arXiv:2611.31014v1 [eess.IV] 23 Nov 2026

人口统计信息编码并不等于暴露：胸部 X 光基础模型中的公平性与隐私

Sahar Askari **Amin Golnari**

摘要

医学基础模型会编码患者人口统计信息，也可能泄露患者是否属于训练集，而这些属性通常被视为同一个问题。我们通过在相同模型上测量人口统计编码、诊断公平性差距和成员推断脆弱性，检验它们是否相互关联。我们在冻结的 BiomedCLIP、DINOv2 和 RAD-DINO 嵌入上训练了 200 个线性或多层感知机（MLP）头，采用三个随机种子，并在 CheXpert 和 NIH ChestX-ray14 上对其进行审计。人口统计编码几乎无法解释患者脆弱性的方差，且 CheXpert 上的全部 72 项关联在统计上均等同于零。抹除线性可解码的人口统计信息后，泄漏程度未变，但年龄相关漏诊差距有所缩小。泄漏取决于头部容量和训练时长。在 RAD-DINO 上，线性头的成员推断曲线下面积（AUC）为 0.694，而训练 30 轮的 MLP 在测试性能相近的情况下，AUC 达到 0.907。因此，应分别审计公平性与隐私；使用线性头或缩短训练时长可以减少泄漏，且几乎不会或完全不会损失效用。代码已在 GitHub 上开放，供公众使用。

关键词 算法公平性 · 胸部放射摄影 · 概念擦除 · 成员推断 · 隐私

1 引言

基础模型正逐渐成为医学影像领域的默认起点，但它们在临床应用中带来了两方面的顾虑。第一是公平性。胸部 X 光分类器会漏诊医疗服务不足的患者群体 [1, 2]。图像表征还会编码性别、年龄和自我报告种族等受保护属性，尤其是种族，临床医生无法从图像中读取这一信息 [3, 4]。第二是隐私。成员推断攻击（MIA）可能揭示患者记录是否被用于训练 [5, 6]。近期一项大规模审计显示，这种风险在患者群体中的分布并不均衡，且会随模型容量的增大而上升 [7]。

这两方面的顾虑通常会被放在一起讨论，其中隐含着一种因果叙事：如果一种表征编码了患者的身份信息，那么该患者可能更容易被单独识别，而编码的属性也可能为偏倚预测提供捷径 [8]。这一叙事预测，人口统计信息编码、诊断公平性差距和成员脆弱性彼此一致。这个问题具有实际影响，因为它决定了某种缓解措施（例如移除人口统计信息）能否同时实现两个目标。如果这三个量彼此一致，一种缓解措施就可能同时实现两个目标；如果它们并不一致，则每种风险都需要单独审计。相关研究考察了编码与其在线性预测中的应用之间的区别 [9], X 光嵌入中的持久性人口统计信息 [10], 缓解措施下的公平性与隐私权衡 [11–14], 以及医学生成模型和基础模型中的记忆化 [15, 16]。然而，这些研究均未将人口统计编码的强度与患者层面的成员风险联系起来。据我们所知，也没有研究在同一批模型上测量全部三个量，并直接检验它们是否一致。

我们测试了人口统计编码、诊断公平性差距和成员推断攻击脆弱性在胸部 X 光基础模型中是否相互一致或彼此独立。我们发现三者彼此独立。编码强度无法预测哪些患者会发生信息泄漏，消除

[a]通讯作者。电子邮件：askarisahar92@gmail.com。伊朗设拉子，设拉子大学计算机科学与工程系。[b]电子邮件：amin.g@vinuni.edu.vn。越南河内，Vin大学工程与计算机科学学院。

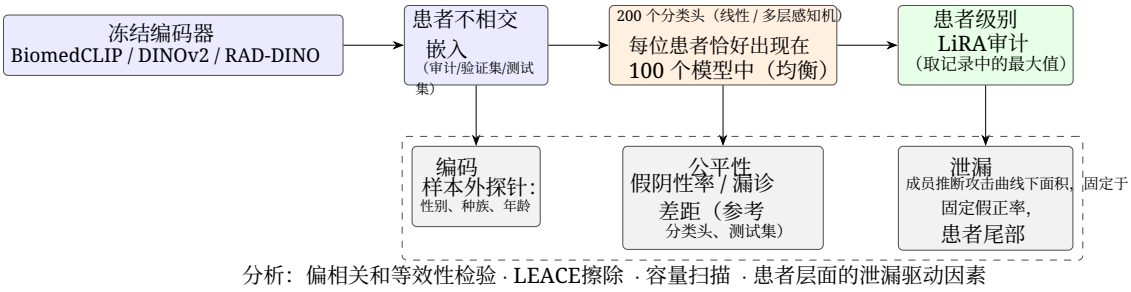

图1：研究设计。冻结嵌入输入200个分类头，每个分类头都在审计患者的随机一半上进行训练。相同的嵌入和分类头产生三项测量结果（编码、公平性和泄漏），随后对这些测量结果进行相互关联分析。折外：折外；LiRA：似然比攻击；FNR：假阴性率；TPR：真正率；FPR：假正率。

线性可解码人口统计信息会改变年龄相关公平性差距，但不会改变泄漏；而泄漏受分类头容量、训练时长和患者非典型性支配。

我们作出四项贡献。首先，我们设计了一项联合审计，在两个公开的胸部X射线数据集和三种编码器上，对同一批200模型集成测量人口统计编码、诊断公平性差距和患者层面成员推断攻击脆弱性；主要配置均采用三个随机种子（第3节和图1）。其次，我们使用等效性检验，而非不显著的 p值，来表明编码强度在统计上与泄漏无关。CheXpert上的全部72个关联均等效于零；在NIH ChestX-ray14上，唯一未通过检验的是年龄相关的六次完全训练的多层感知机运行，且结果的符号均为负（第4.3节）。第三，我们通过线性概念擦除提供干预证据。擦除在不改变泄漏的情况下，去除了大部分线性可解码的人口统计学信息，但会缩小年龄相关漏诊差距，由此将隐私与公平性区分开来（第4.5节）。第四，我们确定了泄漏的驱动因素：分类头容量、训练时长、嵌入家族和患者非典型性可以解释泄漏；而表面上的图像数量效应，很大程度上是对患者记录取最大值所致（第4.2节和第4.4节）。

2 相关工作

2.1 胸部放射摄影中的人口统计编码与公平性

胸部 X 光分类器对女性、年轻人、黑人和低收入患者的漏诊率更高 [1]。深度模型可以从图像中预测自我报告种族，即使相关信号难以定位 [3]。疾病检测模型也会以与性能差异相关的方式编码受保护特征 [4]。公平性差距未必能跨站点迁移 [17], ，而从表征中移除人口统计信息并不能保证行为公平 [18]。与我们的编码分析最为接近的研究包括 Kim 等人 [9], 的工作，他们对比了基础模型编码的内容与其线性头使用的内容；以及 Santos 等人 [10], 的工作，他们表明人口统计信息仍存在于 X 光基础嵌入中。这两项研究都未衡量隐私。

2.2 会 员推断与差异化脆 弱性

成员推断旨在判断某条记录是否属于模型的训练集 [5]。似然比攻击在低假阳性率下对这种推断进行校准 [6]。不同群体的脆弱性存在差异 [12], ，公平性干预也可能改变这种差异 [11]。Knolle 等人 [7] 利用闭式记录级似然比检验，对医学影像模型开展了大规模审计。他们发现，风险集中在一部分患者的尾部群体中，并且会随着架构容量的增加而上升。他们还观察到，某些群体在这一尾部群体中的过度代表，部分可以由群体规模解释。在医学生成模型和微调基础模型中，已有研究表明，记忆化取决于样本的稀有性和独特性 [15, 16]。公平性与隐私权衡在缓解措施下已在亚群体层面接受审计 [13], ，而隐私训练可能保留或削弱公平性 [14]。这些研究将隐私与群体成员身份、缓解措施或样本稀有性联系起来。据我们所知，它们都没有探究人口统计编码是否能够预测患者级泄漏，而这正是我们要研究的问题。

