arXiv:2302.00319v1 [cs.LG] 1 Feb 2023

基于无监督和半监督深度学习方法开发能够反映发病率和死亡率的深度生物年龄

Seong-EunMoon*[1], JiWon Yoon*[2],ShinyoungJoo[1],YoohyungKim[3], JaeHyun Bae[4], SeokhoYoon[5],HaanjuYoo†[1], 以及 Young Min Cho†[2,6,7]

[1]NAVER Cloud AI Lab[2]内科学系，首尔国立大学医院医疗系统江南中心[3]内科学系，首尔国立大学医院[4]内科学系，高丽大学安岩医院，高丽大学医学院[5]DaNaA Data[6]内科学系，首尔国立大学医学院[7]老龄化研究所，首尔国立大学[1]{seongeun.moon, sy.joon, haanju.yoo}@navercorp.com[2, 3]{jwyoonmd, yhkimmd, ymchomd}@snu.ac.kr[4]fermatah@gmail.com[5]seokho.yoon@danai.co.kr

摘要：

背景 深度学习技术能够基于大规模数据获取潜在表示，因此可用于发现新型衰老生物标志物；然而，现有的生物年龄估计深度学习方法通常依赖实际年龄，且未考虑死亡率和发病率，而这两者是衰老最重要的结局。

方法 本文提出了一种新型深度学习模型，用于学习与受试者发病率和死亡率相关的生物衰老潜在表示。该模型除利用健康检查数据外，还使用发病率和死亡信息，以学习衰老与测量所得临床属性之间的复杂关系。

研究结果 与 KDM 及其他基于学习的模型相比，所提出的模型在一个大型一般人群数据集上进行了评估。结果表明，所提出的模型得出的生物年龄在区分受试者的发病率和死亡率方面表现更优。

*贡献相同。
†通讯作者。

关键词： 生物年龄、表征学习

1 引言

随着公众对衰老的关注日益广泛，许多研究尝试对衰老程度进行定量测量。实际年龄是衡量衰老的传统指标，但正如先前研究所示 [1, 2], ，我们的身体会因遗传和环境因素而呈现不同程度的生物衰老。这促使许多研究人员寻找新的衰老生物标志物，他们相信对该标志物进行适当干预能够改善我们的生活质量。与实际年龄相对应，另一种衰老指标通常称为生物年龄。

尽管近年来基于DNA甲基化或端粒长度的生物年龄已得到广泛研究 [3–5], ，但这些方法需要进行DNA测序等昂贵的检查，因此难以获取大规模数据来开发通用的生物年龄，也难以将所开发的生物年龄应用于现实世界。因此，许多研究者也利用常规健康检查所获得的数据，例如血液检测和人体测量 [6, 7]。此外，已有研究提出了从原始数据中提取与衰老相关生物标志物的方法。例如，文献 [6]将主成分分析（PCA）应用于血液和尿液检测等常见检查结果。近期，研究者还提出了基于深度学习的方法，期望借此提取衰老的深层表征，并用于生物年龄估计 [8, 9]。

然而，以往的大多数研究都高度依赖实际年龄。生物年龄通常通过对实际年龄相关信息进行统计分析，或使用经训练以预测实际年龄的模型得出。这从根本上是因为缺乏衰老的真实指标，同时也因为研究者假设大多数（健康）人群的生物年龄与实际年龄相同。然而，这一假设并不适用于某些情况。例如，年老时没有疾病的受试者应被视为比其实际年龄所显示的更健康，而年轻时没有疾病则很常见。此外，由于营养状况改善和医疗保健发展，年轻一代的衰老模式可能不同于老年人群。

因此，要建立能够泛化的生物衰老模型，模型训练无需严格以实际年龄为导向。此外，在进行生物年龄建模时，应考虑年龄和发病率各不相同的人群，使模型能够理解发病率与年龄相关的特征。可以预见，年龄相关发病率的多样性会降低实际年龄预测的表现，因此需要专门用于评估生物年龄的指标。

本文提出一种基于深度学习的生物年龄估计方法，在不依赖实际年龄的情况下，将年龄相关发病率和死亡率纳入生物年龄建模。我们的模型依据发病率和死亡信息捕捉受试者的健康状态，并计算生物年龄与实际年龄之间的差距。将预测得到的差距加到实际年龄上，即可获得生物年龄。在此，所提出的生物年龄以无监督方式学习，从而减少噪声监督目标（即实际年龄）可能造成的误差传播。同时，我们的模型基于以下四项内容进行训练：1）实际年龄预测；2）发病率负样本对；3）与死亡率的相关性；4）差距的语义一致性，以提取对生物年龄有意义的信息。

我们的贡献可总结如下：

- 将一个大规模队列纳入生物年龄建模，并纳入发病率和死亡信息。

- 我们提出一种基于自编码器的深度学习模型，用于根据健康体检数据估算生物年龄与实际年龄之间的差距。无监督结构使预测的生物年龄不易与实际年龄紧密关联，因为实际年龄可能存在较多噪声。

- 我们的模型通过将临床常识嵌入训练损失中，学习与生物年龄相关的发病率和死亡率信息。具体而言，我们为患病数据构造负样本对，并通过对比负样本对中的元素来训练模型，使其捕捉疾病带来的差异。此外，模型还利用死亡数据进行训练，以最大化距死亡天数与预测差距之间的负相关。

- 我们根据年龄相关发病率定义了多种研究人群，即超常组、正常、平均和*Whole*组。我们进行了全面的实验，以评估这些组在生物年龄建模中的代表性。

2 相关工作

传统方法： 生物年龄估计的传统方法包括多元线性回归（MLR）、基于主成分分析的方法（为方便起见，以下简称 PCA）以及 Klemera-Doubal方法（KDM）。MLR 是用于生物年龄估计的初始方法 [10, 11],，它将生物年龄估计为与衰老相关的生物标志物的加权和。尽管 MLR 易于实现和理解，但仍存在若干缺点。例如，MLR 模型往往会向目标值的均值回归，导致生物年龄在回归边缘出现偏差，并造成衰老速率不连续 [2]。

Nakamura等人提出采用主成分分析进行生物年龄估计 [12],，此后该方法已被众多研究采用 [13–17]。该方法会选取与实际年龄最相关且彼此独立的生物标志物，即主成分，用于生成生物年龄评分。主成分分析克服了多元线性回归的问题，但其估计结果的可解释性相对不如多元线性回归模型，因为主成分可能由多个生物标志物组合而成。此外，由于主成分分析是一种线性模型，因此无法学习变量的非线性特征。尽管主成分分析在理论上有所改进，Levine的研究显示，多元线性回归与主成分分析的性能并无显著差异 [6]。

Klemera和Doubal提出了一种称为KDM的数学模型，并称其是生物年龄估计的最优方法 [18]。KDM认为，衰老程度由随实际年龄系统性变化的生物标志物所表征。基于这一假设，生物年龄是各生物标志物相对于年龄相关平均值的距离的加权和。多项研究通过实验表明，KDM是估计死亡率相关生物年龄的最佳方法 [2, 6, 19], ；然而，近期的模型已开始超越KDM [7, 20]。KDM依赖于一个在现实世界中未必成立的假设，即生物标志物与生物年龄之间接近线性关系。此外，该方法依赖实际年龄，因此生物年龄的真实目标仍存在不确定性。

近期方法： 近年来，深度学习技术通过训练能够预测受试者实际年龄的深度神经网络，被用于生物年龄估计任务。值得注意的是，在 [8] 和 [21], 中，研究人员将 XGBoost 和随机森林等机器学习方法与具有多个全连接层的深度神经网络进行了比较。Pyrkov等人采用了一维卷积神经网络，并开展了类似的比较，将其与传统机器学习方法进行对照 [9]。不出所料，在实际年龄预测的准确性方面，深度学习模型优于其他方法。然而，

这些研究侧重于构建一个预测健康个体实际年龄的模型，并假定预测年龄代表其确切的健康状态。然而，众所周知，深度学习模型通常无法妥善处理分布外样本 [22–24]。因此，对于患有疾病的受试者，这类模型的生物年龄估计可能不稳定。我们表明，使用包含各种患病状况的数据，能够更稳健地测量生物年龄。

Rahman 和 Adjeroh 提出了一种基于年龄聚类的算法，用于生物年龄建模 [7]。首先，根据受试者的实际年龄对训练集中的受试者进行聚类，然后针对每个聚类，计算美国国家健康与营养调查（NHANES）中 16 项特征的均值和标准差。给定一名测试集受试者后，通过比较各聚类质心与该受试者特征之间的距离，选出邻近聚类。随后，对这些邻近聚类的实际年龄进行加权平均，得到该受试者的生物年龄。在死亡率预测方面，该方法取得了与深度神经网络相当的结果。然而，这种基于聚类的方法仍依赖实际年龄，且只有在每个年龄组都有充足数据时才能使用。此外，如果特征存在缺失观测，预测出的生物年龄可能不稳定。

此外，文献 [20] 还提出了一种将发病信息嵌入生物年龄的方法。该方法通过自编码器提取电子病历数据的潜在表示；自编码器基于重构数据与用于评估健康风险的风险评分之间的 KL散度进行训练。随后，采用类似主成分分析方法的公式计算生物年龄。基于风险评分的学习方向有助于模型学习给定数据的患病状况。与以往方法相比，这种无监督模型在疾病发病率预测方面表现更优。此外，生物年龄与实际年龄差距较大的群体，其风险评分也较高。

3 所提方法

3.1 目标

我们的目标是利用发病率和死亡率信息对受试者的健康状态进行编码，从而更可靠地估计生物年龄。我们的模型利用与 $[GAP]$ 特殊标记对应的潜在特征，估计生物年龄与实际年龄之间的差距。此外，模型根据 $[CA]$ 标记的特征预测实际年龄，以获取有关受试者健康状态的补充信息。我们利用受试者的发病情况进行生物年龄估计，这是基于这样一个事实：与健康受试者相比，患有疾病的受试者通常生物年龄更高。我们将与疾病相关的生物标志物的值替换为正常范围内的值，从而生成患病受试者数据的对比样本对。根据校正后的正常数据计算出的生物年龄，预期低于根据患病受试者的原始数据计算出的生物年龄。我们还通过训练所提出的模型，使其生成的生物年龄与距死亡天数呈负相关，从而将死亡率纳入考量。通过将与生物年龄相关的规则纳入所提出模型的损失函数，我们证明，相较于以往仅针对健康受试者的研究，所提出的模型能够估计包括患病受试者在内的一般人群的生物年龄。

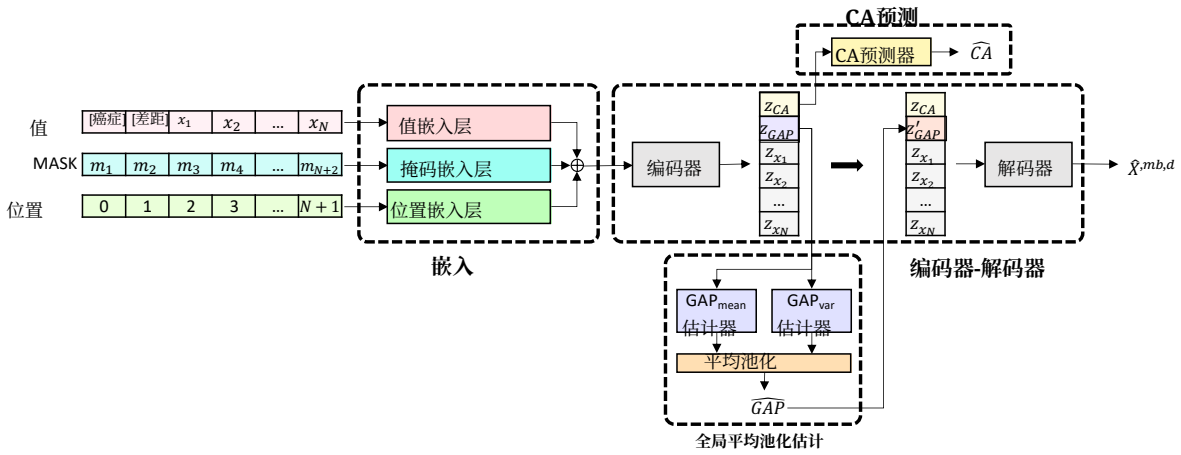


图1：所提出的方法模型概览。

3.2 模型架构

所提出模型的架构如图1所示。所提出的模型主要分为四个部分：嵌入、编码器-解码器、实际年龄预测和全局平均池化估计，且所有部分均联合训练。编码器-解码器模块通过重建任务学习提取用于描述受试者健康状态的潜在表示。实际年龄预测模块经过训练，以最小化实际年龄预测损失；这也会引导编码器将与实际年龄相关的信息放入 $[CA]$ 标记的潜在特征中。全局平均池化估计模块通过反映发病率和死亡率相关规则，学习估计生物年龄与实际年龄之间的差距。各模块的详细信息如下。

嵌入输入数据 $\mathbf{x} = [x_1, x_2, ..., x_N]$ 首先经过嵌入过程。我们在输入数据开头添加两个特殊标记 $[CA]$ 和 $[GAP]$，分别用于实际年龄预测和生物年龄估计。健康检查数据通常包含缺失变量，因此我们假设缺失值已采用适当的插补策略进行填补。通过插补掩码$\mathbf{m} = [m_1, ..., m_{N+2}]$来区分原始测量值和之后插补的特征。对于原始变量，掩码值为零；否则为一。输入数据的位置信息采用绝对方式表示，即两个特殊标记和 N个特征的位置使用整数 $\mathbf{p} = [0, ..., N+1]$进行编码。

虽然插补掩码和位置通过典型的嵌入层 e_m 和e_p进行嵌入，但由于大多数特征值是连续的，传统的嵌入方法难以应用于输入特征。因此，我们采用全连接层 e_f 进行特征值嵌入。嵌入后的特征值、插补掩码和位置信息通过求和操作整合。也就是说，

$$\mathbf{X}^{embed} = e_f(\mathbf{x}) + e_m(\mathbf{m}) + e_p(\mathbf{p}). \tag{1}$$

编码器-解码器 由于所提出的模型对编码器和解码器结构没有任何属性假设，因此，只要具备足够的计算能力，任何类型的学习网络都可以用于编码器-解码器模块。

我们为编码器和解码器选择了 Performer 网络 [25], ，它是一种采用线性自注意力的高效 Transformer。请注意，使用特殊标记来汇总相关信息的方式可能因网络结构而异。例如，对于循环神经网络，将特殊标记附加在特征向量末尾会更加自然。

编码器 E 接收 $\mathbf{X}^{embed}$ 并提取潜在特征 $\mathbf{Z} = [\mathbf{z}_{CA}, \mathbf{z}_{GAP}, \mathbf{z}^{x_1}, ...\mathbf{z}_{x_N}]$。解码器 D 使用潜在特征 $\mathbf{Z}$ 还原嵌入后的输入数据 $\hat{}\mathbf{X}^{embed}$ 。样本 K 的重构训练损失使用均方误差（ MSE ）计算，即

$$\mathcal{L}_{recon} = \frac{1}{K(N+2)} \sum_{i=1}^{K} \sum_{j=1}^{N+2} \left(\hat{\mathbf{X}}_{i,j}^{embed} - \mathbf{X}_{i,j}^{embed} \right)^2, \tag{2}$$

