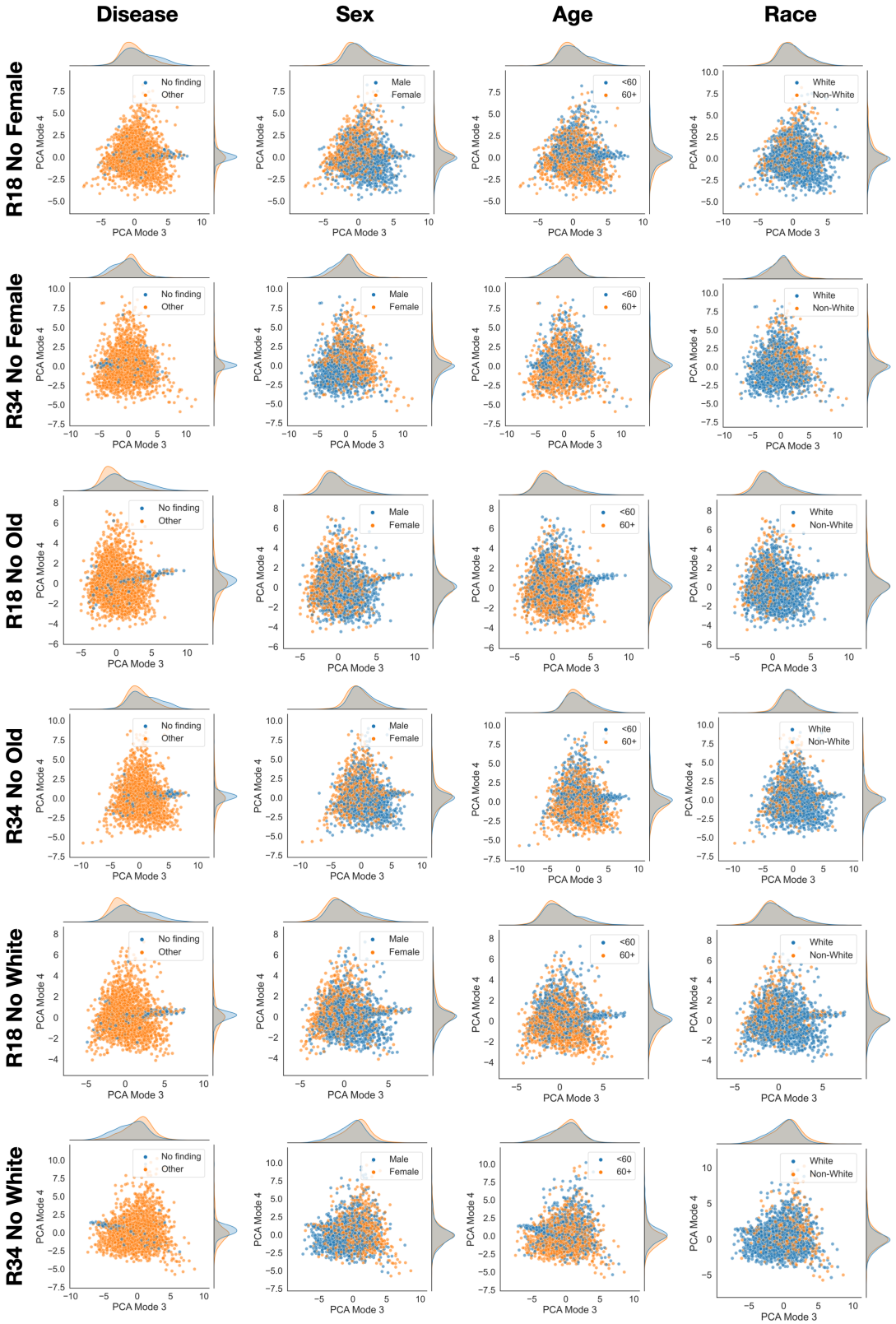

Figure A.22: PCA mode 3 & 4 students scatter plots with marginal distributions for **CheXpert** test data feature representations trained with **Feature KD** on No Female/Old/White unfair compositions. R18 and R34 mean ResNet18 and ResNet34 respectively.

A.6 KS-Tests

We provide two-sample Kolmogorov-Smirnov tests that are performed between the pairs of subgroups indicated in each column. The p-values are adjusted for multiple testing using the Benjamini-Yekutieli procedure and significance is determined at a 95% confidence level. Statistically significant results are marked with * p¡0.05 and ** p¡0.001. Given small input to our overall analysis from KS-tests and overall unclear patterns concerend with Ham10000 we only showcase CheXpert results.