2.3 概念擦除

迭代零空间投影会从表征中移除可线性解码的属性信息 [19]。最小二乘概念擦除（LEACE）以闭式形式执行这一移除，并保证任何线性分类器都无法恢复该

表 1：数据集和编码器。图像数量为抽样选取 15,000 名患者后所使用的数量。

数据集	患者	图像	审计记录	标签	属性	编码器
CheXpert-small	15,000	44,622	31,601	14	性别、种族、年龄	BiomedCLIP、DINOv2、RAD-DINO
NIH ChestX-ray14	15,000	54,521	38,316	15	性别、年龄	BiomedCLIP、RAD-DINO

属性信息 [20]。我们将 LEACE 作为一种干预手段，用于测试移除编码的人口统计信息是否会改变泄漏。

3 方法

3.1 数据与编码器

我们使用了两个公开的胸部X线摄影数据集（表 1）。CheXpert-small [21] 提供性别、年龄组和自我报告种族（种族标签来自 Stanford AIMI 发布的人口统计文件 [3]），而 NIH ChestX-ray14 [22] 仅提供性别和年龄组。我们分别从每个数据集中抽取了 15,000 名患者一次（随机种子 0）。对于每个运行随机种子，我们在患者层级将这些患者划分为审计集（10,500）、验证集（1,500）和测试集（3,000）。患者的年龄取该患者所有记录的均值，并分为四个分组区间：40 岁以下、40 至 60 岁、60 至 80 岁以及 80 岁及以上，其中下限包含在内。CheXpert 包含 14 个标签；我们屏蔽了不确定标签，并将空白标签视为阴性。NIH ChestX-ray14 包含 15 个标签。

嵌入来自三个冻结编码器。BiomedCLIP [23] 通过其图像塔提供 512 维嵌入。DINOv2 [24] 提供 768 维分类标记（CLS）嵌入，并作为通用领域对照。RAD-DINO [25] 提供 768 维 CLS 嵌入，专用于放射学。在 NIH ChestX-ray14 上，我们仅使用 BiomedCLIP 和 RAD-DINO（表 1）。RAD-DINO 的预训练数据包含 CheXpert。因此，我们将其在 CheXpert 上的结果视为领域匹配编码器表现的上界，而非对泛化能力的纯粹测试。我们使用审计集统计量对所有嵌入进行了标准化。

3.2 分类头与成员资格审计

遵循 Knolle 等人 [7], 的方法，我们训练了 $N = 200$ 个模型，每个模型使用随机抽取的一半审计患者进行训练，因此每位患者恰好是 100 个模型的成员。每个模型要么是在冻结嵌入上训练的线性头，要么是单隐藏层多层感知机（MLP）。损失函数、训练设置和指标详见第 3.3 节。

令$\mathbf{x}_i$表示记录 i的嵌入， $f_{mj}(\mathbf{x}_i)$表示模型 m针对标签 j的 logit， $y_{ij} \in \{0, 1\}$ 表示标签， $v_{ij} \in \{0, 1\}$表示标签有效的指示变量。记录级分数是对有效标签的带符号 logit 间隔取均值，

$$s_{mi} = \frac{\sum_j v_{ij} \, (2y_{ij} - 1) \, f_{mj}(\mathbf{x}_i)}{\sum_j v_{ij}}. \tag{1}$$

对于每条记录 i，我们计算闭式似然比攻击（LiRA）统计量

$$\theta_i = \Phi\left(\frac{\mu_{\text{in}} - \mu_{\text{out}}}{\sqrt{\sigma^2_{\text{in}} + \sigma^2_{\text{out}}}} \right), \tag{2}$$

其中， μ_{in} 和 σ^2_{in} 分别是包含该记录的模型中记录分数 s_{mi} 的均值和方差， μ_{out} 和σ^2_{out} 则是排除该记录的模型中的相同统计量（为便于阅读，省略记录索引），而 Φ 是标准正态累积分布函数。统计量 θ_i 等于在两组分数分布均为高斯分布时，用于区分成员与非成员分数分布的曲线下面积（AUC）。

作为校准检查，我们重新随机抽取成员身份。该零假设在每次运行中得到的患者级第 99 百分位数约为 0.606，且闭式 AUC 与经验 AUC 一致。我们还报告了另外两种患者评分，它们消除了对记录取最大值所产生的影响，分别是记录均值和一个随机记录的分数。

3.3 损失、指标和超参数

每个模型都针对其所属记录中的有效标签最小化二元交叉熵，并对标签和记录取平均。我们从全部 R 条审计记录中有放回地抽取 $B = 2048$ 条记录组成小批次，并训练所有

表2：第3.4节的分类头和编码探针的超参数。一个训练轮次包含 $\lceil R/B \rceil$ 小批量步数。表6中的探针将在第3.7节中介绍。

	设置	值
头	架构	线性，或使用 GELU 的多层感知机（512 个隐藏单元；其中 8 个和 64 个在参数扫描）
	模型数量 N	200（每位患者对应 100 个模型）
	优化器、权重衰减	AdamW [26], 10^{-2}
	学习率	3×10^{-3} (线性)， 10^{-3} (多层感知机)
	调度	余弦退火至零
	批量大小 B，训练轮数	2048，30（参数扫描中为 5）
	随机种子	3（主要配置），1（参数扫描）
探针	优化器、权重衰减	AdamW， 10^{-4}
	学习率、批量大小	3×10^{-3}，2048
	训练轮数、交叉验证	30，5 折

N 个模型使用同一小批次进行训练，同时将每个模型的损失限制在其所属记录上（约占每个小批次的一半）。一个训练轮次对应 $\lceil R/B \rceil$ 个步骤。表2列出了分类头以及用于编码的线性探针（第3.4节）的超参数。权重从经过扇入倒数平方根缩放的高斯分布中初始化，偏置设为零。第3.5节中的参考头采用相同的设置，但使用全部审计患者进行训练。

我们将效用衡量为在两个类别均存在的标签上计算的宏平均受试者工作特征曲线下面积（受试者工作特征曲线下面积），该指标由参考头在测试患者上计算得出。

我们在记录、聚合和患者层级衡量泄漏。令 $z_{mi} \in \{0,1\}$ 表示记录 i 是否属于模型 m的训练集。对于聚合攻击，我们针对模型 m 和记录 i 的每一组配对计算对数似然比

$$\Lambda_{mi} = \log \frac{\mathcal{N}\big(s_{mi}; \mu_{\text{in}}^{(-m)}, \sigma_{\text{in}}^{2\,(-m)}\big)}{\mathcal{N}\big(s_{mi}; \mu_{\text{out}}^{(-m)}, \sigma_{\text{out}}^{2\,(-m)}\big)}, \tag{3}$$

其中， $\mathcal{N}(\cdot;\, \mu,\ \sigma^2)$ 是高斯密度，均值和方差根据除 m之外的模型估计。聚合成员推断攻击AUC是区分成员（$z_{mi}=1$）和非成员（$z_{mi}=0$）的曲线下面积，依据为 Λ_{mi}，在最多四百万组随机抽样的配对上计算。我们还报告假正率（FPR）为 $\alpha = 10^{-3}$时的真正率（TPR），

$$\text{TPR@}\alpha = \Pr\big(\Lambda_{mi} > \lambda_\alpha \mid z_{mi}=1\big), \quad \text{with} \quad \Pr\big(\Lambda_{mi} > \lambda_\alpha \mid z_{mi}=0\big) = \alpha, \tag{4}$$