其中，$\mathbf{X}_i^{embed}{}_{,\ j}$ 表示第 i个样本的第 j个特征的嵌入描述。

CA 预测 给定特殊标记 $[CA]$的潜在特征，我们使用实际年龄预测模块中的两个带有非线性激活函数的全连接层来预测实际年龄。实际年龄预测的性能通过均方误差（MSE）和决定系数两种指标进行评估。当 CA 预测网络给出的预测值为 $\hat{}y$ ，而真实实际年龄为 y时，训练损失定义为:

$$\mathcal{L}_{CA} = \frac{1}{K} \sum_{i=1}^{K} (\hat{y}_i - y_i)^2 + \lambda_{CA_{R2}} \left(1 - \frac{\sum_{i=1}^{K} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{K} (y_i - \bar{y})^2} \right). \tag{3}$$

$\lambda_{CA_{R2}}$ 用于平衡均方误差与决定系数的值， $\bar{y}$ 表示 y的平均值。该损失函数旨在提升实际年龄预测性能，同时鼓励编码器提取与实际年龄相关的信息，并将提取的信息聚合到 $[CA]$ 标记的潜在特征中。

全局平均池化估计 我们将作为 $[GAP]$ token 潜在特征的$\mathbf{z}_{GAP}$参数化为概率分布。然后，使用重参数化技巧 [26]从其分布中采样差距估计值。我们使用最大均值差异损失（MMD） [27] ，使估计的差值$\mathbf{g}$ 遵循正态分布 $\mathcal{N}(\mathbf{0}, \mathbf{1})$。

$$\mathcal{L}_{dist} = \text{MMD}\left(p(\mathbf{g}|\mathbf{z}_{GAP}), \mathcal{N}(0,1)\right) \tag{4}$$

除了最大均值差异损失（MMD）之外，我们还引入一致性损失，以强制 z_{GAP} 与z_{CA}相互独立。我们通过在一个小批量中随机打乱 z_{CA} ，为$\mathbf{g}$ 构造一个正样本对。包含打乱后的 z_{CA} 的潜在特征记为$\mathbf{Z}'$。然后，我们最小化从原始$\mathbf{Z}$和 $E(D(\mathbf{Z}'))$:= $\mathbf{Z}^*$中获得的差值之间的均方误差（MSE）。因此，一致性损失可写为:

$$\mathcal{L}_{consist} = \frac{1}{K} \sum_{i=1}^{K} \left(f_{GAP_{mean}}(\mathbf{z}_{GAP}) - f_{GAP_{mean}}(\mathbf{z}^*_{GAP}) \right)^2, \tag{5}$$

其中， $f_{GAP_{mean}}$ 表示用于推断差距均值的网络。

为了引导模型反映受试者的发病率，我们使用对比损失，确保基于患病受试者数据得出的差距估计值与基于将生物标志物替换为正常范围内的值后得到的数据得出的差距估计值之间存在边际差异。例如，当根据某位受试者的血糖水平判断其患有糖尿病（DM）时，我们会构造一对数据，其中包括原始的健康

检查结果，以及将血糖水平调整至正常范围后的数据。发病率的对比损失可定义如下：

$$\mathcal{L}_{contrast} = \max\left(0, \frac{1}{K}\sum_{i=1}^{K}(\gamma - \mathbf{g}_{org} - \mathbf{g}_{corr})\right), \tag{6}$$

其中，γ 用于调节差距估计值的间隔。

此外，我们纳入死亡率损失，以促使估计的差值与距死亡天数呈负相关。相关性通过皮尔逊相关系数（PCC；有关 PCC 的详情，请参见第 4.3.1 节和公式 9）进行评估。

$$\mathcal{L}_{mort} = 1 + \mathrm{PCC}_{\mathbf{g},days_to_death} \tag{7}$$

因此，所提出的模型的总损失函数可概括如下：

$$\mathcal{L} = \lambda_1\mathcal{L}_{recon} + \lambda_2\mathcal{L}_{CA} + \lambda_3\mathcal{L}_{dist} + \lambda_4\mathcal{L}_{consist} + \lambda_5\mathcal{L}_{contrast} + \lambda_6\mathcal{L}_{mort}, \tag{8}$$

其中，λ_1、……、λ_6 是用于调节不同类型损失影响的权重值。

4 数据集与预处理

4.1 研究人群

我们采用了 2013 年 10 月至 2020 年期间收集的 151,283 名受试者的健康检查数据进行分析。数据共包含 88 项属性，包括人口统计信息、人体测量记录、血液检查结果、尿液检查结果等。此外，我们使用通过问卷收集的受试者临床病史来分析患病状态。具体而言，我们纳入了糖尿病（DM）、高血压（HBP）、血脂异常（DLP）、癌症、心血管疾病（CVD）和心血管意外（CVA）的诊断史，以及 DM、HBP 和 DLP 的用药状态。此外，我们还利用受试者的死亡信息（包括死亡日期及相应死因），考察预测的生物年龄与死亡率之间的关系。

受试者的癌症、CVD 和 CVA 患病状态依据诊断史确定；DM、HBP、DLP 和多发性硬化症（MS）的患病状态则依据健康检查和问卷结果确定，具体定义见表 1。未满足 MS 条件的受试者被视为 MS 患病状态正常。此外，对于 DM、HBP 和 DLP，既不符合正常条件也不符合患病条件的受试者，被视为处于相应疾病的前期阶段。提取患病状态的整体流程，以及被划分至各状态的受试者人数，见图 2。

4.2 用于生物年龄估计的代表性人群组

我们根据受试者的发病率考察了四类研究人群，以评估人群的代表性。首先，我们筛选出没有疾病的受试者，这对应于以往研究中常用的超常组 [8, 28]。正常组还包括处于糖尿病、高血压和 DLP 疾病前期的受试者。我们考虑了八项与疾病相关的特征来定义平均组，即糖化血红蛋白（糖尿病）、空腹血糖（糖尿病、多发性硬化症）、收缩压（高血压、多发性硬化症）、舒张压（高血压、多发性硬化症）、低密度脂蛋白胆固醇（DLP）、甘油三酯（DLP、多发性硬化症）、高密度脂蛋白胆固醇（DLP、多发性硬化症）和腰围（多发性硬化症）。疾病相关特征 $X^D = \{x_1^D, ..., x_8^D\}$ 处于以下范围内的受试者

表1：确定糖尿病、高血压、DLP和多发性硬化症患病状态的标准。特征缩写如下：FBS= 空腹血糖，SBP=收缩压，DBP=舒张压，LDLC=低密度脂蛋白胆固醇，TG=甘油三酯，HDLC=高密度脂蛋白胆固醇，WC=腰围。

	正常	疾病
DM	• FBS < 100 毫克/分升 AND • HbA1C < 5.7%	• FBS ≥ 126 毫克/分升 OR • HbA1c ≥ 6.5 % 或 • 正在服用糖尿病药物
HBP	• 收缩压 < 120mmHg 且 • 舒张压 < 80mmHg	• 收缩压 ≤ 140 mmHg 或 • 舒张压 ≥ 90 或 • 服用高血压药物
DLP	• 低密度脂蛋白胆固醇 < 100毫克/分升 且 • 甘油三酯 < 150 毫克/分升 且 • 高密度脂蛋白胆固醇 ≥ 60 毫克/分升	• 低密度脂蛋白胆固醇 ≥ 160 毫克/分升 或 • 甘油三酯 ≥ 200 毫克/分升 或 • 高密度脂蛋白胆固醇 < 40mg/dL 或 • 正在服用血脂异常药物
MS	-	符合以下三项或更多条件 • 腰围 ≥男性：90 cm，女性：85 cm • 甘油三酯 ≥ 150 mg/dL • 高密度脂蛋白胆固醇 <男性：40 mg/dL，女性：50 mg/dL • 收缩压 ≥ 135 mmHg 和舒张压 ≥ 85 mmHg • 空腹血糖 ≥ 100 mg/dL

Ta表 2：数据的数量、平均值和标准差统计数据。

属性	Male			女性		
	数量	Mean	Std.	数量	Mean	Std.
Age	78,847	46.41	12.07	72,434	45.67	12.47
死亡时间 [天]	2,240	2,794	1,676	1,106	2,665	1,689
体质指数 [kg/m^2]	76,431	24.62	11.27	70,762	21.92	3.31
HbA1c [%]	73,470	5.73	0.74	68,649	5.61	0.55

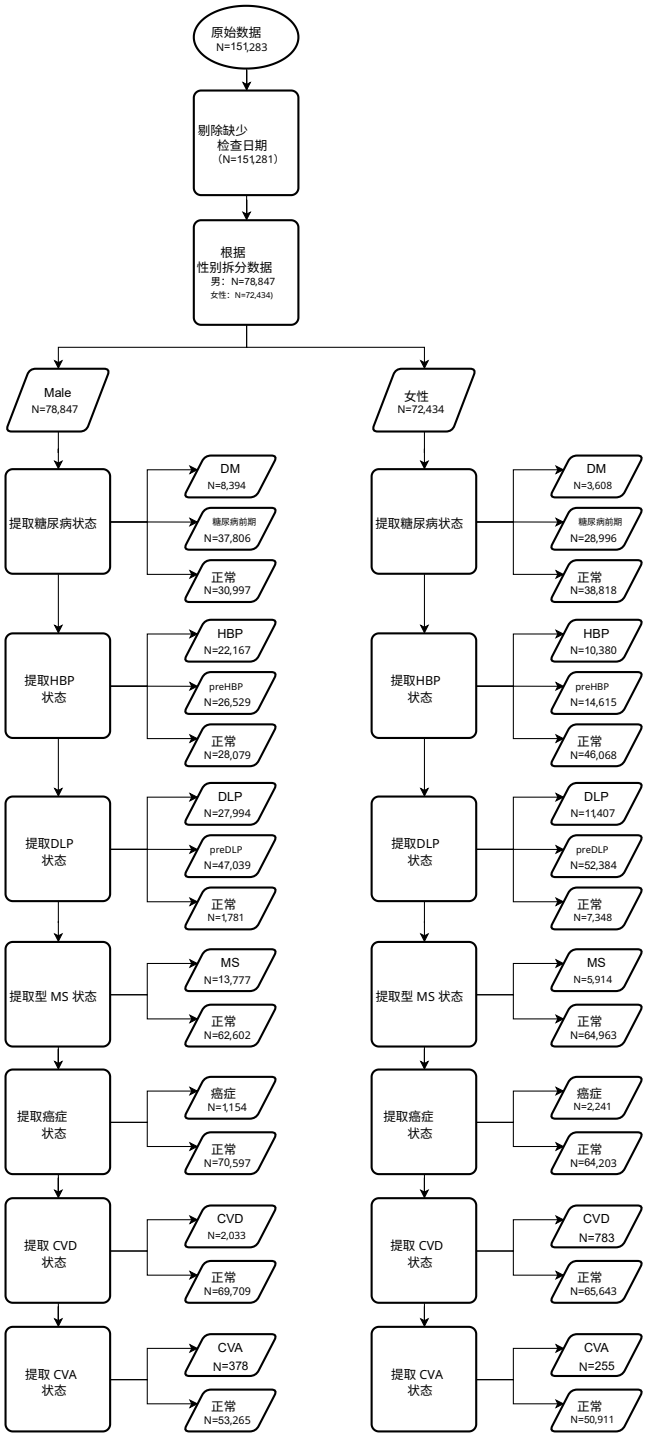


图 2：包含疾病标注的数据处理流程图。

表3：整体受试者（全部）、女性受试者（F）和男性受试者（M）的观测特征与实际年龄之间相关性分析的结果。

	Whole	超常组
All	血尿素氮、估算肾小球滤过率（CKD-EPI）、估算肾小球滤过率（MDRD）、第一秒用力呼气量、第一秒用力呼气容积与用力肺活量比值、用力肺活量、糖化血红蛋白、内脏脂肪面积	白蛋白、估算肾小球滤过率（CKD-EPI）、估算肾小球滤过率（MDRD）、第一秒用力呼气量、第一秒用力呼气容积与用力肺活量比值、用力肺活量百分位数
F	腹部脂肪量、估算肾小球滤过率（CKD-EPI）、估算肾小球滤过率（MDRD）、第一秒用力呼气量、第一秒用力呼气容积与用力肺活量比值、FVC，身高、收缩压、内脏脂肪区域，腰围	白蛋白、钙、估算肾小球滤过率（CKD-EPI）、估算肾小球滤过率（MDRD）、第一秒用力呼气量、第一秒用力呼气容积与用力肺活量比值、FVC，身高、幽门螺杆菌抗体、平均红细胞体积、红细胞计数
M	白蛋白、估算肾小球滤过率（CKD-EPI）、估算肾小球滤过率（MDRD）、第一秒用力呼气量、第一秒用力呼气容积与用力肺活量比值、用力肺活量、身高、糖化血红蛋白、细胞内液、矿物质、蛋白质，红细胞计数	估算肾小球滤过率（CKD-EPI）、估算肾小球滤过率（MDRD）、第一秒用力呼气量、第一秒用力呼气容积百分位数、第一秒用力呼气容积与用力肺活量比值、用力肺活量、内脏脂肪面积

$[\mu_{x_i^D} - 2\sigma_{x_i^D}, \mu_{x_i^D} + 2\sigma_{x_i^D}]$，其中 $\mu_{x_i^D}$和 $\sigma_{x_i^D}$表示特征 x_i^D的均值和标准差，归入平均 组。最后，整体 组指数据集中所有受试者。

对于每种人群，受试者按照 0.70、0.15 和 0.15 的比例随机划分为训练集、验证集和测试集。使用不同的随机种子重复进行数据划分三次。第 5 节中报告的最终结果为三次重复实验结果的平均值。

4.3 特征集的选择

4.3.1 实际年龄的相关性分析

为确定与衰老密切相关的生物标志物，我们使用皮尔逊相关系数（PCC）、斯皮尔曼等级相关系数（SROCC）和互信息（MI），分析了*Whole* 组和超常组 中特征与实际年龄之间的关系。两个变量 $X_1 = \{x_{1,1}, ..., x_{1,n}\}$和 $X_2 = \{x_{2,1}, ..., x_{2,n}\}$的 PCC 定义为:

$$\text{PCC}_{X_1,X_2} = \frac{\text{cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}, \tag{9}$$

其中，cov 表示协方差。PCC 的取值范围为 -1 至 1，分别表示完全负线性关系和完全正线性关系。如果两个变量之间不存在关系，则 PCC 为零。

SROCC 衡量排序变量的 PCC，用于评估变量之间的单调性，而非线性关系。也就是说，SROCC 的计算方式为:

$$\text{SROCC}_{X_1,X_2} = \frac{\text{cov}(R_{X_1}, R_{X_2})}{\sigma_{R_{X_1}}\sigma_{R_{X_2}}}, \tag{10}$$

其中， R_{X_1} 和 R_{X_2} 分别指 X_1 和 X_2的排序变量。如果不存在秩并列，因此变量具有 n 个不同的秩，SROCC 也可以表示为 $1 - \frac{6\sum_{i=1}^{n} R_{x_1,i} - R_{x_2,n}}{n^2-1}$。由于

SROCC 的计算方式与 PCC 相同，其取值也介于 -1 和 1 之间，但结果的解释依据是单调关系，而非线性关系。

MI 用于评估两个变量之间的信息相关性。当 X_1 和 X_2 相互独立时，它们之间的 MI 为零。X_1 和 X_2 的 MI 值越大，表明 X_1 包含的有关 X_2 的信息越多，反之亦然。对于概率分布 P_{X_1}、P_{X_2}及其联合分布 $P_{X_1X_2}$，MI 的计算公式为

$$\mathrm{MI}_{X_1,X_2} = \sum_{i=1}^{n} P_{X_1X_2}(x_{1,i}, x_{2,i}) \log \frac{P_{X_1X_2}(x_{1,i}, x_{2,i})}{P_{X_1}(x_{1,i}) P_{X_2}(x_{2,i})}. \tag{11}$$