CheXpert ResNet34 Teacher				
	Label	Sex	Age	Race
PCA Mode	p-values			
1	<0.0001**	1.0	<0.0001**	1.0
2	<0.0001**	<0.0001**	<0.0001**	<0.0001**
3	<0.0001**	<0.0001**	1.0	<0.0001**
4	<0.0001**	0.7032	<0.0001**	0.0041*

Table A.5: Kolmogorov-Smirnov tests for comparing marginal distributions across PCA modes for the CheXpert **teacher** model.

CheXpert ResNet18 No Female Student w/o KD				
	Label	Sex	Age	Race
PCA Mode	p-values			
1	<0.0001**	0.0031*	<0.0001**	0.6694
2	<0.0001**	0.0002*	0.02171*	0.2655
3	<0.0001**	0.0126*	<0.0001**	0.3262
4	0.0126*	<0.0001**	1.0	0.5160
CheXpert ResNet18 No Female Student w/ KD				
1	<0.0001**	0.3326	<0.0001**	1.0
2	<0.0001**	0.9208	<0.0001**	1.0
3	0.0009*	<0.0001**	<0.0001**	<0.0001**
4	0.3291	<0.0001**	0.0092*	<0.0001**

Table A.6: Kolmogorov-Smirnov tests for comparing marginal distributions across PCA modes for the **ResNet18 No Female** student model with and without KD.

CheXpert ResNet34 No Female Student w/o KD				
	Label	Sex	Age	Race
PCA Mode	p-values			
1	<0.0001**	0.0808	<0.0001**	0.7863
2	<0.0001**	<0.0001**	<0.0001**	0.0027*
3	<0.0001**	1	<0.0001**	0.0671
4	<0.0001**	0.1007	<0.0001**	0.0455*
CheXpert ResNet34 No Female Student w/ KD				
1	<0.0001**	0.4937	<0.0001**	1
2	0.0212*	<0.0001**	<0.0001**	0.8716
3	<0.0001**	<0.0001**	<0.0001**	0.398
4	<0.0001**	<0.0001**	<0.0001**	<0.0001**

Table A.7: Kolmogorov-Smirnov tests for comparing marginal distributions across PCA modes for the **ResNet34 No Female** student model with and without KD.

CheXpert ResNet18 No Old Student w/o KD				
	Label	Sex	Age	Race
PCA Mode	p-values			
1	<0.0001**	<0.0001**	<0.0001**	0.106
2	<0.0001**	<0.0001**	<0.0001**	<0.0001**
3	<0.0001**	<0.0001**	<0.0001**	<0.0001**
4	0.0002*	<0.0001**	<0.0001**	0.9823
CheXpert ResNet18 No Old Student w/ KD				
1	<0.0001**	1	<0.0001**	1
2	<0.0001**	0.2831	<0.0001**	1
3	0.0018*	<0.0001**	<0.0001**	<0.0001**
4	<0.0001**	0.0828	0.0973	0.0047*

Table A.8: Kolmogorov-Smirnov tests for comparing marginal distributions across PCA modes for the **ResNet18 No Old** student model with and without KD.

CheXpert ResNet34 No Old Student w/o KD				
	Label	Sex	Age	Race
PCA Mode	p-values			
1	<0.0001**	0.1159	<0.0001**	1
2	<0.0001**	0.0007*	<0.0001**	0.0298*
3	<0.0001**	0.9061	<0.0001**	1
4	<0.0001**	<0.0001**	<0.0001**	0.0009*
CheXpert ResNet34 No Old Student w/ KD				
1	<0.0001**	0.3914	<0.0001**	1
2	<0.0001**	0.0006*	<0.0001**	0.8779
3	<0.0001**	0.3914	<0.0001**	1
4	<0.0001**	<0.0001**	<0.0001**	<0.0001**

Table A.9: Kolmogorov-Smirnov tests for comparing marginal distributions across PCA modes for the **ResNet34 No Old** student model with and without KD.

CheXpert ResNet18 No White Student w/o KD				
	Label	Sex	Age	Race
PCA Mode	p-values			
1	<0.0001**	1	<0.0001**	1
2	<0.0001**	0.0013*	<0.0001**	1
3	<0.0001**	<0.0001**	<0.0001**	1
4	<0.0001**	<0.0001**	<0.0001**	<0.0001**
CheXpert ResNet18 No White Student w/ KD				
1	<0.0001**	1	<0.0001**	1
2	<0.0001**	0.0031*	<0.0001**	0.2144
3	0.0336*	<0.0001**	<0.0001**	<0.0001**
4	0.0165*	<0.0001**	<0.0001**	<0.0001**

Table A.10: Kolmogorov-Smirnov tests for comparing marginal distributions across PCA modes for the **ResNet18 No White** student model with and without KD.