其中 λ_α 为决策阈值。记录级统计量 θ_i 通过以下方式聚合至患者层级：

$$\theta_p = \max_{i \in \mathcal{R}_p} \theta_i, \tag{5}$$

其中 $\mathcal{R}_p$ 为患者 p的记录集合。我们通过其第99百分位数以及满足$\theta_p \geq 0.9$的患者占比来概括 θ_p。

对于具有 C 个类别的属性，患者级编码得分是患者 p真实类别 c_p 的先验校正后验概率，

$$e_p = \frac{\exp\big(o_{pc_p} - \log \pi_{c_p}\big)}{\sum_{c=1}^{C} \exp\big(o_{pc} - \log \pi_c\big)}, \tag{6}$$

其中 o_{pc} 是线性探针中类别 c 的折外逻辑值， π_c 是训练折中的类别先验。 e_p 的随机水平为 $1/C$。总体编码强度是后验概率的一对其余宏平均受试者工作特征曲线下面积。

对于测试患者组 g，并使用第3.5节中的逐标签阈值，宏平均假阴性率（FNR）和漏诊率分别为

$$\text{FNR}_g = \frac{1}{|\mathcal{J}_g|} \sum_{j \in \mathcal{J}_g} \frac{\text{FN}_{gj}}{\text{FN}_{gj} + \text{TP}_{gj}}, \qquad \text{UD}_g = \frac{\text{FP}_{g0}}{\text{FP}_{g0} + \text{TN}_{g0}}, \tag{7}$$

其中，FN、TP、FP 和 TN 表示在这些阈值下的记录数， $\mathcal{J}_g$ 是 g中至少包含20条阳性记录的疾病标签集合，索引0表示标签“No Finding”。因此，漏诊率就是以下记录层面的假正率：

“No Finding”，该指标基于具有有效阴性“No Finding”标签的记录计算。相对于参考组r 的差距为

$$\Delta_g^{\mathrm{FNR}} = \mathrm{FNR}_g - \mathrm{FNR}_r, \qquad \Delta_g^{\mathrm{UD}} = \mathrm{UD}_g - \mathrm{UD}_r. \tag{8}$$

3.4 人口统计编码

对于每个属性，我们在患者均值嵌入上拟合线性探针，并通过五折交叉验证获得折外（OOF）预测结果。编码得分和编码强度的定义见公式（6）和第 3.3 节。探针使用了全部 15,000名抽样患者，审计患者的得分则用于关联分析。对于 CheXpert 种族，我们排除了“Unknown”和“Other”类别。

3.5 诊断公平性

我们在所有审计患者上训练了一个参考头，并使用 Youden 指数 [27] 在验证集上固定逐标签阈值。在留出测试患者上，我们以每个属性中规模最大的群体为参照，计算相对于该参考组的假阴性率差距和漏诊差距（公式（8））。漏诊率是指患病患者被判定为健康的比例。在每次运行中，我们通过进行 500 次重抽样的患者聚类自助法，估计差距的置信区间（CIs）。对于不同随机种子下原始嵌入与擦除后嵌入之间的差异（第 4.5 节），我们在三个随机种子上使用 t 区间。结果部分重点讨论漏诊差距，这是 Seyyed-Kalantari 等人 [1] 提出的主要衡量指标。假阴性率差距采用相同流程计算，但不再作进一步分析。

3.6 关联与等价性分析

我们通过两种方式考察患者编码得分与患者 MIA 脆弱性之间的关系。第一种是秩变换偏相关 ρ。第二种是编码在回归中增加的增量 ΔR^2；该回归还包含记录数量的对数、k近邻稀疏度（$k = 10$，余弦距离）、发现数量和组大小的对数。非显著的 p 值并不能证明不存在关联，因此我们采用了等效性检验 [28, 29]。若满足以下两个条件，我们就认为关联等效于零： ρ 的 95%置信区间必须落在 ± 0.1 范围内，这符合保守的双单侧检验标准；且 ΔR^2 置信区间的上界必须低于 0.01。所有变量均经过秩变换。稀疏度是在审计集内根据患者均值嵌入计算的；发现数量是每条记录中阳性发现数量的均值；组大小的对数则是患者所属组中审计患者数量的对数。 ρ 和 ΔR^2 的置信区间是基于患者级自助法进行 500 次重采样得到的百分位区间。

针对第4.4节的分析，我们在 seed-0运行上拟合了一个包含八个预测变量块的更广义秩回归。这些变量块包括记录数量的对数、稀疏性、发现数量、每条记录中不确定标签的平均比例、每条记录中罕见发现的平均数量（患病率低于5%的标签）、包含支持设备的记录比例、患者年龄，以及人口统计编码（属性编码得分的均值）。对于每个变量块，我们报告移除一个变量块后的 ΔR^2，即从完整模型中移除该变量块时 R^2 的下降幅度。由于这些变量块彼此相关，这些数值之和并不等于总 R^2。

3.7 干预措施与参数扫描

我们应用 LEACE [20] 在 CheXpert 上联合擦除独热编码性别、种族和年龄概念，然后重复进行审计。我们在15,000名抽样患者的所有记录上，以记录级拟合 LEACE，其中包括验证集患者和测试患者，并为所有观测到的种族类别设置独热编码列。由于 LEACE 在此作为分析性干预而非可部署的缓解措施，因此这种传导式拟合会将同一擦除映射应用于每条经过审计和评估的记录。为衡量残余人口统计信息（表6），我们在所有抽样患者经重新标准化的患者均值嵌入上，分别拟合了五折探针。这些探针包括一个线性探针和一个具有一个隐藏层的 MLP探针，该隐藏层包含256个 ReLU 单元；探针采用类别加权交叉熵、AdamW（学习率 10^{-3}、权重衰减 10^{-2}）、20个训练轮数和512的批量大小进行训练。它们报告的是未经校正的后验概率的宏平均 AUROC，因此其数值不同于第4.3节的编码强度；后者使用的是第3.4节中的探针。

为研究容量和训练时长，我们在随机种子 0 下，针对 BiomedCLIP 和 RAD-DINO 调整了 MLP 宽度（8、64 和 512 个隐藏单元）以及训练轮数（5 和 30）。在 NIH ChestX-ray14 上，参数扫描仅包括 5 轮 MLP。表5将这些参数扫描运行汇总为“MLP精简版”。主要配置（线性头和 30轮MLP 分类头）使用了三个随机种子；每个随机种子都会重新抽取患者划分、成员分配和头部初始化。我们报告均值 ± 标准差（s.d.）。