根据香农熵 H，公式 11 可改写为 $H(X_1) - H(X_1|X_2)$ 或 $H(X_2) - H(X_2|X_1)$。据此，我们可以将 MI 值解释为：通过观测 X_2 (X_1)，关于 X_1(X_2)的不确定性减少量。

与实际年龄的 PCC、SROCC 或 MI 值位于前 10% 的特征列于表 3。我们选取了通常在健康检查中测量的特征，并将缺失率超过 50% 的特征排除在分析之外。

值得注意的是，与肺功能相关的特征，即第一秒用力呼气量、第一秒用力呼气量与用力肺活量之比以及用力肺活量，在所有情况下均被选中。比较*Whole* 人群与超常组 的结果可见，糖化血红蛋白和收缩压等与慢性疾病相关的特征仅在*Whole* 组中具有显著性。

由于用力肺活量百分位数和估算肾小球滤过率的计算包含实际年龄，因此不将其视为与实际年龄相关的特征。此外，基于生物电阻抗测量计算得出的内脏脂肪面积、腹部脂肪量、细胞内液、矿物质和蛋白质，因其准确性存在不确定性而被舍弃。身高也与实际年龄呈显著相关。然而，这可能反映的是生长时期营养状况的改善，而非衰老的影响。

4.3.2 特征选择

根据相关性分析结果和专家知识，我们选取了十二个特征来描述健康状态的各个方面。

- **贫血：** 我们选取红细胞计数（RBC）和平均红细胞体积（MCV）这两个特征来描述贫血状况；贫血在老年人中较为常见。
- **肥胖：** 我们使用腰围（WC）、体质指数（BMI）、骨骼肌质量（SMM）、体脂肪量（BFM）和去脂体重（FFM）作为肥胖指标。请注意，我们采用了FR和MR，而非基于计算的特征。
- **炎症：** 我们加入了高敏感性C反应蛋白（hsCRP），用于描述体内的炎症程度。
- **肾功能：** 我们没有选用eGFR特征，而是选择了肌酐（Cr），因为它是计算eGFR所需的原始数据。
- **肺功能：** 根据相关性分析的结果，我们选取第一秒用力呼气量和用力肺活量来评估肺功能。
- **代谢：** 糖化血红蛋白（HbA1c）被认为能够代表身体的代谢状态。代谢紊乱，即糖尿病，也可以通过 HbA1c 反映出来。

- **营养：** 白蛋白（Alb）主要在超常组人群中与实际年龄显著相关，因此被用作身体的营养标志物。

这些特征被视为了解受试者健康状态所需的最少特征。因此，我们将此特征集称为基线。

除基线 特征集外，我们还构建了与发病率相关 特征集，以考虑受试者的发病情况。第 4.2 节中列出的所有疾病相关特征均加入基线 特征集。具体而言，纳入了空腹血糖（FBS）、收缩压（SBP）、舒张压（DBP）、低密度脂蛋白胆固醇（LDLC）、甘油三酯（TG）和高密度脂蛋白胆固醇（HDLC）。因此，与发病率相关特征集共包含十八个特征。

最后，我们考虑除基于实际年龄或生物电阻抗测量计算的特征之外的全部生物标志物。整体 特征集共包含 88 个特征。不过，实际用于建模的特征数量会因缺失率而异。对于训练受试者未能完整测量的特征，不纳入建模；而对于部分缺失的测量值，则使用训练数据的均值进行填补。

5 结果

5.1 实施

为考虑生理性别相关的生物学差异，我们分别针对女性和男性受试者构建了模型。实验使用 PyTorch 进行。输入数据在嵌入过程中被转换为 128 维向量。编码器和解码器均采用 3 层 Performer 网络，配备 4 头注意力和 1024 维隐藏层前馈层。我们采用 AdamP优化器 [29] 训练所提出的模型。训练时批量大小设为 2,000，学习率设为 $1e-3$，共训练 1000 轮；若连续 100 轮验证数据的性能没有提升，则提前停止训练。

我们采用 *KDM*、CA 聚类法（*CAC*） [7], 和 *DNN* [8] 方法进行比较。我们根据各自的原始论文实现了基线模型。DNN 的超参数，例如隐藏层和隐藏节点的数量、学习率以及训练轮次，均设为与 [8]中最佳模型相同。按照原始论文，我们从整体 特征集中选取了与实际年龄相关的特征用于 KDM。简而言之，对于 KDM 方法，我们从整体 特征集中剔除了肿瘤标志物、视力测量、分类变量以及其他计算得出的特征。我们还使用标准整体 特征集评估了 KDM；与专门的整体 特征集相比，该评估结果相对不稳定。

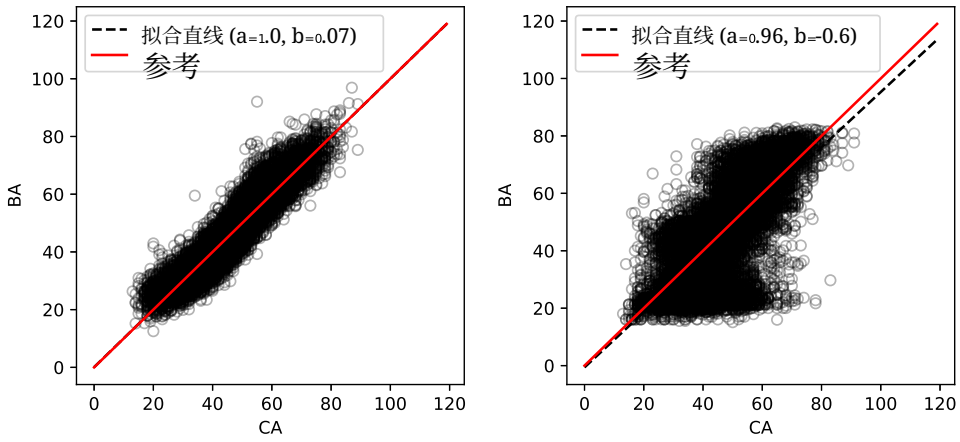
5.2 疾病状态的区分能力

表4描述了受试者总体健康状态不同的情况下，生物年龄与实际年龄之间的差距；按疾病类型划分的结果见附录A。差距为负表示，受试者比平均而言实际年龄相同的人更健康；差值为正则意味着相反的情况。如果生物年龄估算准确，健康受试者的差值应为负，而平均受试者的差值可能更接近于零。更重要的是，合理的生物年龄应能凸显受试者的疾病状态，这可以用正差值表示。

表4中的结果表明，所提出的模型能够稳健地估计受试者的健康状态。仅当模型使用正常 和超常组进行训练时，才会观察到特殊情况

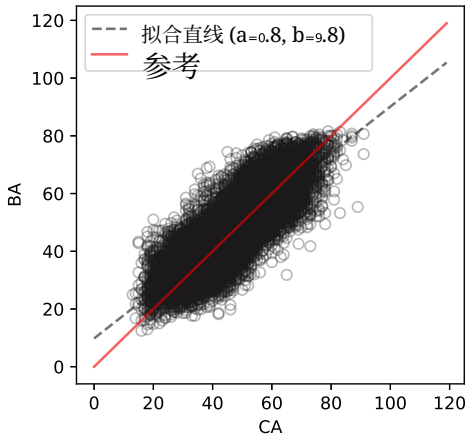
表4：男性（M）和女性（F）受试者总体疾病状态的平均 BA 间隔值。健康（Hlty）人群指没有任何疾病或疾病风险的受试者，而平均（Avg）人群包括疾病高风险组中的受试者，但不包括已确诊患有疾病的受试者（Unhlty）。灰色背景的数字表示相应测试结果所对应的数据分布与训练数据的分布不同（分布外案例）；粗体数字标示满足以下条件的情况：1）健康受试者的差距为负；2）平均受试者的绝对差距较小（即 < 1）；3）不健康受试者的差距为正。满足受试者组间相对条件（健康 <平均 <不健康）的情况以<u>下划线</u>标出。

人群		模型	整体			与发病率相关			*base*		
			Hlty	Avg	不健康	Hlty	Avg	不健康	Hlty	Avg	不健康
M	*Whole*	KDM	3.65	-0.94	0.33	-0.80	-5.55	2.35	5.59	-2.47	0.80
		CAC	-1.62	-2.06	-3.45	-10.74	-5.39	0.12	-0.04	-2.85	-1.66
		DNN	0.02	0.60	0.19	1.19	0.56	-0.10	3.31	1.08	-0.57
		所提出的方法	**-3.73**	**-0.81**	**0.47**	-4.21	-1.54	0.29	**-2.23**	**-0.75**	**0.49**
	平均	KDM	2.72	-0.29	0.41	-4.73	-3.38	2.63	0.40	-0.84	0.34
		CAC	-3.44	-1.94	-1.71	-12.85	-4.60	0.85	-6.35	-3.71	-1.83
		DNN	-0.24	0.05	-0.63	1.62	1.05	-0.15	2.68	0.79	-0.85
		所提出的方法	-4.28	-1.20	1.27	-5.32	-1.93	0.82	**-2.51**	**-0.79**	**0.93**
	正常	KDM	**-2.90**	**0.04**	**1.08**	**-3.93**	**-0.16**	**5.91**	2.37	0.03	0.32
		CAC	**-3.60**	**-0.07**	**0.65**	**-9.21**	**-0.08**	**4.56**	-1.94	-0.89	-0.22
		DNN	-0.24	-0.21	-1.89	1.57	0.14	-2.20	3.73	1.71	-1.31
		所提出的方法	**-2.13**	**-0.06**	**1.34**	-0.71	1.70	4.10	-0.67	-0.16	-0.03
	超常组	KDM	**-0.81**	**0.22**	**0.36**	-0.10	4.46	9.79	0.85	-2.65	-1.58
		CAC	-0.96	-3.73	-8.28	1.64	-1.35	-5.69	0.80	-1.75	-5.97
		DNN	-0.73	-0.17	-6.31	0.26	-4.95	-16.86	-0.44	-2.25	-6.02
		所提出的方法	0.54	-0.05	0.07	1.87	3.45	5.89	0.59	2.21	4.34
F	*Whole*	KDM	**-1.31**	**-0.59**	**0.54**	-22.68	-10.00	8.51	-22.25	-10.44	9.50
		CAC	-4.15	-4.79	-4.30	-7.78	-5.04	0.19	-3.14	-3.58	-1.28
		DNN	0.31	0.40	-0.37	0.52	0.57	-0.02	1.76	0.49	-1.06
		所提出的方法	**-3.29**	**-0.99**	**1.19**	-3.11	-1.11	0.69	**-0.59**	**-0.31**	**0.66**
	平均	KDM	**-2.05**	**-0.57**	**1.67**	-9.43	-2.51	7.76	-7.62	-2.20	6.68
		CAC	-6.03	-3.74	1.23	-7.98	-3.22	2.42	-4.67	-2.80	0.37
		DNN	0.85	1.09	0.08	0.08	0.01	-1.39	3.22	1.91	-1.74
		所提出的方法	**-3.47**	**-0.79**	**2.17**	**-4.04**	**-0.62**	**2.81**	**-1.70**	**-0.54**	**2.64**
	正常	KDM	**-8.60**	**0.17**	**10.62**	**-12.29**	**-0.21**	**15.75**	**-9.70**	**-0.01**	**13.17**
		CAC	-6.01	-1.75	4.05	-9.80	-1.44	5.57	-4.47	-2.24	0.08
		DNN	1.44	1.20	-1.78	-0.11	0.60	-2.62	1.31	0.44	-4.47
		所提出的方法	**-1.56**	**0.26**	**3.27**	0.31	0.22	0.28	**-0.66**	**0.19**	**0.91**
	超常组	KDM	-0.16	6.40	11.30	-0.79	12.84	28.03	-0.24	7.97	18.83
		CAC	-0.28	1.48	-1.77	1.10	5.64	0.95	1.17	1.76	-2.25
		DNN	-1.82	-3.80	-11.19	-0.33	-2.95	-13.09	-0.10	-4.04	-13.27
		所提出的方法	0.05	1.19	1.89	2.08	6.02	8.20	**-0.22**	**0.03**	**0.69**

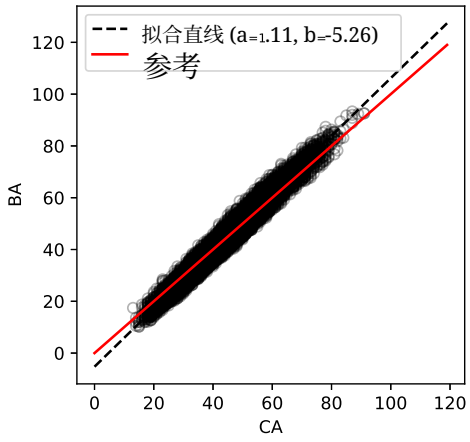


(a) KDM

(b) CAC

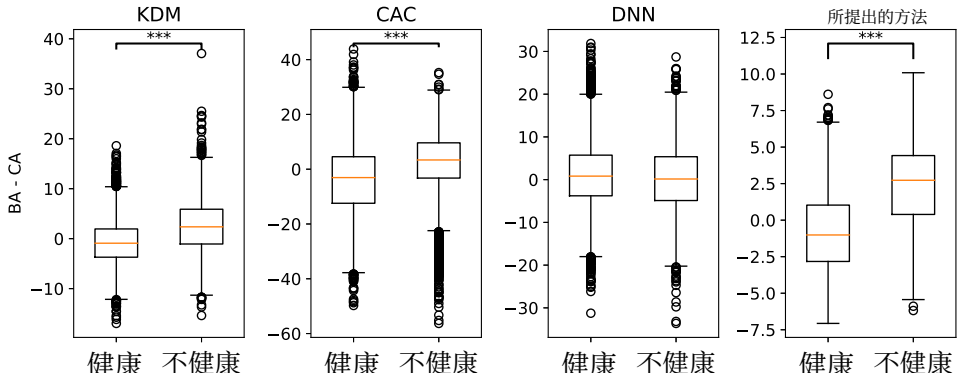


(c) 深度神经网络

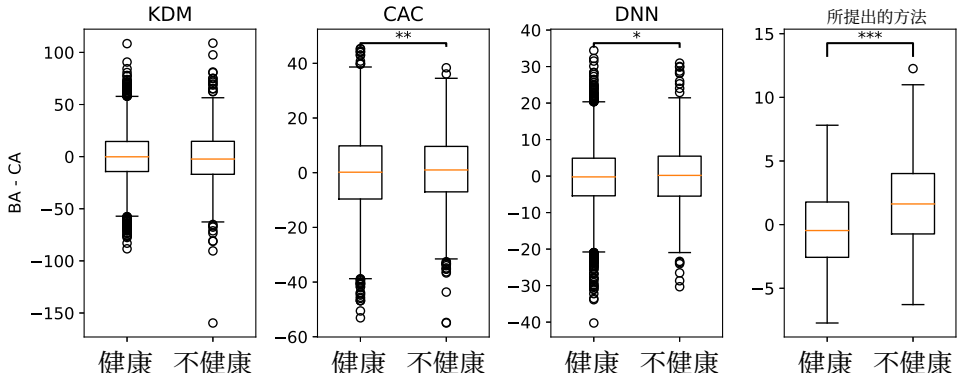


(d) 所提出的方法

图3：基线模型和所提出的模型相对于实际年龄（CA）的估计生物年龄（BA）。此处展示了女性-平均-整体 病例的全部三次运行的测试结果。黑色虚线拟合实际年龄与所得生物年龄，表示为 $BA = a \times CA + b$；红线对应$BA = CA$病例。