CheXpert ResNet18 No White Student w/o KD				
	Label	Sex	Age	Race
PCA Mode	p-values			
1	<0.0001**	1	<0.0001**	0.188
2	0.0206*	<0.0001**	<0.0001**	0.9203
3	0.9979	<0.0001**	<0.0001**	<0.0001**
4	<0.0001**	<0.0001**	0.5746	0.2431
CheXpert ResNet18 No White Student w/ KD				
1	<0.0001**	0.0766	<0.0001**	0.53
2	<0.0001**	0.2314	<0.0001**	1
3	0.1064	<0.0001**	<0.0001**	<0.0001**
4	1	1	0.0110*	0.0094*

Table A.11: Kolmogorov-Smirnov tests for comparing marginal distributions across PCA modes for the **ResNet34 No White** student model with and without KD.

A.7 Capacity Gap

Data	Model	AUC	Avg AUC^{Gap}	Subgroup-AUC Distance to Teacher
No Female	ResNet18	-0.43%	17.77%	19.18%
	ResNet34	-0.39%	16.83%	35.98%
No Old	ResNet18	-0.38%	26.22%	40.05%
	ResNet34	-0.36%	21.36%	86.18%
No White	ResNet18	-0.47%	10.04%	-22.20%
	ResNet34	-0.20%	4.62%	-11.93%

Table A.12: CheXpert: Relative difference in student performance between capacity gap setting (ResNet101 Teacher) and no capacity gap setting (ResNet34 Teacher). 'Subgroup-AUC distance to Teacher' is calculated as Euclidean distance between the student's and the teacher's subgroup performance metrics (e.g. Male AUC). Overall we see decrease in performance, increase in subgroup disparities and subgroup performance. Surprisingly, we also see better subgroup performance resemblance to the teacher in No White high capacity gap setting.

Data	Model	AUC	Avg AUC^{Gap}	Subgroup-AUC Distance to Teacher
No Female	ResNet18	-2.22%	40.46%	244.41%
	ResNet34	-1.42%	19.63%	253.87%
No Old	ResNet18	-1.16%	17.44%	111.65%
	ResNet34	-2.24%	37.62%	327.22%

Table A.13: Ham10000: Relative difference in student performance between capacity gap setting (ResNet101 Teacher) and no capacity gap setting (ResNet34 Teacher). 'Subgroup-AUC distance to Teacher' is calculated as Euclidean distance between the student's and the teacher's subgroup performance metrics (e.g. Male AUC). Overall we see decrease in performance, increase in subgroup disparities and subgroup performance.