表3：效用和泄漏（三个随机种子的均值 ± 标准差）。“患者 AUC≥0.9”列给出了其记录最大值成员推断攻击曲线下面积至少为 0.9 的患者占比。

数据集	编码器 / 分类头	测试 AUROC	成员推断攻击曲线下面积	真正率@10^{-3}假正率（%）	患者曲线下面积 p99	患者曲线下面积≥0.9 (%)
CheXpert	BiomedCLIP / 线性	0.733 ± 0.003	0.568 ± 0.000	0.47 ± 0.04	0.725 ± 0.001	0.0 ± 0.0
	BiomedCLIP / 多层感知机	0.731 ± 0.004	0.698 ± 0.001	1.62 ± 0.06	0.939 ± 0.003	2.9 ± 0.0
	DINOv2 / 线性	0.714 ± 0.003	0.653 ± 0.000	1.37 ± 0.05	0.869 ± 0.005	0.6 ± 0.0
	DINOv2 / 多层感知机	0.715 ± 0.005	0.827 ± 0.001	5.34 ± 0.09	0.993 ± 0.001	28.3 ± 0.1
	RAD-DINO / 线性	0.772 ± 0.006	0.694 ± 0.001	2.88 ± 0.12	0.918 ± 0.004	1.4 ± 0.1
	RAD-DINO / 多层感知机	0.766 ± 0.005	0.907 ± 0.000	12.43 ± 0.22	1.000 ± 0.000	59.6 ± 0.4
NIH CXR14	BiomedCLIP / 线性	0.751 ± 0.010	0.566 ± 0.001	0.94 ± 0.10	0.725 ± 0.005	0.1 ± 0.0
	BiomedCLIP / 多层感知机	0.768 ± 0.006	0.618 ± 0.001	1.30 ± 0.02	0.908 ± 0.001	1.1 ± 0.0
	RAD-DINO / 线性	0.799 ± 0.008	0.692 ± 0.001	4.80 ± 0.25	0.932 ± 0.004	1.5 ± 0.0
	RAD-DINO / 多层感知机	0.804 ± 0.003	0.810 ± 0.002	7.16 ± 0.09	0.998 ± 0.000	16.8 ± 0.3

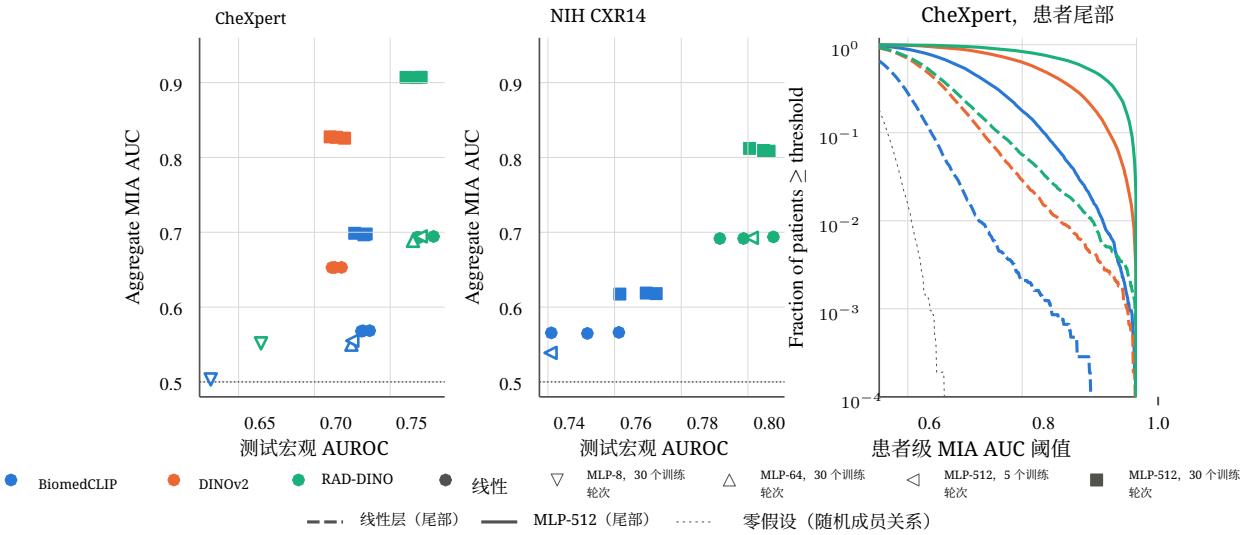

图2：泄漏与效用相互独立。(a, b) CheXpert 和 NIH ChestX-ray14 上聚合成员推断攻击AUC与测试 AUROC 的关系。(c) CheXpert 上患者级别成员关系 AUC 的生存曲线；灰色曲线显示随机成员关系下的零假设。

4 结果

4.1 泄漏取决于编码器和分类头，而非效用

泄漏程度差异很大，而测试性能变化甚微（表3和图2）。在 CheXpert 上，采用 RAD-DINO 的 30轮MLP 达到 0.907 的聚合成员推断攻击AUC。其患者中有 59.6% 的AUC高于 0.9，且在假正率为 10^{-3} 时，真正率达到 12.4 %。采用 BiomedCLIP 的线性头对应的数值分别为 0.568、0.0% 和 0.5%。其测试 AUROC 略低，为 0.733；相比之下，RAD-DINO MLP 为 0.766，RAD-DINO 线性头为 0.772。

对于这三种编码器，多层感知机头比线性头泄漏更多，但测试集 AUROC 并未提高。BiomedCLIP 的成员推断攻击曲线下面积从 0.568 升至 0.698，DINOv2 从 0.653 升至 0.827，RAD-DINO 从 0.694 升至 0.907。在 NIH ChestX-ray14 上也观察到了相同趋势：RAD-DINO 的成员推断攻击曲线下面积从 0.692 升至 0.810，BiomedCLIP 从 0.566 升至 0.618。这一结果与端到端模型中报告的容量效应 [7]一致，并表明冻结特征上的小型头部也存在相同效应。不同随机种子间的差异可以忽略，成员推断攻击曲线下面积的标准差至多为 0.002。

表 4：针对 MLP 宽度和训练轮数的容量与训练时长扫描（随机种子 0）。

编码器	Head	测试 AUROC	成员推断攻击曲线下面积	真正率@10^{-3}假正率（%）	患者曲线下面积≥0.9（%）
CheXpert / BiomedCLIP	线性	0.736	0.568	0.51	0.0
	MLP-8，30 个训练轮次	0.632	0.503	0.11	0.0
	MLP-64，30 个训练轮次	0.724	0.550	0.24	0.0
	MLP-512，5 个训练轮次	0.725	0.555	0.35	0.0
	MLP-512，30 个训练轮次	0.734	0.698	1.67	2.9
CheXpert / RAD-DINO	线性	0.779	0.694	3.02	1.4
	MLP-8，30 个训练轮次	0.665	0.552	0.41	0.0
	MLP-64，30 个训练轮次	0.765	0.689	1.54	1.0
	MLP-512，5 个训练轮次	0.770	0.694	2.46	2.9
	MLP-512，30 个训练轮次	0.770	0.907	12.63	59.3
NIH CXR14 / BiomedCLIP	线性	0.752	0.565	0.82	0.1
	MLP-512，5 个训练轮次	0.741	0.539	0.37	0.0
	MLP-512，30 个训练轮次	0.762	0.617	1.28	1.2
NIH CXR14 / RAD-DINO	线性	0.799	0.692	4.58	1.4
	MLP-512，5 个训练轮次	0.801	0.693	4.02	2.8
	MLP-512，30 个训练轮次	0.806	0.808	7.05	16.9

4.2 训练时长与容量控制的泄漏

泄漏并非仅由编码器决定（表 4；本小节中的所有值均来自随机种子 0）。在 CheXpert 上使用 RAD-DINO 时，训练 5 轮的 512 单元多层感知机，其成员推断攻击曲线下面积与线性头相同（0.694），测试集 AUROC 为 0.770。将同一多层感知机训练 30 轮后，成员推断攻击曲线下面积升至 0.907，但效用没有提升。只有 8 个单元的窄型多层感知机几乎没有泄漏（ 0.552 ），但效用有所下降（测试集 AUROC 为 0.665）。64 单元的多层感知机泄漏程度已与线性头大致相当（0.689）。Bio medCLIP 也呈现相同趋势：512 单元多层感知机训练 5 轮后的成员推断攻击曲线下面积为 0.555，训练 30 轮后为 0.698；相应的测试集 AUROC 分别为 0.725 和 0.734。在 NIH ChestX-ray14 上，RAD-DINO 的成员推断攻击曲线下面积从 0.693 升至 0.808。在该数据集上，缩短训练并非总是没有代价：对于 BiomedCLIP，训练 5 轮的多层感知机相较于训练 30 轮的多层感知机，测试集 AUROC 低约 0.02（0.741 对 0.762）。5轮设置是提前停止的固定替代方案，不采用基于验证集的早停。