(a) 女性-平均-整体



(b) 男性-正常-整体

图4：根据受试者的健康状态（健康与不健康）绘制的生理年龄与实际年龄之间差距的箱线图。使用两个独立样本的 t 检验评估差距差异的统计显著性（*： p-值 < 0.05，**： p-值 < 0.01，***： p-值 < 0.001）。

.....人群中，由于训练人群本身的局限性，模型无法探索发病信息。相比之下，基线模型在大多数情况下无法反映健康状态之间的差异。KDM 和 CAC 似乎在少数情况下能够奏效，例如训练人群为女性且平均，并且使用整体 特征集时。然而，正如图3和图4所示，即使对于分布内测试数据，基线模型得出的生物年龄也存在过大的变异性。尤其是，DNN 由于严重依赖实际年龄，未能捕捉受试者的发病信息。此外，DNN 不仅会为缺乏说服力的分布外数据得出较小的差值，还倾向于低估发病程度。例如，在女性-平均-整体情况下，DNN 得出的不健康受试者的平均差距值为 0.08，而健康受试者和平均受试者的平均差距值分别为 0.85 和 1.09。

从图3中还能观察到一个有趣的现象：所提出的模型倾向于将老年人群的生物年龄估计得略高于年轻一代，即 $a > 1$。相比之下，CAC 和 DNN 的斜率值较小（分别为 0.96 和 0.80）；考虑到年轻人群受益于改善的营养和医疗保健服务，这并不合乎逻辑。一种可能的解释是，CAC 和 DNN 在训练时仅预测受试者的实际年龄，而未考虑实际年龄分布不平衡，因此模型倾向于将生物年龄预测为接近实际年龄均值（图3训练数据中的均值为 45.6），以实现学习目标。

比较所提出的模型在不同训练人群下获得的结果可以发现，使用*Whole*和平均人群进行训练的模型具有更强的区分受试者健康状态的能力[1]，，这也体现为表4中差值之间更明显的差异。当使用正常和超常组人群训练模型时，由于这些人群组中不包含患病受试者，发病对比损失无法采用，因此这一结果是预期之中的。此外，还可以观察到，对于未采用发病情况指导的CAC，使用更大规模的训练人群时，其区分能力有所提升。因此，正如我们在第1节中所揭示的，训练人群的多样性对于获得可靠的生物年龄至关重要。

对于男性-*Whole*-整体情况，分别根据受试者患糖尿病、高血压和癌症的发病情况，在图5、图6和图7中对获得的生物年龄进行了评估。需要注意的是，关于癌症的信息（包括附录B中的心血管疾病和脑血管意外）在训练期间并未提供给所提出的模型，但在此处用于展示该模型能够考虑总体健康状态。无论疾病类型如何，所提出的模型始终表现出对受试者发病情况显著的区分能力。基线模型无法区分不同的疾病状态，这一点从其较大的差值方差中可以明显看出。我们还可以从图5、图6和图7中发现CAC和DNN生物年龄估计不稳定的其他证据，即随着实际年龄增加，其差值明显减小。女性受试者以及其他疾病（DLP、多发性硬化症、心血管疾病和脑血管意外）的结果见附录B。

5.3 与死亡率的关系

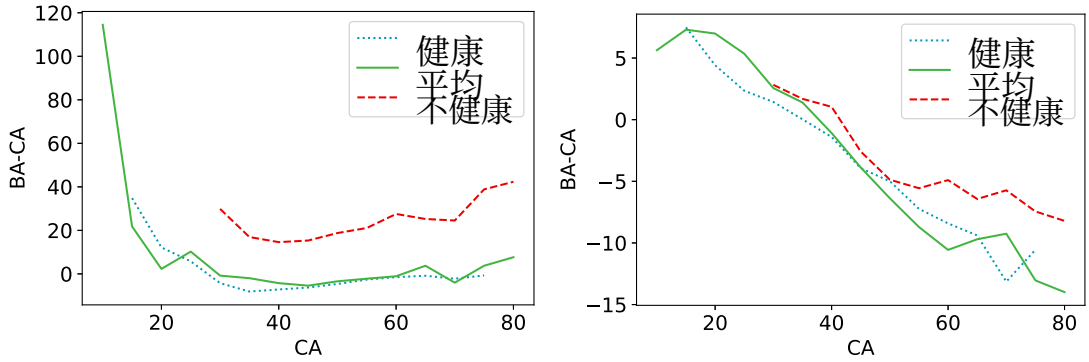
预计生物年龄具有预测受试者死亡率的能力；也就是说，生物年龄越大，死亡时间应越短。我们在表 5 中比较了使用实际年龄和估算生物年龄进行线性回归所得结果与死亡时间之间的关系。通过两种方式考察所提出的模型得到的生物年龄，即使用生物年龄进行单变量线性回归

[1]例如，尽管健康、平均和不健康受试者之间的关系经过了良好校准，但在男性-超常组-整体、男性-超常组-与发病率相关、男性-超常组-基线、女性-正常-与发病率相关、女性-超常组-整体以及女性-超常组-与发病率相关 的情况下，健康受试者的平均差距值为正。

表 5：基于《年龄》和生物年龄对整体人群的死亡时间进行线性回归的结果。通过线性回归线的斜率、回归结果与实际死亡时间之间的决定系数 (R^2)，以及年龄与死亡时间之间的皮尔逊相关系数（PCC），评估其与死亡率的关系。粗体斜率值表示统计显著性最高的情况，即所得斜率值显著小于零。粗体数字对应的 R^2 和 PCC 结果为最佳情况。所提出的方法（差值）模型的斜率和 PCC 使用差值计算。

人群	特征集	模型	斜率		R^2		PCC	
			CA	BA	CA	BA	CA	BA
M	整体	KDM（生物年龄）	-4.683	**-6.092*****	0.001	0.048	-0.032	-0.219
		CAC（生物年龄）		**-13.487*****		0.030		-0.172
		深度神经网络（生物年龄）		-3.732		0.001		-0.024
		所提出的方法（生物年龄）		**-19.938*****		0.020		-0.141
	与发病率相关	KDM	-1.601	**-5.922*****	0.000	0.017	-0.012	-0.130
		CAC		-8.886**		0.009		-0.093
		DNN		-14.674**		0.009		-0.097
		所提出的方法（生物年龄）		**-13.812*****		0.011		-0.105
		所提出的方法（差距）		**-139.756*****		**0.101**		**-0.317**
	base	KDM	6.959	**-6.498*****	0.002	0.018	0.048	-0.133
		CAC		-2.863		0.001		-0.031
		DNN		-6.953		0.002		-0.046
		所提出的方法（生物年龄）		-5.946		0.002		-0.045
		所提出的方法（差距）		**-112.617*****		**0.080**		**-0.273**
F	整体	KDM	3.794	4.867	0.001	0.001	0.029	0.037
		CAC		-3.779		0.002		-0.046
		DNN		2.654		0.000		0.017
		所提出的方法（生物年龄）		-3.823		0.001		-0.032
		所提出的方法（差距）		**-167.651*****		**0.086**		**-0.274**
	与发病率相关	KDM	13.708	0.789	0.011	0.003	0.106	0.058
		CAC		6.991		0.005		0.071
		DNN		23.668		0.026		0.160
		所提出的方法（生物年龄）		6.530		0.003		0.054
		所提出的方法（差距）		**-152.662*****		**0.074**		**-0.221**
	base	KDM	11.272	4.196	0.007	0.028	0.082	0.167
		CAC		11.086		0.015		0.122
		DNN		23.056		0.023		0.153
		所提出的方法（生物年龄）		2.601		0.000		0.019
		所提出的方法（差距）		**-95.941*****		**0.049**		**-0.210**

*： p-值< 0.05, **： p-值< 0.01, ***： p-值< 0.001



(a) KDM (b) CAC

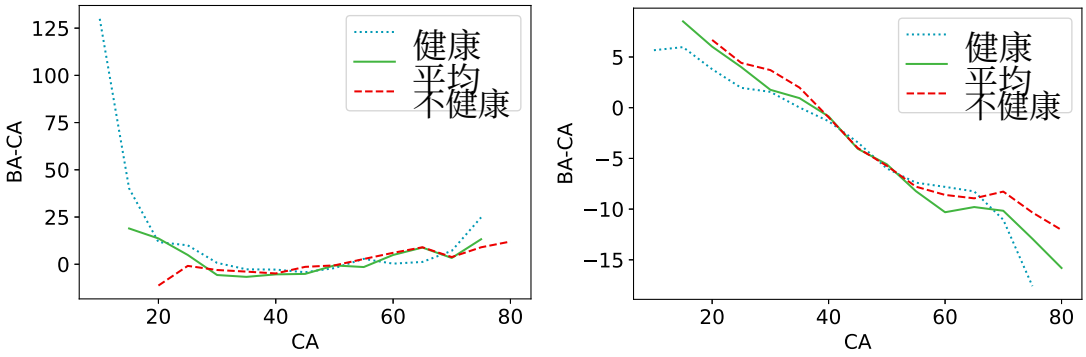
(c) 深度神经网络 (d) 所提出的方法

图5：男性—Whole—整体病例中，生物年龄与实际年龄之间的差距随**糖尿病** 发病情况的变化。差值按5年区间取平均，即实际年龄在40至44岁范围内的受试者的平均差距值显示在 $CA = 40$处。绿色点线表示健康受试者的结果，红色虚线表示不健康受试者的结果。平均受试者的结果以黄色实线表示。彩色区域覆盖第0.25至第0.75分位数。

（= 癌症 +差距），以及使用癌症和差距进行的多元线性回归。所提出的模型，尤其是在使用差值进行评估时，在所有情况下均表现最佳；而基线模型采用多元策略后仅获得边际改善（结果见附录C）。在基线方法中，已知对死亡率预测有效的KDM总体上优于其他方法。这表明，与直接对生物年龄建模相比，基于年龄差的方法更为有效。

与此前发病率分析的结果不同，在三种不同特征集（整体、与发病率相关和基线）中，没有一种表现出明显优势；使用整体 和与发病率相关 特征集训练的模型，分别在预测男性和女性受试者的死亡率方面表现出更强的能力。这可能表明，尽管额外信息较为稀疏，使用这些信息仍能带来有意义的提升。

为进一步验证所得生物年龄与死亡率之间的关系，我们以男性-*Whole*-整体 案例为基础进行生存分析。由于我们采用了三次重复运行，并且每次使用不同的



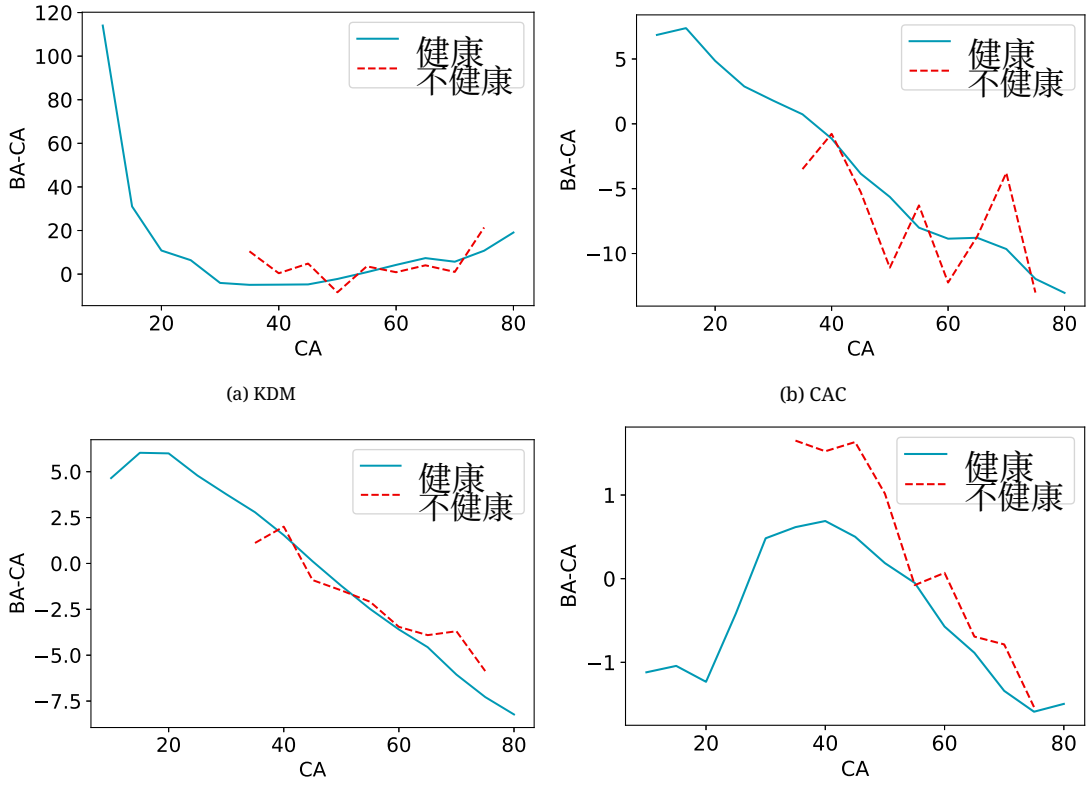
(a) KDM

(b) CAC

(c) 深度神经网络

(d) 所提出的方法

图 6：在男性-**高血压** 患病的*Whole*-整体 病例中，生物学年龄与实际年龄之间的差距。



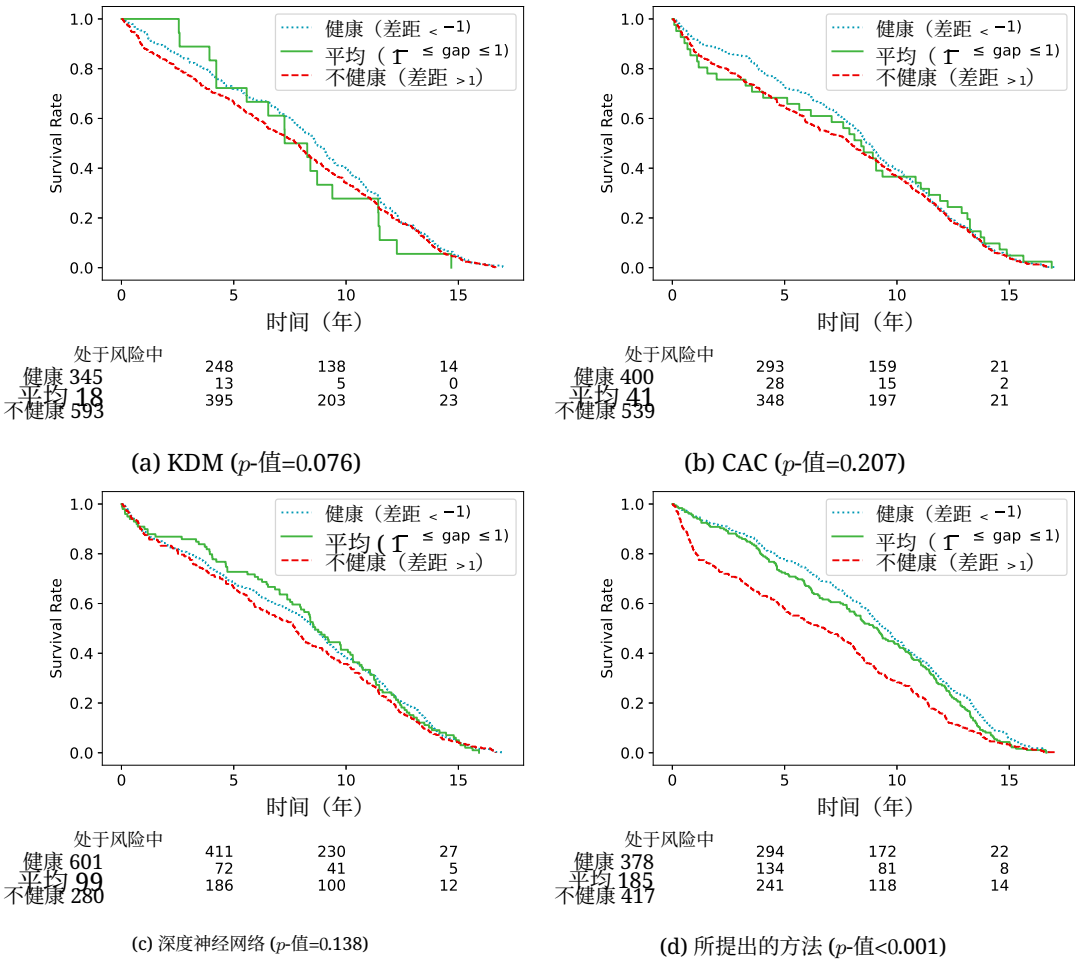
(a) KDM

(b) CAC

(c) 深度神经网络

(d) 所提出的方法

图 7：男性-*Whole*-整体 病例中，生物学年龄与实际年龄之间的差距随**癌症** 发病率的变化。绿色实线表示健康受试者的结果，红色虚线表示不健康受试者的结果。