A.8 Correlation Heatmaps

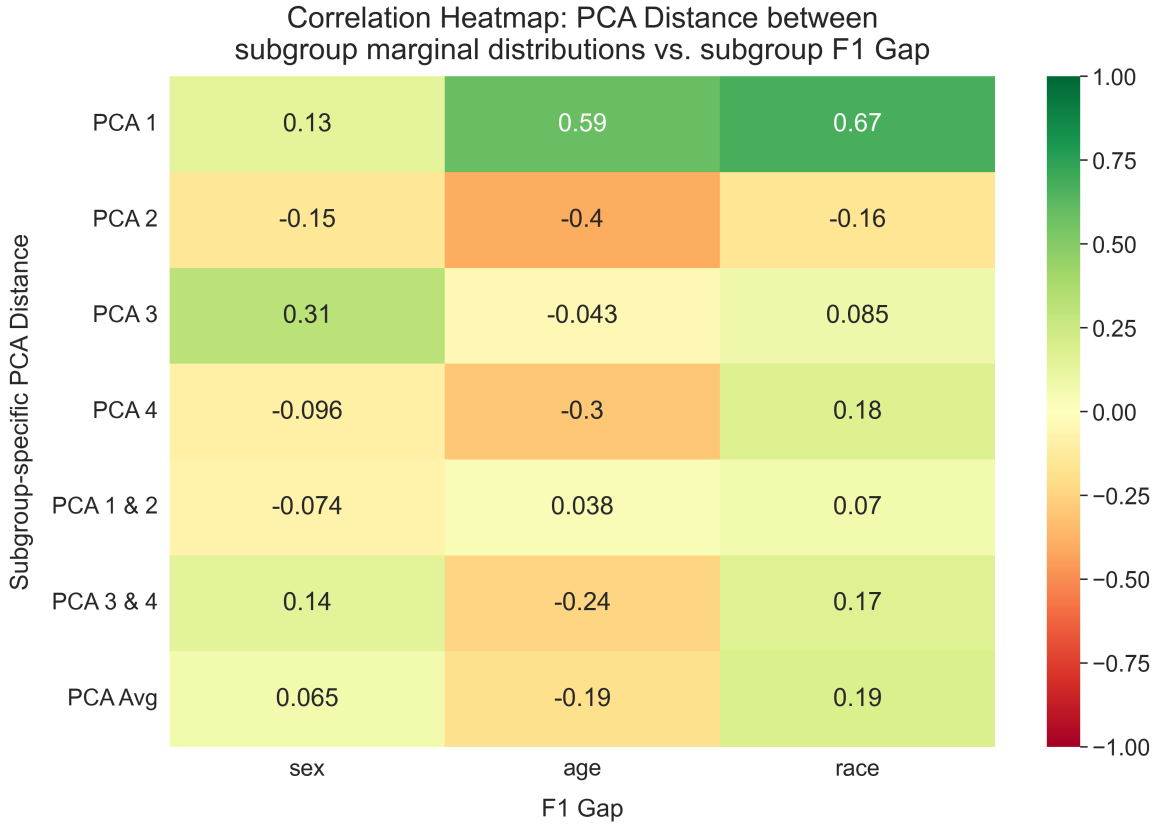

Figure A.23: Heatmap visualising the correlation between the PCA distributional differences in subgroups, captured by PCA distance, and their associated performance disparities, represented by the **F1** gap.

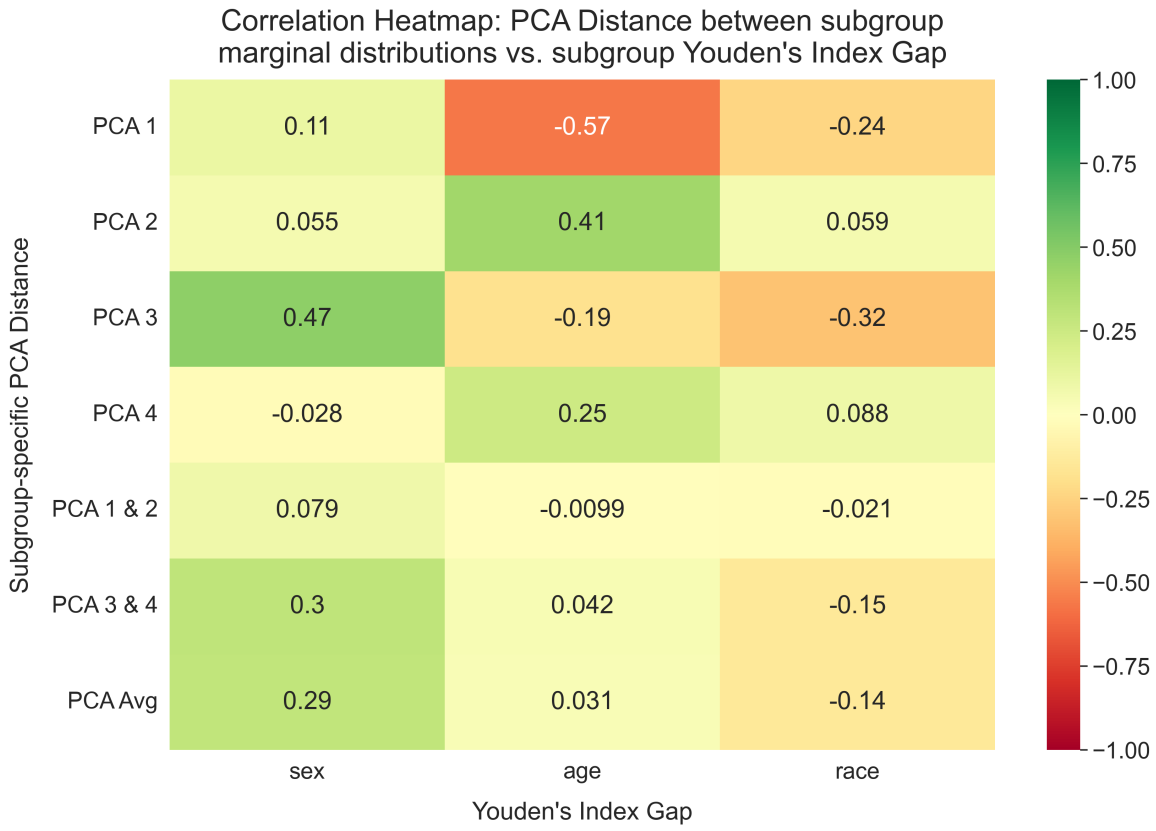

Figure A.24: Heatmap visualising the correlation between the PCA distributional differences in subgroups, captured by PCA distance, and their associated performance disparities, represented by the **Youden Index** gap.