4.3 人口统计特征编码与患者脆弱性在统计上无关

三个编码器都以线性方式对人口统计特征进行了编码。在 CheXpert 上，BiomedCLIP 对年龄、种族和性别的编码宏平均 AUROC 分别为 0.78、0.68 和 0.89；DINOv2 分别为 0.85、0.76 和 0.98；RAD-DINO 分别为 0.90、0.78 和 1.00。在 NIH ChestX-ray14 上，BiomedCLIP 对年龄和性别的值分别为 0.72 和 0.94，RAD-DINO 分别为 0.87 和 1.00。
然而，在同一模型内，对患者人口统计特征的编码强度并不能预测该患者的脆弱性（图3和表5）。在 CheXpert 上，全部 72 个关联（属性 $\times$ 头部配置 $\times$编码器 $\times$ 随机种子）均等于零。它们的绝对偏相关最大为 0.035，最宽95%置信区间为 ± 0.056，而 ΔR^2 上的最大上界为 0.0018。在 NIH ChestX-ray14 上，28 个关联中有 22 个通过了ρ 标准，28 个中有 27 个通过了 ΔR^2 标准。 ρ 标准的全部 6 个例外均为针对年龄的完全训练 MLP 运行，其 ρ 介于 -0.145 和 -0.088 之间； ΔR^2 标准的唯一例外（上界为 0.0114）也出现在同一组运行中。其符号为负，因此更强的年龄编码与更低的泄漏相伴。对于这些运行，我们因此只能声称编码强度没有增加泄漏，而不能声称该关联等于零。

在模型层面，编码强度和泄漏对编码器的排序一致：RAD-DINO 排名第一，DINOv2 第二，Bio medCLIP 第三。这一观察虽有启示性，但尚无定论，因为它仅基于三种编码器，且患者层面关联为零假设。我们将其解读为可能存在某种共同驱动因素（例如嵌入丰富度），而非因果关系的证据。

4.4 患者非典型性而非人口统计特征可预测脆弱性

在记录最大值评分下，仅记录数量的对数就解释了患者脆弱性的 ΔR^2 0.19 至 0.33（图4）。当根据记录均值或一条随机记录（$\Delta R^2 \leq 0.04$）对患者进行评分时，这一效应基本消失。例外是 RAD-DINO 线性头，其值分别为 0.143 和 0.108。因此，记录数量效应主要反映了从更多次抽取中取最大值。其余解释方差来自患者非典型性的特征。这些特征包括发现数量（RAD-DINO 最高可达 0.124

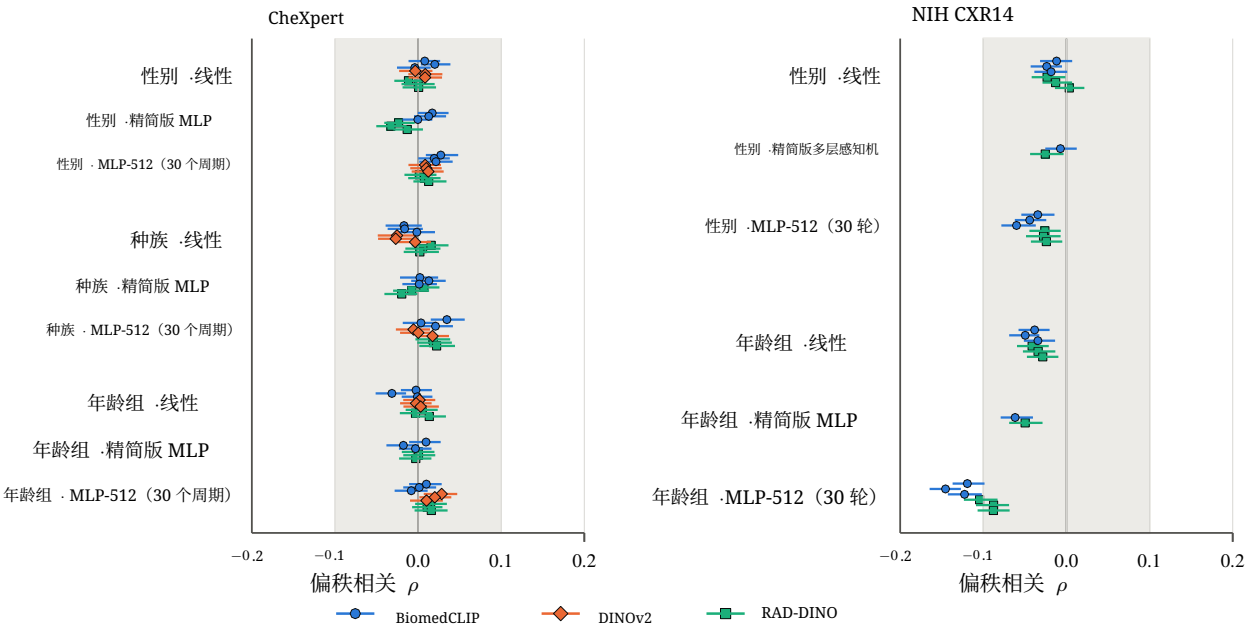

图3：人口统计编码与成员脆弱性之间的偏秩相关。各点表示单次运行及其95%置信区间，阴影带表示 ±0.1的等效区域。

表5：按数据集、分类头配置和属性划分的等效性检验。对于线性和30轮MLP配置，Runs统计编码器与随机种子组合的数量。“精简版 MLP”汇总了BiomedCLIP和RAD-DINO的seed-0参数扫描运行结果（CheXpert上的8单元和64单元MLP，以及5训练轮次的512单元MLP；NIH ChestX-ray14上的5训练轮次MLP），因此Runs统计编码器与配置组合的数量。

数据集	分类头配置 / 属性	Runs	部分 ρ范围	最宽95%置信区间	ρ等效	max ΔR^2 （置信区间上限）	ΔR^2等效
CheXpert	线性 / 性别	9	[-0.012, +0.020]	[-0.029, +0.039]	9/9	0.0009	9/9
	线性 / 种族	9	[-0.027, +0.016]	[-0.049, +0.037]	9/9	0.0013	9/9
	线性 / 年龄组	9	[-0.032, +0.013]	[-0.051, +0.034]	9/9	0.0016	9/9
	MLP-512，30 个训练轮次 / 性别	9	[+0.001, +0.027]	[-0.016, +0.048]	9/9	0.0014	9/9
	MLP-512，30 个训练轮次 / 种族	9	[-0.005, +0.035]	[-0.027, +0.056]	9/9	0.0018	9/9
	MLP-512，30 个训练轮次 / 年龄组	9	[-0.008, +0.028]	[-0.028, +0.047]	9/9	0.0011	9/9
	精简版 MLP / 性别	6	[-0.033, +0.017]	[-0.050, +0.037]	6/6	0.0015	6/6
	精简版 MLP / 种族	6	[-0.020, +0.013]	[-0.041, +0.033]	6/6	0.0012	6/6
	精简版 MLP / 年龄组	6	[-0.018, +0.010]	[-0.038, +0.027]	6/6	0.0009	6/6
NIH CXR14	线性 / 性别	6	[-0.024, +0.003]	[-0.043, +0.022]	6/6	0.0012	6/6
	线性 / 年龄组	6	[-0.050, -0.029]	[-0.069, -0.010]	6/6	0.0030	6/6
	MLP-512，30 个训练轮次 / 性别	6	[-0.060, -0.024]	[-0.078, -0.005]	6/6	0.0027	6/6
	MLP-512，30 个训练轮次 / 年龄组	6	[-0.145, -0.088]	[-0.165, -0.068]	0/6	0.0114	5/6
	精简版 MLP / 性别	2	[-0.025, -0.007]	[-0.044, +0.012]	2/2	0.0010	2/2
	精简版 MLP / 年龄组	2	[-0.062, -0.050]	[-0.079, -0.029]	2/2	0.0039	2/2