(a) KDM (p-值=0.076)

(b) CAC (p-值=0.207)

(c) 深度神经网络 (p-值=0.138)

(d) 所提出的方法 (p-值<0.001)

图8：男性-*Whole*-整体病例中使用生理年龄与实际年龄之间的差距进行生存分析的结果。括号中的 p值表示通过对数秩检验得出的健康受试者与不健康受试者Kaplan-Meier曲线之间差异的显著性。

单次实验设置的随机种子下，从三次运行中提取包含死亡信息的测试数据进行分析，共计956个样本（分别为334、306和316个样本）。图8展示了根据生物年龄与实际年龄之间的差值划分的健康、平均和不健康受试者组的Kaplan-Meier曲线。具体而言，差值小于-1（大于1）的受试者分别被视为健康（不健康），其余受试者则归入平均组。我们还使用对数秩检验评估健康与不健康受试者的Kaplan-Meier曲线差异。

分析表明，所提出的模型能够有效区分未来死亡情况不同的受试者。值得注意的是，不健康受试者的生存率在早期阶段急剧下降，这表明我们的生物年龄模型在死亡率短期预测方面甚至更具优势。相比之下，基线模型得到的生物年龄无法区分死亡预期不同的受试者。

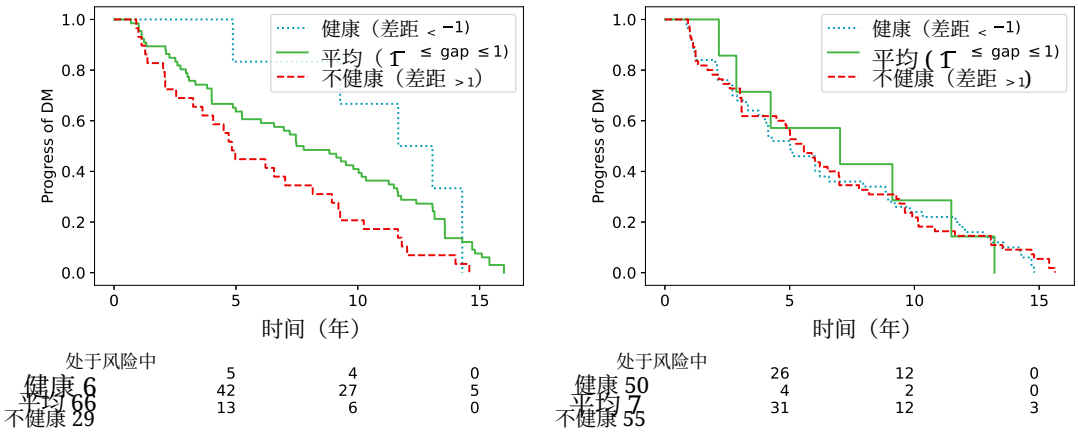
同样地，图9根据预测的生物年龄评估了已知与生物年龄 [30, 31],密切相关的糖尿病发病情况。尽管依据KDM生物年龄划分的健康组和不健康组在糖尿病发病模式上也存在统计学显著差异，但通过所提出的模型估算的生物年龄显示出其作为糖尿病发病中介指标的潜力。

6 结论

总之，所提出的模型在考虑受试者群体发病率和死亡率的同时，有效地学习了生物衰老的潜在表示。我们证明，即使某种疾病未纳入模型训练，所提出的模型估算的生物年龄仍能成功反映多种疾病的患病状态。所得生物年龄还能够准确预测衰老相关结局，即死亡率和糖尿病发病，其表现优于该领域现有方法。本研究有望增进我们对衰老过程的理解，并助力开发改善老年人健康结局的干预措施。未来仍需针对衰老的不同方面开展进一步实验和验证，以增强模型的稳健性。

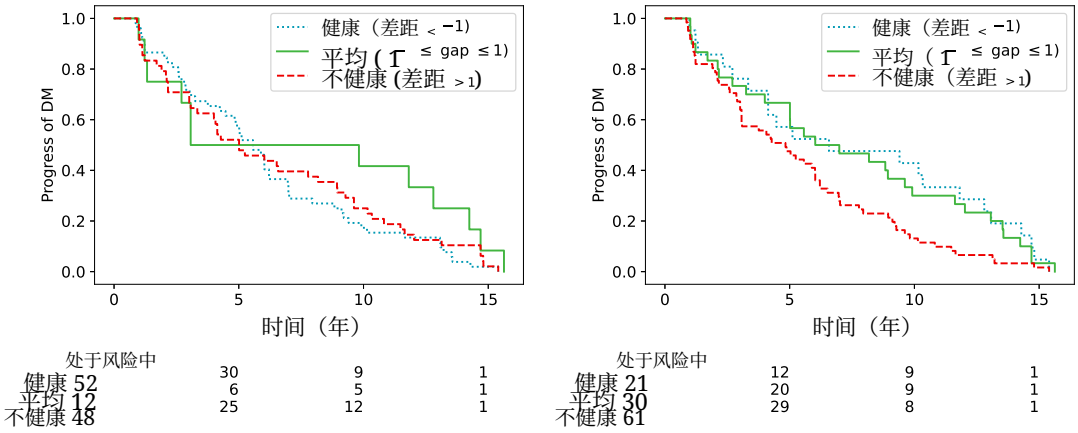
参考文献

[1] Juulia Jylhava、Nancy L. Pedersen 和 Sara Hagg。生物年龄预测因子。*EBioMedicine*，21：29–36，2017 年。[2] Linpei Jia、Weiguang Zhang 和 Xiangmei Chen。生物年龄估计的常用方法。*Clinical Interventions in Aging*，12：759–772，2017 年。[3] Steve Horvath 和 Kenneth Raj。基于DNA甲基化的生物标志物与衰老的表观遗传时钟理论。*Nature Reviews Genetics*，19：371–384，2018 年。[4] Alexander Vaiserman 和 Dmytro Krasnienkov。端粒长度作为生物年龄标志物：最新进展、未决问题与未来展望。*Frontiers inGenetics*，11：630186:1–20，2021 年。[5] Jarod Rutledge、Hamilton Oh 和 Wyss-Coray Tony。利用组学数据测量生物年龄。*Nature Reviews Genetics*，23：715–727，2022 年。



(a) KDM (p-值=0.048)

(b) CAC (p-值=0.687)



(c) 深度神经网络 (p-值=0.603)

(d) 所提出的方法 (p-值=0.032)

图 9：针对女性-*Whole*-整体 病例，使用生物年龄与实际年龄之间的差距进行糖尿病发病分析的结果。括号中的 p值表示通过对数秩检验得到的健康受试者与不健康受试者Kaplan-Meier曲线差异的显著性。

[6] Morgan E Levine. 衰老速率建模：估算的生物年龄能否比实际年龄更准确地预测死亡率？ 老年学期刊 *A* 系列：生物医学与医学科学, 68(6):667–674, 2013.[7] Syed Ahiqur Rahman and Donald A. Adjeroh. 年龄邻域的质心：一种估算生物年龄的新方法。 *IEEE* 生物医学与健康信息学期刊, 24(4):1226–12234, 2019.[8] Chul-Young Bae, Yoori Im, Jonghoon Lee, Choong-Shik Park, Miyoung Kim, Hojeong Kwon, Boseon Kim, Hye ri Park, Chun-Koo Lee, Inhee Kim, and JeongHoon Kim. 使用临床实践环境中常规测量的临床生物标志物比较生物年龄预测模型：人工智能技术与传统统计方法。 《分析科学前沿》, 1:709589:1–12, 2021.[9] Timothy V Pyrkov, Konstantin Slipensky, Mikhail Barg, Alexey Kondrashin, Boris Zhurov, Alexander Zenin, Mikhail Pyatnitskiy, Leonid Menshikov, Sergei Markov, and Peter O Fedichev. 通过深度学习从生物医学数据中提取生物年龄：好事是否过犹不及？ 《科学报告》, 8 (1):1–11, 2018.[10] James William Hollingsworth, Asaji Hashizume, and Seymour Jablon. 广岛受试者衰老测试之间的相关性——尝试定义“生理年龄”。 《耶鲁生物学与医学杂志》, 38(1):11, 1965.[11] TL Dubina, VA Dyundikova, and EV Zhuk. 生物年龄及其估算。II. 采用多元回归分析评估白化大鼠的生物年龄。 《实验老年学》, 18(1):5–18, 1983.[12] E Nakamura, K Miyao, and T Ozeki. 采用主成分分析评估生物年龄。《衰老与发育机制》, 46(1-3):1–18, 1988.[13] E Nakamura, T Moritani, and A Kanetaka. 生物年龄与体能年龄。《欧洲应用生理学与职业生理学杂志》 , 58(7):778–785, 1989.[14] JinHo Park, BeLong Cho, HyukTae Kwon, and CheolMin Lee. 利用主成分分析和衰老临床生物标志物，为韩国男性开发生物年龄评估方程。《老年学与老年医学档案》, 49(1):7–12, 2009.[15] Eitaro Nakamura and Kenji Miyao. 一种识别衰老生物标志物并构建人类生物年龄指数的方法。 *The Journals of Gerontology Series A:Biological Sciences andMedical Sciences*, 62(10):1096–1105, 2007.[16] Wei-Guang Zhang, Shu-Ying Zhu, Xiao-Juan Bai, De-Long Zhao, Shi-Min Jiang, Juan Li, Zuo-Xiang Li, Bo Fu, Guang-Yan Cai, Xue-Feng Sun, et al. 基于端粒长度和实际年龄筛选衰老生物标志物，以构建生物年龄方程。 《年龄》, 36(3):1201–1211, 2014.[17] Eitaro Nakamura and Kenji Miyao. 人类生物衰老中的性别差异。 *The Journals ofGerontology Series A: BiologicalSciences and Medical Sciences*, 63(9):936–944, 2008.[18] Petr Klemera and Stanislav Doubal. 关于生物年龄概念及其计算的新方法。 《衰老与发育机制》, 127(3):240–248, 2006.[19] Il Haeng Cho, Kyung S Park, and Chang Joo Lim. 一项关于生物年龄估计算法的实证比较研究，并应用于工作能力指数（wai）。 《衰老与发育机制》, 131(2):69–78, 2010.

[20] Suhyeon Kim, Hangyeol Kim, Eun-Sol Lee, Chiehyeon Lim, and Junghye Lee. 用于生物年龄估计的嵌入风险评分的深度学习：开发与验证。 *Information Sciences*, 586: 628–643, 2022.[21] Evgeny Putin, Polina Mamoshina, Alexander Aliper, Mikhail Korzinkin, Alexey Moskalev, Alexey Kolosov, Alexander Ostrovskiy, Charles Cantor, Jan Vijg, and Alex Zhavoronkov. 人类衰老的深度生物标志物：将深度神经网络应用于生物标志物开发。 *Aging (AlbanyNY)*, 8(5):1021, 2016.[22] Edardo H. P. Pooch, Pedro L. Ballester, and Rodrigo C. Barros. 我们能否信任基于深度学习的诊断? 领域偏移对胸部X线片分类的影响。载于*International Workshop on Thoracic Image Analysis* 论文集, pages 74–83, 2020.[23] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. 面向分布外泛化的深度稳定学习。载于计算机视觉与模式识别会议论文集, pages 5372–5382, 2021.[24] David Berend, Xiaofei Xie, Lei Ma, Lingjun Zhou, Yang Liu, Chi Xu, and Jianjun Zhao. 猫不是鱼：深度学习测试需要具备分布外意识。载于国际自动化软件工程会议论文集, pages 1041–1052, 2020.[25] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. 重新思考 Performers 中的注意力机制。载于国际学习表征会议论文集, 2021.[26] Diederic P. Kingma and Max Welling. 自编码变分贝叶斯。载于*arXiv* 预印本, pages arXiv:1312.6114:1–14, 2014.[27] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander Smola. 一种核两样本检验。机器学习研究期刊, 13:723–773, 2012.[28] Seokyung An, Choonghyun Ahn, Sungji Moon, En Ji Sim, and Sue-Kyung Park. 个体化生物年龄作为疾病预测因子：韩国基因组与流行病学研究（KoGES）队列。个性化医学杂志, 12(3):505:1–11, 2022.[29] Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdu Yoon, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. AdamP：为尺度不变权重上的动量优化器减缓“减速”。载于国际学习表征会议论文集, 2021.[30] Crystal D. Grant, Nadereh Jafari, Lifang Hou, Yun Li, James D. Stewart, Guosheng Zhang, Archana Lamichhane, JoAnn E. Manson, Andrea A. Baccarelli, Eric A. Whitsel, and Karen N. Conneely. 一项关于DNA甲基化作为年龄相关糖尿病风险潜在中介因素的纵向研究。*GeroScience*, 39(5):475–489, 2017.[31] Nadine Bahour, Briana Cortez, Hui Pan, Hetal Shah, Alessandro Doria, and Cristina Aguayo- Mazzucato. 糖尿病与临床生物标志物所指示的生物年龄增加相关。*GeroScience*, 44(1):415–427, 2022.

A 按发病率分类的平均差距值

表 A1：患有糖尿病、高血压和 DLP 的男性受试者中，健康（Hlty）、患病前（preD）和患病（withD）组的平均 BA 间隔值。

人群	模型	疾病	整体			与发病率相关			*base*		
			健康糖尿病前期		withD	健康糖尿病前期		withD	健康糖尿病前期		withD
Whole	KDM	DM		0.64	16.0		-3.48	20.4		-1.47	11.43
		HBP	6.14	-1.10	0.18	-0.80	-4.16	4.82	5.59	-4.58	-1.45
		DLP		-0.96	-3.78		-5.94	-0.06		-2.85	-1.67
	CAC	DM		-2.62	-3.87		-4.90	1.95		-3.00	-1.31
		HBP	-1.62	-1.72	-4.63	-10.74	-1.03	1.41	-0.04	-3.45	-3.13
		DLP		-2.16	-4.04		-5.24	-0.58		-3.01	-2.86
	DNN	DM		0.77	0.23		0.88	-0.18		0.54	-1.40
		HBP	0.02	0.91	-0.26	1.19	0.77	-1.06	3.31	0.70	-2.49
		DLP		0.59	0.14		0.48	-0.26		0.98	-0.45
	所提出的方法	DM		**-0.47**	**1.26**		**-0.89**	**1.69**		**0.39**	**4.35**
		HBP	**-3.73**	**-0.09**	**0.98**	**-4.21**	**-0.79**	**0.92**	**-2.23**	**-0.65**	**0.69**
		DLP		**-0.64**	**1.47**		-1.41	1.01		**-0.71**	**0.80**
平均	KDM	DM		0.93	11.41		**-0.15**	**18.75**		1.19	11.97
		HBP	2.72	0.46	1.36	**-4.73**	**-0.31**	**5.06**	0.40	-1.81	0.18
		DLP		-0.47	-0.49		-3.44	1.74		-0.93	0.03
	CAC	DM		-2.53	-1.78		-3.45	3.29		-2.59	2.36
		HBP	-3.44	-0.12	-1.41	-12.85	1.35	3.36	-6.35	-3.06	-2.40
		DLP		-1.88	-1.62		-4.37	0.24		-3.64	-1.49
	DNN	DM		0.09	-0.44		1.12	-0.10		0.48	-0.62
		HBP	-0.24	0.24	-1.06	1.62	1.03	-0.74	2.68	0.41	-2.19
		DLP		0.05	-0.58		1.02	0.08		0.74	-0.26
	所提出的方法	DM		**-0.77**	**0.94**		-1.35	0.53		**0.42**	**3.85**
		HBP	**-4.28**	**-0.47**	**1.04**	-5.32	-1.27	0.64	**-2.51**	**-0.64**	**1.13**
		DLP		-1.07	1.80		-1.79	1.35		**-0.74**	**0.90**
正常	KDM	DM		1.77	7.37		3.68	20.85		2.18	13.91
		HBP	**-2.90**	**0.12**	**0.69**	**-3.93**	1.15	6.74	2.37	-1.52	0.26
		DLP		**0.14**	**0.86**		**0.01**	**5.42**		-0.13	-0.29
	CAC	DM		**0.76**	**5.15**		2.50	10.8		1.49	9.62
		HBP	**-3.60**	1.26	0.67	**-9.21**	4.06	6.21	-1.94	-1.24	-0.88
		DLP		**0.06**	**0.45**		**0.21**	**3.67**		-0.92	-0.21
	DNN	DM		-0.42	-2.77		0.19	-3.79		1.10	-2.30
		HBP	-0.24	-0.42	-2.80	1.57	-0.08	-3.28	3.73	1.26	-2.96
		DLP		-0.20	-1.63		0.11	-1.83		1.63	-0.65
	所提出的方法	DM		**0.37**	**1.97**		2.90	8.59		**0.14**	**0.81**
		HBP	**-2.13**	**0.07**	**1.51**	-0.71	1.91	5.43	**-0.67**	**-0.34**	**0.44**
		DLP		**0.02**	**1.46**		1.80	3.59		-0.14	-0.10

（续）患有糖尿病、高血压和 DLP 的男性受试者中，健康（Hlty）、患病前（preD）和患病（withD）人群的平均 BA 间隔值。

人群	模型	疾病	整体			与发病率相关			*base*		
			健康糖尿病前期		withD	健康糖尿病前期		withD	健康糖尿病前期		withD
超级正常	KDM	DM	**-0.81**	**0.88**	**1.87**	-0.10	7.53	17.67	0.85	0.33	12.15
		HBP		0.47	-0.39		6.25	10.93		-3.78	-1.93
		DLP		**0.21**	**0.39**		4.47	9.30		-2.72	-2.22
	CAC	DM	-0.96	-5.72	-13.32	1.64	-3.36	-11.97	0.80	-2.44	-10.6
		HBP		-4.29	-10.75		-1.76	-8.25		-2.46	-8.26
		DLP		-3.76	-7.28		-1.38	-4.49		-1.79	-5.11
	DNN	DM	-0.73	-1.73	-13.86	0.26	-9.60	-30.88	-0.44	-2.50	-5.34
		HBP		-1.24	-8.60		-8.28	-21.58		-2.87	-8.60
		DLP		-0.14	-5.83		-4.95	-15.76		-2.28	-5.20
	所提出的方法	DM	0.54	0.04	1.51	1.87	4.56	7.90	0.59	3.16	6.59
		HBP		-0.42	-0.19		3.63	6.80		2.40	4.99
		DLP		-0.06	0.24		3.47	5.55		2.23	4.13

表A2：女性受试者中健康（Hlty）、患病前（preD）和患病（withD）组在糖尿病、高血压和DLP发病率方面的平均BA差距值。

人群	模型	疾病	整体			与发病率相关			*base*		
			健康糖尿病前期		withD	健康糖尿病前期		withD	健康糖尿病前期		withD
Whole	KDM	DM	**-1.80**	**-0.25**	**2.92**	-22.68	-3.71	45.89	-22.25	1.24	61.12
		HBP		**0.20**	**1.60**		4.22	24.52		-4.96	21.13
		DLP		**-0.60**	**1.30**		-7.78	17.29		-8.63	17.29
	CAC	DM	-4.15	-5.46	-3.42	-7.78	-4.41	4.29	-3.14	-3.38	1.04
		HBP		-4.97	-4.21		2.73	4.12		-4.69	-2.07
		DLP		-4.90	-5.78		-4.64	-0.03		-3.62	-3.04
	DNN	DM	0.31	0.43	-0.60	0.52	1.16	-0.06	1.76	0.78	-1.12
		HBP		0.45	-1.33		0.86	-1.37		-1.36	-3.69
		DLP		0.39	-0.77		0.53	-0.59		0.21	-2.32
	所提出的方法	DM	**-3.29**	**-0.20**	**3.39**	**-3.11**	**-0.05**	**2.98**	**-0.59**	**0.83**	**3.29**
		HBP		1.03	2.80		**0.80**	**2.29**		**-0.72**	**0.27**
		DLP		**-0.54**	**2.49**		**-0.73**	**2.17**		**-0.35**	**0.13**
平均	KDM	DM	**-2.05**	**0.34**	**3.80**	-9.43	1.60	22.29	**-7.62**	3.30	23.78
		HBP		1.01	2.26		6.22	11.39		**-0.60**	**8.83**
		DLP		**-0.28**	**1.93**		-1.26	8.50		-1.20	7.55
	CAC	DM	-6.03	-2.93	3.36	-7.98	-1.87	5.49	**-4.67**	**-0.50**	**5.09**
		HBP		1.97	3.63		5.58	5.09		-2.54	0.55
		DLP		-3.18	1.46		-2.22	2.11		-2.44	0.88
	DNN	DM	0.85	1.11	0.09	0.08	0.37	-0.41	3.22	1.54	-0.50
		HBP		1.44	-0.33		0.51	-1.97		-0.20	-3.03
		DLP		1.11	0.04		-0.07	-1.54		1.65	-1.75
	所提出的方法	DM	**-3.47**	**0.21**	**3.08**	**-4.04**	**0.69**	**3.60**	**-1.70**	1.88	10.64
		HBP		1.06	2.37		1.49	2.99		**0.19**	**3.65**
		DLP		**-0.19**	**2.71**		**0.12**	**3.37**		**-0.42**	**2.83**

（续）女性受试者中健康（Hlty）、患病前（preD）和患病（withD）群体在糖尿病、HBP 和 DLP 发病率方面的平均 BA 间隔值。

人群	模型	疾病	整体		与发病率相关		*base*	
			健康糖尿病前期	withD	健康糖尿病前期	withD	健康糖尿病前期	withD
正常	KDM	DM	5.10	19.34	7.11	34.95	9.34	40.16
		HBP	-8.60	7.42	11.95	-12.29	11.28	20.06	-9.70	4.43	17.09
		DLP	1.80	12.32	1.80	17.19	1.46	14.09
	CAC	DM	**0.38**	**5.69**	1.00	8.06	**0.91**	**5.72**
		HBP	**-6.01**	5.97	5.71	**-9.80**	10.76	7.93	**-4.47**	-1.90	0.40
		DLP	**-0.98**	**4.36**	**0.07**	**5.62**	-1.92	0.30
	DNN	DM	1.02	-3.31	1.11	-3.19	0.73	-4.39
		HBP	1.44	1.04	-2.98	-0.11	0.78	-4.03	1.31	-1.43	-5.69
		DLP	1.15	-1.76	0.66	-2.32	0.19	-4.68
	所提出的方法	DM	1.34	4.39	0.30	-0.06	1.10	1.17
		HBP	**-1.56**	1.55	3.82	0.31	-0.47	0.39	**-0.66**	**0.39**	**1.51**
		DLP	**0.58**	**3.53**	0.24	0.38	**0.26**	**1.01**
超级正常	KDM	DM	9.34	14.63	19.71	43.96	13.74	34.35
		HBP	-0.16	8.39	10.80	-0.79	19.60	31.30	-0.24	11.25	21.83
		DLP	6.59	12.22	13.13	29.38	8.07	19.15
	CAC	DM	1.52	-5.15	6.41	-3.47	3.53	-4.27
		HBP	-0.28	2.00	-3.56	1.10	6.03	-1.33	1.17	-0.24	-4.22
		DLP	1.50	-1.88	5.58	0.80	1.63	-2.54
	DNN	DM	-5.07	-16.44	-5.00	-20.13	-7.19	-21.52
		HBP	-1.82	-5.61	-13.27	-0.33	-6.17	-16.07	-0.10	-7.33	-15.93
		DLP	-3.89	-12.06	-3.00	-13.99	-4.08	-13.72
	所提出的方法	DM	1.39	1.71	7.28	10.09	0.38	0.22
		HBP	0.05	1.13	1.58	2.08	6.16	8.26	**-0.22**	**0.27**	**1.11**
		DLP	1.28	2.21	6.25	8.79	**0.02**	**0.73**

表A3：健康组（Hlty）和患病组（withD）男性受试者在多发性硬化症、癌症、心血管疾病和脑血管意外发病率方面的AverageBA间隔值。

人群	模型	疾病	整体		与发病率相关		*base*	
			Hlty	withD	Hlty	withD	Hlty	withD
Whole	KDM	MS	6.14	-5.05	**-0.80**	**4.58**	5.59	-2.09
		癌症		2.23		**0.46**		0.48
		CVD		7.50		**2.12**		3.68
		CVA		0.29		**5.90**		-0.55
	CAC	MS	-1.62	-4.07	**-10.74**	**0.84**	-0.04	-1.99
		癌症		-7.38		-2.81		-3.94
		CVD		-4.06		-1.53		-2.01
		CVA		-6.62		-0.12		-5.30
	DNN	MS	0.02	0.33	1.19	0.05	3.31	-0.56
		癌症		-1.87		-2.78		-3.34
		CVD		-0.70		-2.27		-2.80
		CVA		-2.32		-3.79		-5.11
	所提出的方法	MS	**-3.73**	**2.07**	**-4.21**	**1.91**	**-2.23**	**1.91**
		癌症		**0.38**		**0.13**		**0.56**
		CVD		-0.32		-0.06		**1.19**
		CVA		**0.03**		**0.23**		**1.62**

（续）健康（Hlty）和患病（withD）男性受试者中，多发性硬化症、癌症、心血管疾病和脑血管意外的平均 BA 间隔值。

人群	模型	疾病	整体		与发病率相关		*base*	
			Hlty	withD	Hlty	withD	Hlty	withD
平均	KDM	MS	2.72	1.09	**-4.73**	**6.85**	0.40	0.73
		癌症		1.66		**1.70**		1.44
		CVD		3.29		**3.53**		1.79
		CVA		5.36		**2.78**		5.45
	CAC	MS	-3.44	-1.53	**-12.85**	**2.05**	-6.35	-0.02
		癌症		-2.97		-3.21		-3.23
		CVD		-1.65		**0.43**		-3.01
		CVA		-1.24		-1.40		-1.77
	DNN	MS	-0.24	-0.12	1.62	0.46	2.68	0.17
		癌症		-2.30		-2.08		-2.19
		CVD		-2.12		-1.95		-2.62
		CVA		-2.77		-3.91		-1.23
	所提出的方法	MS	**-4.28**	**2.71**	**-5.32**	**2.52**	**-2.51**	**2.02**
		癌症		**0.14**		-0.63		**1.01**
		CVD		-0.64		-1.30		**1.58**
		CVA		-0.11		-0.81		**2.04**
正常	KDM	MS	**-2.90**	**0.96**	**-3.93**	**8.92**	2.37	0.05
		癌症		**1.58**		**5.46**		2.96
		CVD		**0.84**		**5.33**		4.40
		CVA		**1.46**		**5.50**		4.34
	CAC	MS	**-3.60**	**0.77**	**-9.21**	**4.82**	**-1.94**	**1.09**
		癌症		-0.86		**2.31**		-0.65
		CVD		-0.38		**2.57**		**0.86**
		CVA		-0.50		**2.47**		-0.31
	DNN	MS	-0.24	-1.61	1.57	-1.50	3.73	-0.43
		癌症		-3.50		-4.90		-3.28
		CVD		-3.66		-4.34		-3.48
		CVA		-3.87		-5.13		-3.51
	所提出的方法	MS	**-2.13**	**1.95**	**-0.71**	**3.75**	**-0.67**	-0.21
		癌症		**1.69**		**5.78**		**0.43**
		CVD		**1.89**		**7.56**		**1.51**
		CVA		**2.10**		**9.10**		**2.69**
超级正常	KDM	MS	**-0.81**	**0.61**	**-0.10**	**12.45**	0.85	-1.22
		癌症		-0.81		**8.38**		0.95
		CVD		-2.26		**7.32**		1.98
		CVA		-1.85		**7.49**		2.00
	CAC	MS	-0.96	-7.71	1.64	-4.31	0.80	-5.37
		癌症		-12.02		-9.67		-8.86
		CVD		-13.03		-10.72		-9.77
		CVA		-13.74		-11.59		-10.85
	DNN	MS	-0.73	-7.94	0.26	-21.64	-0.44	-4.71
		癌症		-9.30		-19.70		-9.13
		CVD		-11.19		-23.04		-10.35
		CVA		-11.67		-24.05		-10.72
	所提出的方法	MS	0.54	0.41	1.87	5.56	0.59	4.65
		癌症		0.28		7.31		5.29
		CVD		0.65		7.57		5.71
		CVA		0.67		8.05		6.01

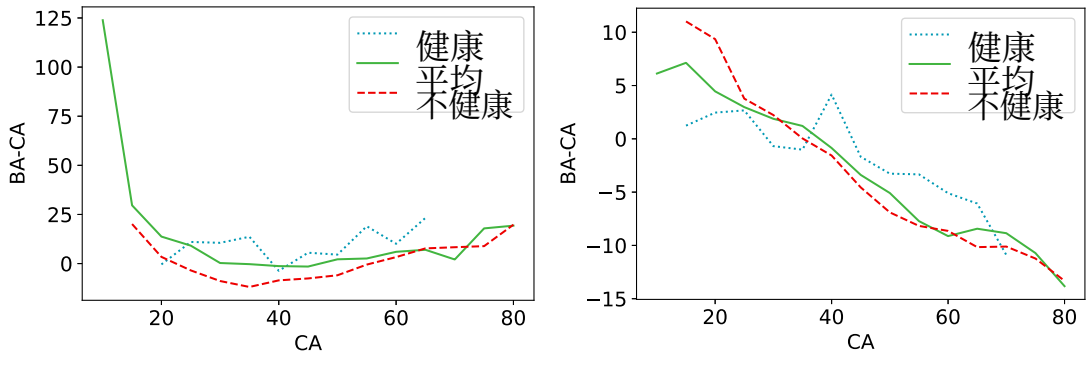
表 A4：女性受试者中健康（Hlty）和患病（withD）者在 MS、CA、CVD 和 CVA 发病率方面的平均 BA 间隔值。