多层感知机）、罕见发现和稀疏嵌入邻域（BiomedCLIP 多层感知机最高可达 0.064），以及支持设备（最高为 0.021）。在每种评分下，人口统计编码的贡献均约为零。替代评分下的总 R^2 为 0.07 至 0.30，因此，这些特征无法解释患者层面方差中的大部分。

4.5 消除人口统计特征会改变公平性，而非泄漏

LEACE 显著减少了可通过线性方式解码的人口统计信息（表6）。非线性探针仍然能够恢复性别信息，其中 RAD-DINO 的多层感知器探针宏平均AUROC 为 0.994，因此擦除并不完全。线性探针也保留了一些信息（例如，RAD-DINO 性别信息的值为 0.683）。一个可能的原因是，LEACE 是基于记录拟合的，而探针使用的是患者均值嵌入。泄漏没有发生变化（图5a）。成员推断攻击曲线下面积为

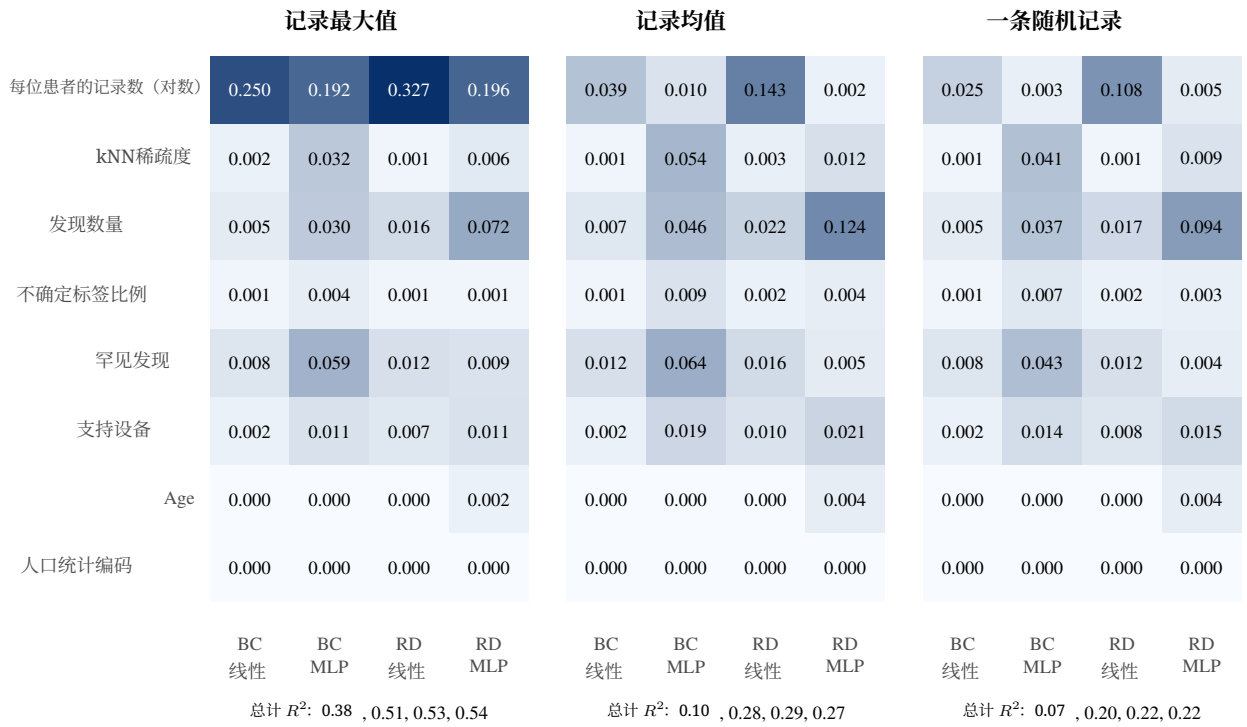

图4：在 CheXpert 上，针对患者脆弱性，在三种患者评分（记录最大值、记录均值和一个随机记录）下，每个特征块的增量 R^2 。单元格显示 seed-0 运行中移除每个特征块后 ΔR^2 的变化，Total R^2 为完整模型的 R^2 。BC：BiomedCLIP；RD：RAD-DINO。

表6：使用第3.7节所述探针，在 CheXpert 上对 LEACE 处理前后的患者均值嵌入进行探针宏平均AUROC评估。宏平均AUROC机会水平为 0.5。

		线性探针			MLP探针		
编码器	嵌入	Age	Race	Sex	Age	Race	Sex
BiomedCLIP	原始	0.746	0.704	0.893	0.820	0.683	0.914
	LEACE	0.442	0.512	0.564	0.561	0.591	0.794
RAD-DINO	原始	0.872	0.802	0.998	0.853	0.804	0.998
	LEACE	0.511	0.541	0.683	0.692	0.666	0.994

在擦除前后，BiomedCLIP 线性头的值均为 0.568。BiomedCLIP 多层感知机的值从 0.698 变为 0.700，RAD-DINO 线性头的值从 0.694 变为 0.695，而 RAD-DINO 多层感知机的值从 0.907 变为 0.913。线性头的值来自一个随机种子，多层感知机的值来自三个随机种子。测试集 AUROC 的代价对于 BiomedCLIP 约为 0.02，对于 RAD-DINO 约为 0.01。

相比之下，擦除会影响公平性（图 5b）。在三个随机种子下使用 MLP 头，BiomedCLIP 上 40 岁以下患者的漏诊差距从 0.167 降至 0.054。差值为 0.113（95%置信区间 [0.073, 0.153]），且三个随机种子下方向一致。在 RAD-DINO 上，同一差距从 0.136 降至 0.101（差值 0.035，95%置信区间 [0.027, 0.042]）。BiomedCLIP 上 40 至 60 岁患者的差距从 0.086 降至 0.035，RAD-DINO 上则从 0.069 降至 0.051。对于 80 岁及以上患者，BiomedCLIP 上的差距改变了符号，从 -0.027 变为 $+0.035$。BiomedCLIP 上的黑人与白人差距从 0.095 缩小至 0.044，但其置信区间包含零，因此我们仅将这一变化报告为方向性变化。在 RAD-DINO 上，这一差距变化不大，从 0.106 降至 0.096。性别差距和亚裔差距始终接近于零。这些置信区间基于三个随机种子的 a t-区间（自由度为 2），因此较为粗略。