人群	模型	疾病	整体		与发病率相关		*base*	
			Hlty	withD	Hlty	withD	Hlty	withD
Whole	KDM	MS	**-1.80**	**3.01**	**-22.68**	**42.92**	**-22.25**	**49.36**
		癌症		-0.25		-2.40		-4.92
		CVD		**1.11**		**13.36**		**20.07**
		CVA		**0.55**		**16.04**		**22.21**
	CAC	MS	-4.15	-1.67	**-7.78**	**3.97**	**-3.14**	**0.45**
		癌症		-8.33		-2.89		-4.30
		CVD		-4.90		**0.01**		-1.16
		CVA		-5.74		-1.30		-2.67
	DNN	MS	0.31	-0.64	0.52	0.11	1.76	-1.08
		癌症		-1.32		-2.01		-3.56
		CVD		-1.45		-2.20		-3.14
		CVA		-1.63		-2.22		-2.39
	所提出的方法	MS	**-3.29**	**3.66**	**-3.11**	**2.96**	**-0.59**	**0.97**
		癌症		**0.84**		**0.84**		-0.08
		CVD		**1.73**		**0.90**		**0.57**
		CVA		**1.32**		**1.21**		**0.92**
平均	KDM	MS	**-2.05**	**4.74**	**-9.43**	**24.81**	**-7.62**	**23.37**
		癌症		-0.07		**0.24**		**0.89**
		CVD		**1.21**		**2.98**		**4.74**
		CVA		**1.52**		**4.46**		**1.92**
	CAC	MS	**-6.03**	**7.81**	**-7.98**	**8.61**	**-4.67**	**7.19**
		癌症		-3.38		-1.58		-2.52
		CVD		-0.82		**0.50**		-1.17
		CVA		**0.25**		**1.62**		-1.14
	DNN	MS	0.85	0.42	0.08	-0.11	3.22	-0.63
		癌症		0.13		-2.06		-1.46
		CVD		-0.92		-2.50		-2.11
		CVA		-0.46		-1.42		-3.42
	所提出的方法	MS	**-3.47**	**3.93**	**-4.04**	**4.30**	**-1.70**	**5.84**
		癌症		**0.70**		**1.22**		**0.93**
		CVD		**1.59**		**1.89**		**1.97**
		CVA		**0.41**		**1.47**		**2.03**
正常	KDM	MS	**-8.60**	**20.48**	**-12.29**	**33.79**	**-9.70**	**32.16**
		癌症		**4.54**		**6.14**		**4.93**
		CVD		**8.53**		**12.87**		**15.46**
		CVA		**5.16**		**8.59**		**8.42**
	CAC	MS	**-6.01**	**8.71**	**-9.80**	**10.07**	**-4.47**	**4.26**
		癌症		-0.45		**1.03**		-2.24
		CVD		**1.67**		**2.46**		-0.40
		CVA		**0.38**		**1.92**		-1.92
	DNN	MS	1.44	-2.65	-0.11	-2.68	1.31	-4.11
		癌症		-1.12		-3.15		-3.99
		CVD		-2.74		-4.71		-5.23
		CVA		-1.62		-4.05		-5.04
	所提出的方法	MS	**-1.56**	**3.75**	0.31	-0.88	**-0.66**	**0.94**
		癌症		**2.27**		0.29		**0.44**
		CVD		**3.60**		0.96		**1.63**
		CVA		**2.89**		0.62		**1.12**

（续）健康（Hlty）和患病（withD）女性受试者的平均生物年龄差距值，按多发性硬化症、癌症、心血管疾病和脑血管意外的发病率划分。

人群	模型	疾病	整体		与发病率相关		*base*	
			Hlty	withD	Hlty	withD	Hlty	withD
超级正常	KDM	MS	**-0.16**	**17.08**	**-0.79**	**43.75**	**-0.24**	**33.54**
		癌症		**8.78**		**19.70**		**13.13**
		CVD		**9.79**		**27.27**		**20.66**
		CVA		**7.47**		**22.57**		**16.38**
	CAC	MS	-0.28	-2.38	1.10	0.33	1.17	-1.91
		癌症		-1.22		1.74		-1.37
		CVD		-4.10		-2.08		-4.38
		CVA		-3.61		-1.11		-3.66
	DNN	MS	-1.82	-13.00	-0.33	-16.53	-0.10	-16.35
		癌症		-8.66		-9.99		-10.63
		CVD		-13.17		-16.15		-15.91
		CVA		-11.46		-13.06		-14.08
	所提出的方法	MS	0.05	1.72	2.08	9.29	-0.22	0.85
		癌症		1.51		6.90		0.46
		CVD		1.75		8.16		1.24
		CVA		1.15		7.63		1.05

B 根据实际年龄和发病率划分的平均差值



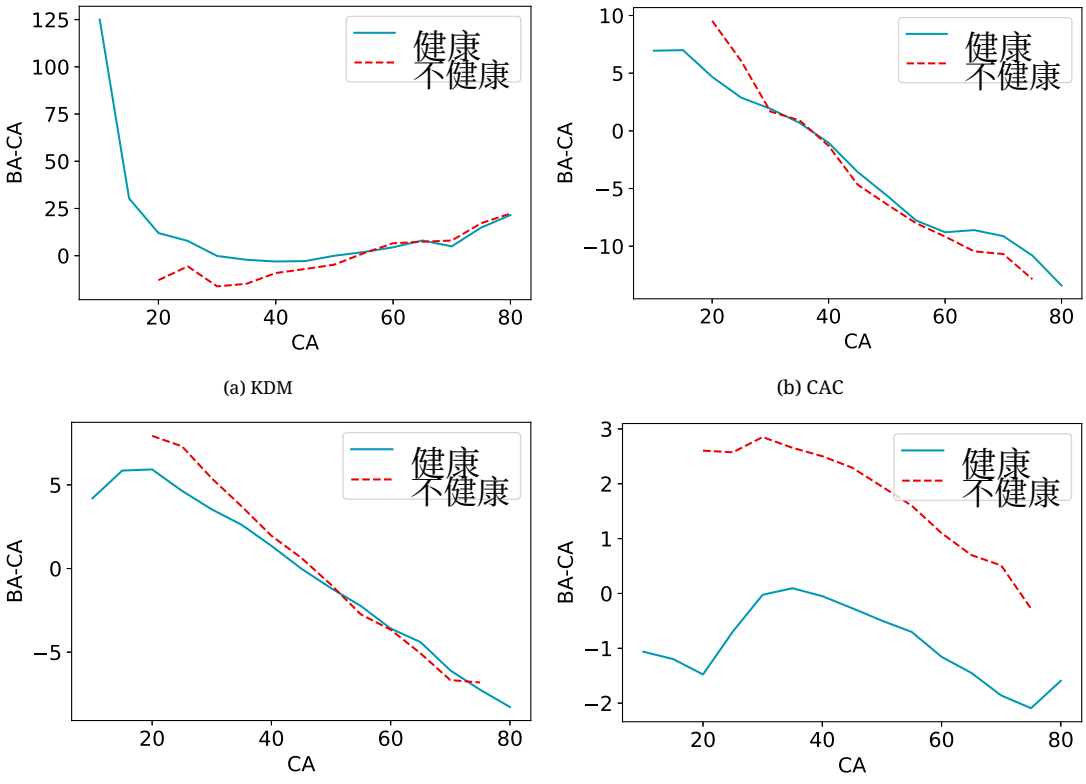
(a) KDM

(b) CAC

(c) 深度神经网络

(d) 所提出的方法

图 B1：男性整体病例中，生物学年龄与实际年龄之间的差距取决于**DLP**的发病率。



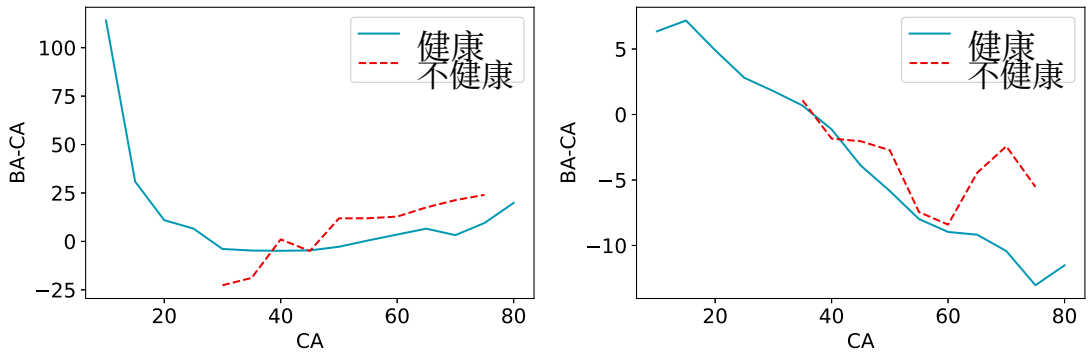
(a) KDM

(b) CAC

(c) 深度神经网络

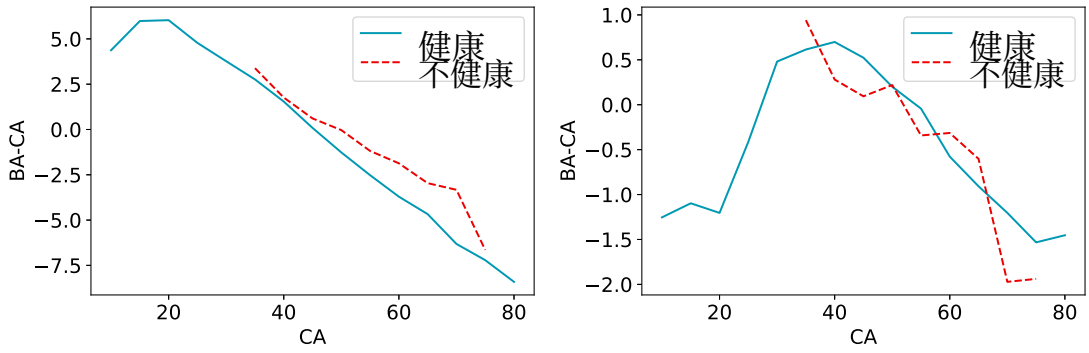
(d) 所提出的方法

图 B2：男性-*Whole*-整体病例中，生物年龄与实际年龄之间的差距取决于**多发性硬化症** 的发病率。



(a) KDM

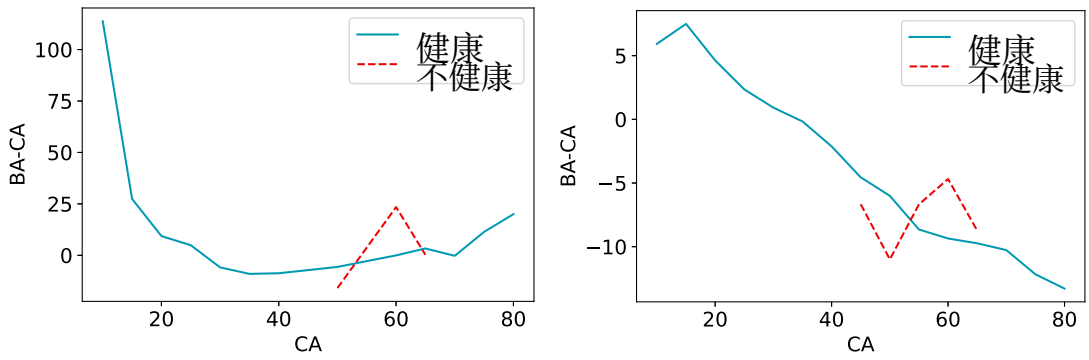
(b) CAC



(c) 深度神经网络

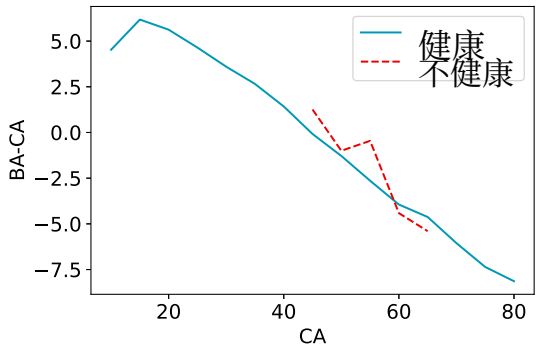
(d) 所提出的方法

图 B3：男性-*whole*-整体 病例中生物年龄与实际年龄之间的差距，具体取决于**心血管疾病** 的发病率。

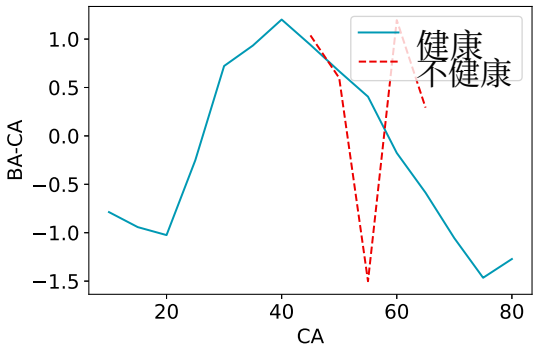


(a) KDM

(b) CAC

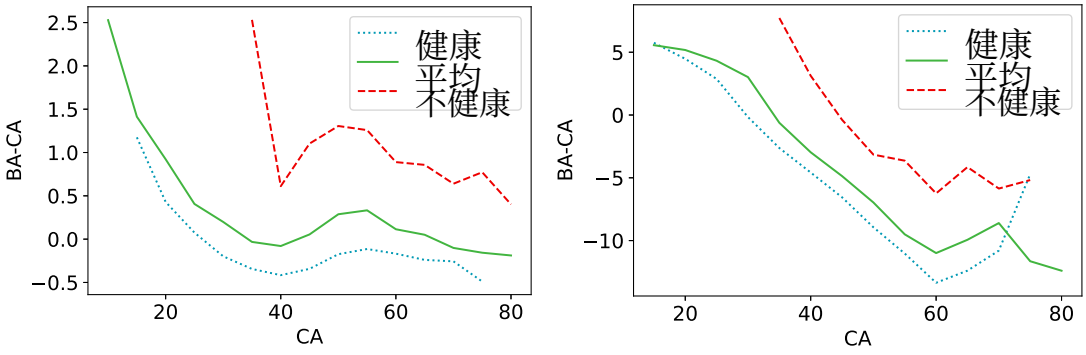


(c) 深度神经网络



(d) 所提出的方法

图 B4：男性-*Whole*-整体病例中生物年龄与实际年龄之间的差距，取决于**脑血管意外** 的发病率。



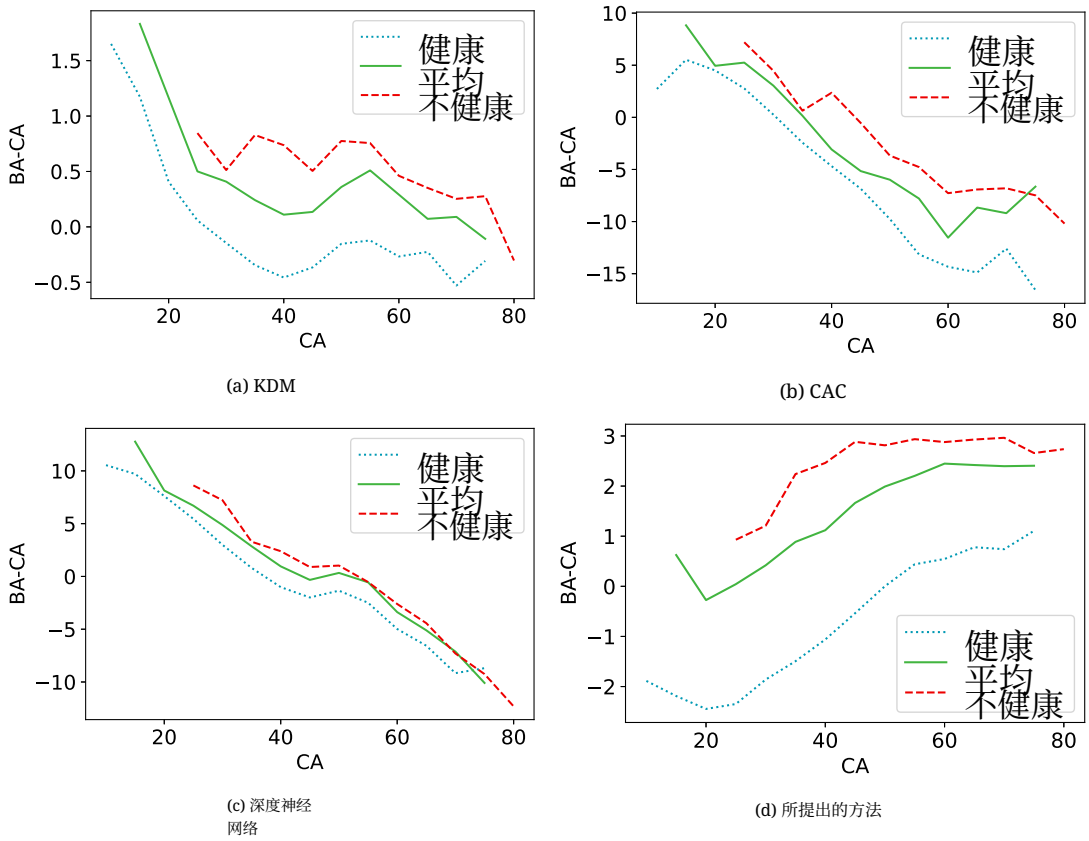
(a) KDM

(b) CAC

(c) 深度神经网络

(d) 所提出的方法

图 B5：在女性-**糖尿病** 女性-*Whole*-整体 病例中，生物学年龄与实际年龄之间的差距随发病率变化的情况。



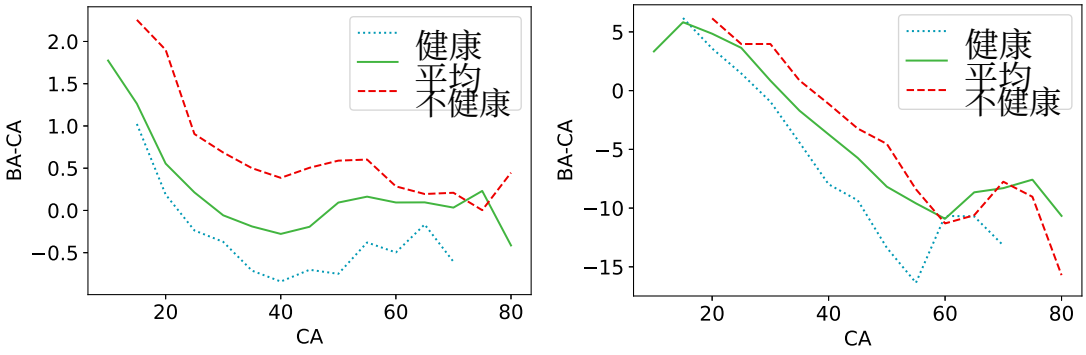
(a) KDM

(b) CAC

(c) 深度神经网络

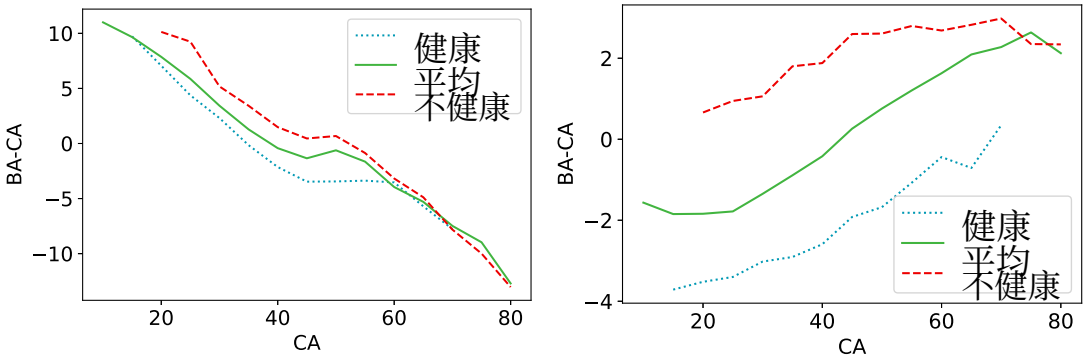
(d) 所提出的方法

图 B6：在女性-**高血压** -*Whole*-整体 病例中，生物学年龄与实际年龄之间的差距随发病率变化的情况。



(a) KDM

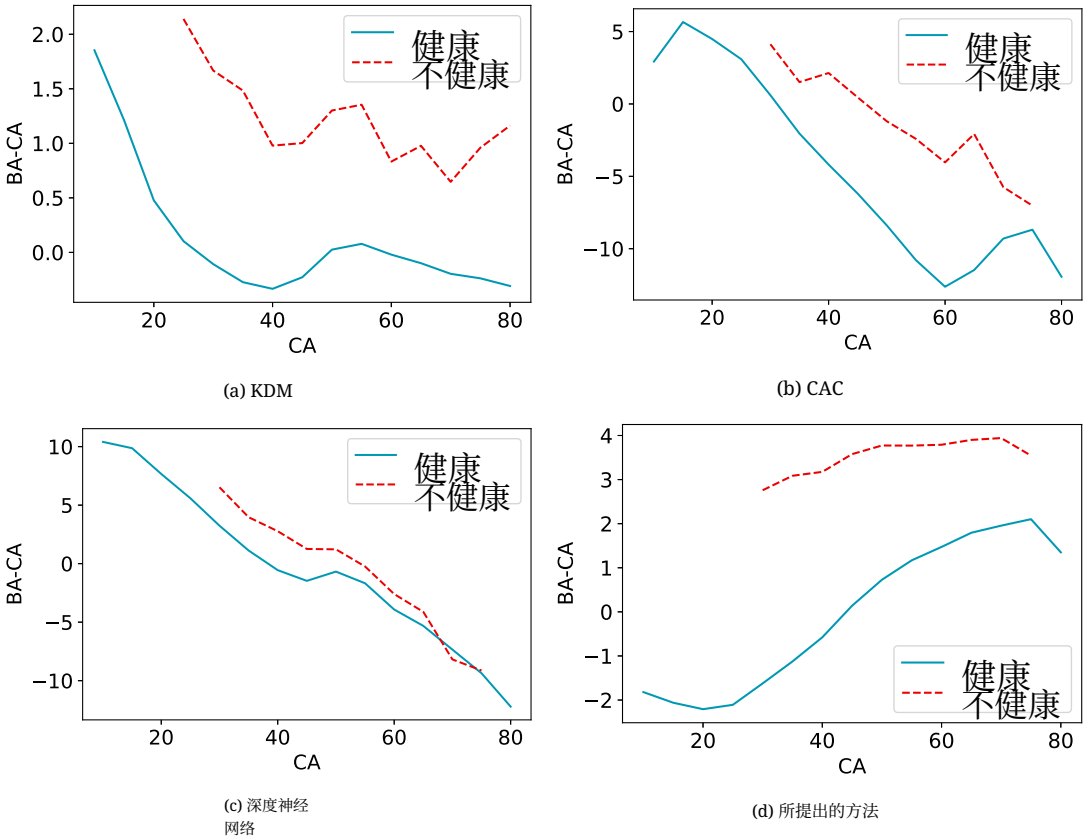
(b) CAC



(c) DNN

(d) 所提出的方法

图 B7：在女性-*Whole*-整体 病例中，生物学年龄与实际年龄之间的差距取决于**DLP**的发病率。



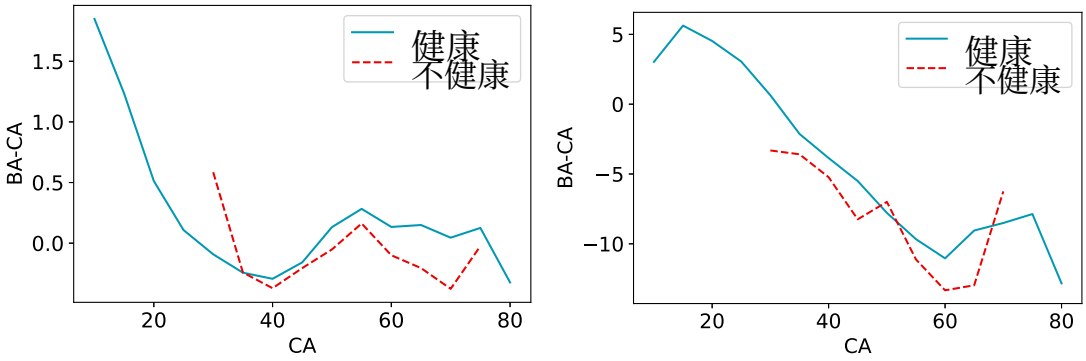
(a) KDM

(b) CAC

(c) 深度神经网络

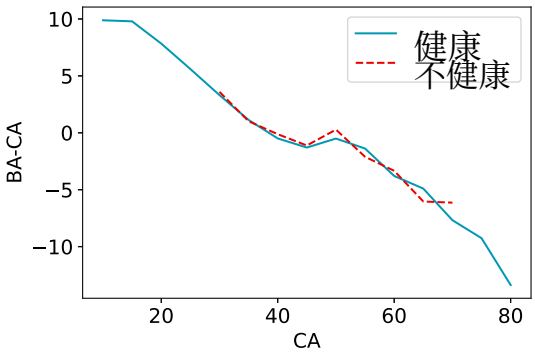
(d) 所提出的方法

图 B8：女性-*Whole*-整体病例中，生物年龄与实际年龄之间的差距取决于**多发性硬化症** 的发病率。

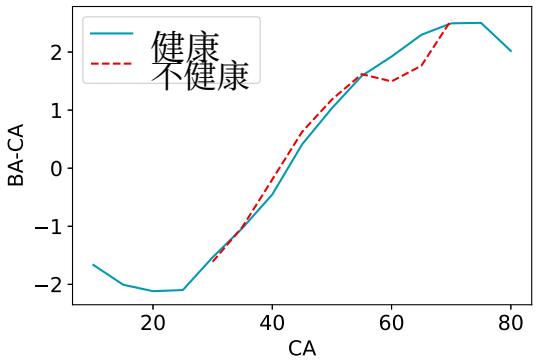


(a) KDM

(b) CAC

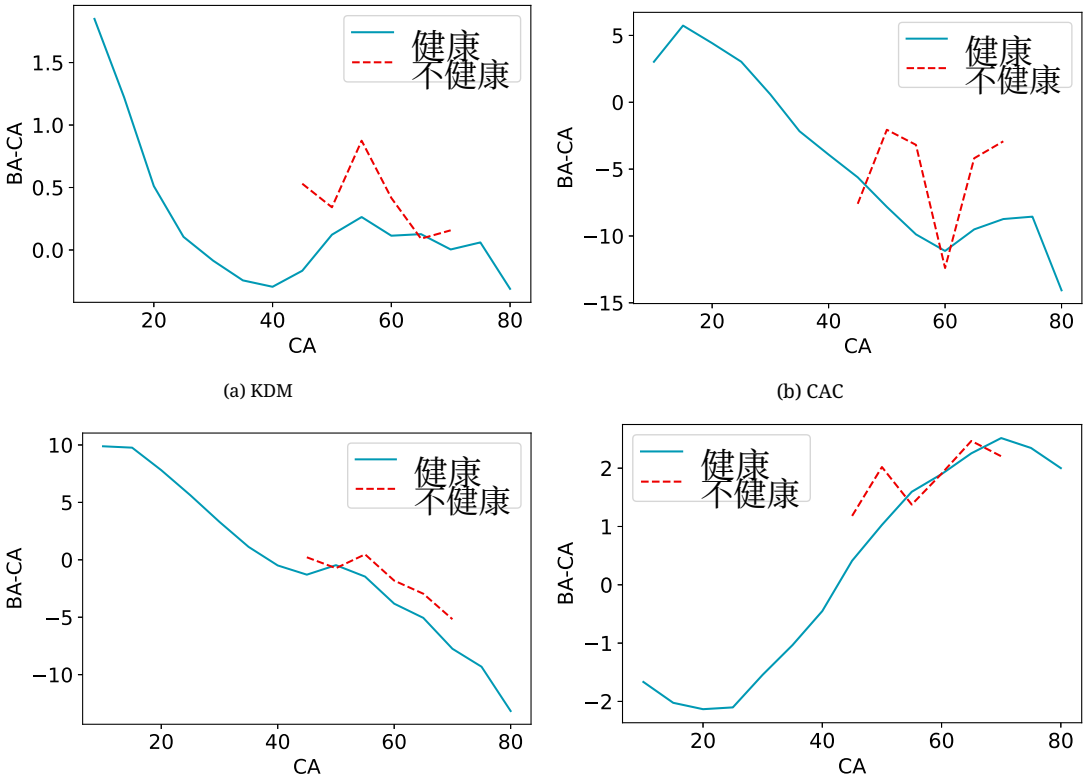


(c) 深度神经网络



(d) 所提出的方法

图 B9：女性-*Whole*-整体病例中，癌症发病率所对应的生物年龄与实际年龄之间的差距。



(a) KDM

(b) CAC

(c) 深度神经网络

(d) 所提出的方法

图 B10：女性-*whole*-整体 病例中，生物年龄与实际年龄之间的差距取决于**心血管疾病** 的发病率。

C 死亡时间的多元线性回归结果 h

表 C1：使用实际年龄和差值评估全体人群死亡时间的线性回归结果。斜率和皮尔逊相关系数根据差值计算。

人群	特征集	模型	斜率		R^2		PCC	
			CA	BA	CA	BA	CA	BA
M	整体	KDM	-4.683	**-2.737*****	0.001	0.026	-0.032	-0.144
		CAC		**-14.891*****		0.031		-0.168
		DNN		-0.602		0.001		0.013
		所提出的方法		**-153.599*****		**0.120**		**-0.342**
	与发病率相关	KDM	-1.601	**-6.302*****	0.000	0.018	-0.012	-0.132
		CAC		-11.034**		0.010		-0.096
		DNN		**-19.891*****		0.012		-0.086
		所提出的方法		**-139.756*****		**0.101**		**-0.317**
	base	KDM	6.959	**-8.209*****	0.002	0.027	0.048	-0.158
		CAC		-6.394		0.006		-0.069
		DNN		-14.627**		0.009		-0.096
		所提出的方法		**-112.617*****		**0.080**		**-0.273**
F	整体	KDM	3.794	12.141	0.001	0.008	0.029	0.011
		CAC		-6.2519		0.005		-0.072
		DNN		-0.998		0.001		-0.021
		所提出的方法		**-167.651*****		**0.086**		**-0.274**
	与发病率相关	KDM	13.708	0.706	0.011	0.014	0.106	0.047
		CAC		2.625		0.012		-0.010
		DNN		21.222		0.026		0.037
		所提出的方法		**-152.662*****		**0.074**		**-0.221**
	base	KDM	11.272	4.018	0.007	0.032	0.082	0.153
		CAC		9.750		0.016		0.076
		DNN		23.433		0.023		0.064
		所提出的方法		**-95.941*****		**0.049**		**-0.210**