综合来看，这些结果区分了这两个问题。移除线性模型可读取的年龄信息，降低了年龄相关差异，但并未降低患者作为训练集成员被识别的可能性。

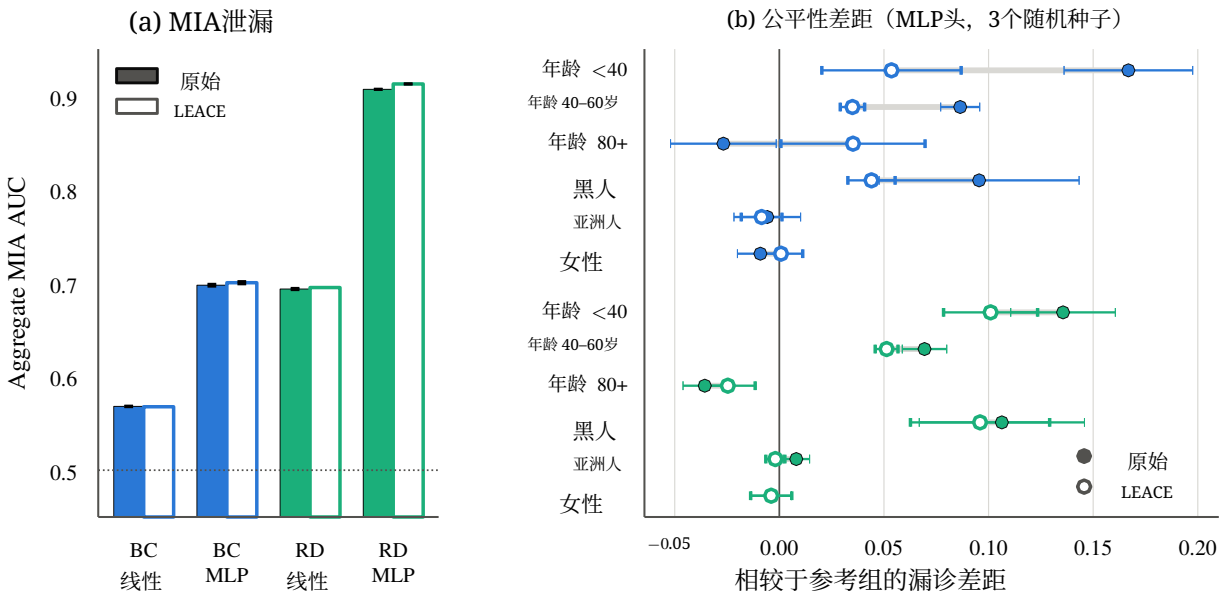

图 5：LEACE 对 CheXpert 上的信息泄漏和公平性的影响。(a) 擦除前后的聚合成员推断攻击AUC（线性头：一个随机种子；MLP头：三个随机种子的均值）。(b) 相对于参考组的 MLP 头漏诊差距。数值为三个随机种子的均值，误差棒表示随机种子间的 ± 1 标准差。

4.6 NIHChestX-ray14 上的结果

NIH ChestX-ray14 重现了在 CheXpert 上观察到的泄漏随编码器和容量变化的排序（表 3 和表 4）。然而，公平性结果有所不同。差距较小，参照年龄组为 40 至 60 岁。年龄为 60 至 80 岁的患者所对应的差距范围为 -0.057 至 -0.098，女性的差距范围为 $+0.008$ 至 $+0.041$。因此，我们不会合并不同数据集的公平性结果。

5 讨论

直觉上认为编码会导致暴露，因此公平性和隐私问题可以通过同一种补救措施解决，但我们的数据并不支持这一观点。编码强度无法预测哪些患者会发生泄漏（第4.3节），擦除线性可解码的人口统计学信息也不会改变泄漏情况（第4.5节），尽管非线性人口统计信息仍然存在。相反，泄漏取决于分类头的训练方式，以及患者的特征有多 atypical（第 4.2 和 4.4 节）。不过，擦除确实减少了年龄相关的诊断不足。因此，在我们的设置中，编码与公平性相关，却与成员身份暴露无关；从业者应分别审计这两类风险。

这些结果对冻结嵌入管线有三项实际启示。首先，使用线性头或缩短训练时间，可以大幅减少患者级泄漏，同时几乎不产生效用损失。对于 CheXpert 上的 RAD-DINO，成员推断攻击曲线下面积从训练30轮的多层感知机的0.907降至线性头的0.694，而测试 AUROC 没有损失；相比之下，在 NIH ChestX-ray14 上使用 BiomedCLIP 时，采用线性头或缩短训练时间会使测试 AUROC 下降约0.02。其次，基于患者记录中最大值的分析会夸大图像数量的作用。第三，最容易受到影响的患者具有不寻常的检查发现或稀疏嵌入邻域。这些特征并非人口统计属性，这表明最容易受到成员推断暴露影响的患者群体，未必与面临漏诊风险的群体重合。

我们的容量结果与 Knolle 等人 [7], 的发现一致：泄漏会随着模型容量的增加而增加，患者尾部的群体富集程度部分取决于群体规模。关于冻结特征，我们补充了两点观察：训练时长与头部宽度同样重要；一旦模型纳入了异常性特征，人口统计编码便无法预测患者脆弱性。

5.1 局限性

我们的结论适用于我们审计的设置：在两个公共数据集的胸部 X 光片冻结嵌入上训练小型分类头。这种设置相当于将基础模型用作固定的特征提取器，并且

这使我们能够训练成员身份受控的 200模型集成。相同的分离现象是否也适用于端到端微调模型、其他成像模态或差分隐私训练，仍有待测试。冻结嵌入也可能丢弃小病灶信号 [30]，因此骨干网络层级的泄漏情况可能有所不同。

若干设计选择决定了该干预和编码结果的适用范围。LEACE 仅移除线性信息，而非线性探针仍能恢复部分人口统计学信号（表6）。我们在所有记录上拟合了它，包括测试患者，因为我们将其用作分析性干预，而非可部署的缓解措施。同样，我们的编码得分也是线性的，因此衡量的是线性可解码的人口统计学信息。容量扫描和线性头部 LEACE 实验使用了一个随机种子，公平性差异的置信区间基于三个随机种子，而编码与泄漏在编码器层级上的一致性则基于三个编码器。因此，我们将这些结果视为支持性证据，主要结论则基于三个随机种子的分析。5轮设置是提前停止的固定代理。RAD-DINO 的预训练数据包括 CheXpert，因此其 CheXpert 结果构成领域匹配编码器的上界，而 DINOv2 仅是部分控制。

人口统计学标签也限定了分析范围。CheXpert 种族标签来自 Stanford AIMI 人口统计文件中的粗粒度、自我报告类别 [3]，其中西班牙裔族裔被归入种族群体，而“未知”和“其他”则被排除。NIH ChestX-ray14 没有种族标签，其公平性模式也不同于 CheXpert。我们未分析交叉性亚组。

6 结论

在人口统计特征编码、诊断公平性和成员脆弱性之间，基于胸部X光基础模型嵌入训练的分类头能够将三者区分开来。在全部72项 CheXpert 测试中，作为患者信息泄漏预测因素的编码强度在统计上均等于零。对人口统计特征进行线性擦除后，泄漏程度保持不变，但年龄相关的诊断不足有所减少，尽管非线性人口统计信息仍然存在。相反，泄漏程度与分类头容量、训练时长、嵌入家族和患者非典型性相关。因此，应分别对公平性和隐私进行审计，而分类头和训练时长等简单选择便可显著降低暴露风险。未来的研究应检验这些发现是否适用于端到端微调模型、其他成像模态以及差分隐私训练。

数据和代码可用性

两个数据集均已公开。CheXpert [21] 由 Stanford ML Group 发布，NIH ChestX-ray14 [22] 由 NIH Clinical Center 发布。CheXpert 的种族标签取自 Stanford AIMI 发布的人口统计文件 [3]。我们通过 Kaggle 获取了数据副本：CheXpert-v1.0-small（https://www.kaggle.com/datasets/ashery/chexpert）和 NIH Chest X-rays（https://www.kaggle.com/datasets/nih-chest-xrays/data）。源代码可在 https://github.com/SAMZ2610/privacy_bias_fairness 获取。

科学写作中生成式人工智能的使用声明

在准备这项工作期间，作者使用 Claude Sonnet 5.5 改进英文表达、纠正语法并提升文本的整体流畅度。使用该工具后，作者根据需要审阅并编辑了内容，并对出版物的最终内容承担全部责任。

利益冲突声明

作者声明，不存在可能影响本文所报告工作的已知竞争性财务利益或个人关系。

CRediT 作者贡献声明

SaharAskari：概念化、方法学、软件、可视化、调查、撰写—初稿、撰写—审阅与编辑。**AminGolnari**：撰写—审阅与编辑。

参考文献

[1] Laleh Seyyed-Kalantari、Haoran Zhang、Matthew B. A. McDermott、Irene Y. Chen 和 Marzyeh Ghassemi。医疗服务不足的患者群体中应用于胸部X光片的人工智能算法的漏诊偏差。《自然医学》，27：2176–2182，2021。[2] Laleh Seyyed-Kalantari、Guanxiong Liu、Matthew McDermott、Irene Y. Chen 和 Marzyeh Ghassemi。CheXclu-sion：深度胸部X线分类器中的公平性差距。载于 太平洋生物计算研讨会，第26卷，第232–243页，2021。[3] Judy Wawira Gichoya、Imon Banerjee、Ananth Reddy Bhimireddy 等。医学影像中的人工智能患者种族识别：一项建模研究。《柳叶刀数字健康》，4(6)：e406–e414，2022。[4] Ben Glocker、Charles Jones、Mélanie Bernhardt 和 Stefan Winzeck。胸部X线疾病检测模型中对受保护特征的算法编码。《*eBioMedicine*》，89：104467，2023。[5] Reza Shokri、Marco Stronati、Congzheng Song 和 Vitaly Shmatikov。针对机器学习模型的成员推断攻击。载于 *IEEE*安全与隐私研讨会（*S&P*），2017。[6] Nicholas Carlini、Steve Chien、Milad Nasr、Shuang Song、Andreas Terzis 和 Florian Tramèr。基于第一性原理的成员推断攻击。载于 *IEEE*安全与隐私研讨会（*S&P*），2022。arXiv:2112.03570。[7] Moritz A. Knolle、Martin J. Menten、Friederike Jungmann、Felix Meissen、Ben Glocker、Daniel Rueckert 和 Georgios Kaissis。医疗人工智能中的差异性隐私风险。《自然》，656：192–198，2026。[8] Robert Geirhos、Jörn-Henrik Jacobsen、Claudio Michaelis、Richard Zemel、Wieland Brendel、Matthias Bethge 和 Felix A. Wichmann。深度神经网络中的捷径学习。《自然·机器智能》，2：665–673，2020。[9] Yeonsu Kim、Yangwon Kim、Yoojin Nam、Namjoon Kim 和 Pa Hong。MIMIC-CXR胸部 X 光基础模型中患者特征的编码与线性使用。《*Diagnostics*》，16(13)：2030，2026。[10] Filipe Santos、Aldo Marzullo、Alessandro Quarta、João M. C. Sousa、Susana M. Vieira、Leo Anthony Celi、Francesco Calimeri 和 Laleh Seyyed-Kalantari。X光基础嵌入中的持久性人口统计信息：医疗保健领域安全且公平部署的风险。载于 *ICML 2025*可操作解释性研讨会，2025。[11] Hongyan Chang 和 Reza Shokri。论算法公平性的隐私风险。载于*IEEE*欧洲安全与隐私研讨会（*EuroS&P*），第292–303页，2021。[12] Bogdan Kulynych、Mohammad Yaghini、Giovanni Cherubin、Michael Veale 和 Carmela Troncoso。对成员推断攻击的差异性脆弱性。《隐私增强技术论文集》，2022(1)：460–480，2022。[13] Umid Suleymanov、Ilhama Novruzova、Khalid Mammadov、Natavan Hasanova 和 Murat Kantarcioglu。审计公平性与隐私权衡：公平性增强算法的亚群体层面影响，2026。arXiv:2607.14607。[14] Soroosh Tayebi Arasteh、Alexander Ziller、Christiane Kuhl 等。维护医学影像私有大规模人工智能模型中的公平性与诊断准确性。《医学通讯》，4：46，2024。[15] Salman Ul Hassan Dar 等。无条件潜在扩散模型会记忆患者影像数据。《自然·生物医学工程》，2025。[16] Santhosh Parampottupadam、Sinem Sav、Dimitrios Bounias、Saikat Roy、Klaus Maier-Hein、Adam Dziedzic、Franziska Boenisch 和 Ralf Floca。罕见、独特且被记忆：审计微调医学基础模型中的记忆化。载于 国际机器学习会议（*ICML*），2026。[17] Yuzhe Yang、Haoran Zhang、Judy W. Gichoya、Dina Katabi 和 Marzyeh Ghassemi。现实世界泛化中公平医疗影像人工智能的局限。《自然医学》，30(10)：2838–2848，2024。[18] Eike Petersen、Enzo Ferrante、Melanie Ganz 和 Aasa Feragen。医学影像中的人口统计学不变模型和表征是否公平？，2023。arXiv:2305.01397。[19] Shauli Ravfogel、Yanai Elazar、Hila Gonen、Michael Twiton 和 Yoav Goldberg。将其归零：通过迭代零空间投影保护受保护属性。载于 计算语言学协会第*58*届年会论文集，第7237–7256页，2020。[20] Nora Belrose、David Schneider-Joseph、Shauli Ravfogel、Ryan Cotterell、Edward Raff 和 Stella Biderman。LEACE：以闭式形式实现完美线性概念擦除。载于 《神经信息处理进展》，第36卷，2023。arXiv:2306.03819。

[21] Jeremy Irvin、Pranav Rajpurkar、Michael Ko 等。CheXpert：包含不确定性标签并与专家进行比较的大型胸部X线片数据集。载于*AAAI*人工智能会议论文集，第33卷，第590–597页，2019年。[22] Xiaosong Wang、Yifan Peng、Le Lu、Zhiyong Lu、Mohammadhadi Bagheri 和 Ronald M. Summers。ChestX-Ray8：医院规模的胸部X线数据库，以及常见胸部疾病弱监督分类与定位的基准。在*IEEE*计算机视觉与模式识别会议（*CVPR*）上发表，第2097–2106页，2017年。[23] Sheng Zhang、Yanbo Xu、Naoto Usuyama 等。BiomedCLIP：基于一千五百万对科学图像-文本对预训练的多模态生物医学基础模型，2023年。arXiv:2303.00915。[24] Maxime Oquab、Timothée Darcet、Théo Moutakanni 等。DINOv2：无监督学习稳健视觉特征，2023年。arXiv:2304.07193。[25] Fernando Pérez-García、Harshita Sharma、Sam Bond-Taylor 等。探索超越文本监督的可扩展医学图像编码器。《自然·机器智能》，2025年。[26] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于国际学习表征会议（*ICLR*），2019年。arXiv:1711.05101。[27] W. J. Youden。诊断试验评价指数。《癌症》，3(1)：32–35，1950年。[28] Donald J. Schuirmann。比较双单侧检验程序与效能法以评估平均生物利用度的等效性。《药代动力学与生物药剂学杂志》，15(6)：657–680，1987年。[29] Daniël Lakens。等效性检验：关于 t 测试、相关分析和荟萃分析的实用入门。《社会心理学与人格科学》，8(4)：355–362，2017年。[30] Raajitha Muthyala、Zhenan Yin、Alekhya Jilla、Frank Li、Theo Dapamede、Bardia Khosravi、Mohammadreza Chavoshi、Judy Gichoya 和 Saptarshi Purkayastha。冻结的基础模型嵌入会丢失胸部放射摄影中的小病灶信号：对部署前评估的启示，2026年。arXiv:2606.11606。