一种基于Transformer的深度学习方法，用于公平预测肝移植后风险因素

李灿，硕士[1], 姜晓倩，博士[2] 和张凯，博士[2]

[1]德克萨斯大学休斯顿健康科学中心公共卫生学院生物统计学与数据科学系，德克萨斯州休斯顿

[2]德克萨斯大学休斯顿健康科学中心生物医学信息学麦克威廉姆斯学院健康数据科学与人工智能系，德克萨斯州休斯顿市

通讯作者：Kai Zhang（Kai.Zhang.1@uth.tmc.edu）

摘要

肝移植是终末期肝病患者的救命治疗。肝移植面临两大主要挑战：为供者找到最匹配的患者，以及确保不同亚群之间的移植公平性。现行终末期肝病模型（MELD）评分系统用于评估患者在90天内未获得器官时的死亡风险。然而，供体-患者匹配还应考虑移植后的风险因素，例如心血管疾病、慢性排斥反应等，这些都是移植后的常见并发症。准确预测这些风险评分仍是一项重大挑战。本研究采用预测模型来解决上述挑战。具体而言，我们提出了一种深度学习模型，用于预测肝移植后的多种风险因素。通过将其构建为多任务学习问题，所提出的深度神经网络经过训练，可同时预测五种 移植后风险，并利用任务平衡技术，实现各项任务同样出色的表现。我们还提出了一种新颖的公平实现算法，以确保不同亚群之间的预测公平性。我们使用了电子健康记录

这些记录涵盖160,360名肝移植患者的人口统计学信息、临床变量和实验室数值，采集自美国1987年至2018年的肝移植记录。我们使用AUROC和AUPRC等多种性能指标评估模型表现。实验结果凸显了多任务模型在实现任务平衡的同时保持准确率的优势。该模型将任务差异显著降低了39%。进一步应用公平实现算法后，各风险因素在所有敏感属性（性别、年龄组和种族/族裔）上的公平差异均大幅降低。这凸显了将公平性考量融入任务平衡框架的效用，从而确保模型能够针对多项任务和不同人口群体做出稳健且公平的预测。

关键词：公平性、肝移植、风险预测、多任务学习

1. 引言

肝移植是严重肝病患者的重要治疗方案。尽管它为终末期肝病患者提供了挽救生命的治疗，但人们仍对潜在的死亡风险和移植后并发症存有担忧。肝移植后的风险预测是一个重要的研究领域，而机器学习在开发预测模型方面日益重要。

多项研究探讨了使用机器学习算法预测移植后存活情况并识别高风险患者。例如，Nitski et al. 指出，肝移植受者在移植一年后患癌症的风险会增加；在预测移植后1年和5年因心血管原因、感染、移植物失败或癌症导致的死亡方面，深度学习模型表现更佳 [1]。Kantidakis et al. 使用 Cox 和机器学习模型预测移植后存活率，并发现再次移植是移植物失败或死亡的最强预测因素 [2]。

我们的工作首次尝试同时预测肝移植患者的多种风险因素。以往研究采用简单的临床评分系统（如 MELD（终末期肝病模型）评分），或建立模型来预测单一风险。例如，MELD 评分仅在患者尚未接受肝移植时预测其90天死亡率。然而，肝移植的成功不仅取决于患者的存活情况，还取决于移植后风险，例如心血管疾病、排斥反应、糖尿病等。在为患者匹配供者时，这些风险因素至关重要。

目前针对每个风险因素分别建立预测模型的策略，缺乏确保这些模型性能相当的标准化方法。因此，由于这种方法未能充分利用各风险之间的相互关系，某些风险预测的表现不如其他预测。研究表明，移植后的风险因素，如恶性肿瘤、糖尿病、排斥反应、感染和心血管并发症，具有高度的相互依赖性 [3], 这表明，联合预测方法将比单一风险预测更为有效。当任务之间存在相关性时，多任务预测模型已展现出卓越的性能 [4]。然而，多任务学习的有效性取决于是否拥有规模足够大且质量良好的数据集，而现有方法受到数据稀缺和过拟合风险的限制。在本研究中，我们提出采用多任务平衡技术，对肝移植后的多个临床关键风险因素（恶性肿瘤、糖尿病、排斥反应、感染和心血管并发症）进行联合预测，并取得了较高的预测性能。

此外，由于器官捐赠有限，肝移植还面临着如何确保不同亚群体之间公平性的挑战。大量研究指出，肝移植获取机会存在地区和人口统计学方面的差异 [5,6]。例如，由于肝移植获取机会存在地域差异，全国各地的供需比各不相同，导致肝移植率存在显著的地区差异 [7]。此外，人口统计学方面的差异

在肝移植获取机会方面存在差异。在终末期肝病模型时代之前，黑人患者的移植率低于白人患者 [8]。在终末期肝病模型时代，女性与男性之间的肝移植率差异有所扩大 [9]。解决这些差异对于确保所有人都能公平地获得肝移植至关重要。

人们已采取措施解决肝移植获取机会中的公平性与差异问题。为平衡器官分配的效率与公平性，OPTN/UNOS（器官获取与移植网络/美国器官共享联合网络）已实施相关政策 [10]。然而，仍需进一步努力，以确保所有人都能公平地获得肝移植。风险预测系统尤为重要，因为它们可应用于现实场景，实现有效的供体-患者匹配。因此，我们提出一种追求公平性的算法，以确保由性别、种族、族裔和年龄组等敏感属性定义的不同亚群体具有同等的预测性能。

我们提出的模型利用了大型肝移植数据库，其中包括全面的供者和受体人口统计信息、生命体征、实验室检查结果、诊断及治疗数据。该模型利用供者和受体双方的生物标志物来预测移植后风险，从而能够基于各种生物标志物更准确地预测供体-患者匹配。复杂的底层生物过程决定了供者与患者之间匹配的良好程度。

因此，我们研究的创新之处在于：

- 所提出的模型整合了供者和受体的生物标志物，以改进供者-受体匹配系统。
- 我们开发了一种基于 Transformer 神经网络和任务平衡技术的精准多任务预测模型，以实现多项风险预测的性能均等。

● 我们提出了一种公平实现算法，以确保由敏感属性界定的不同亚群体获得同等的预测性能。

2. 相关研究

预测肝移植后的风险十分重要，因为这有助于更好地指导患者筛选流程，从而实现器官的最佳匹配。同时，确保不同亚群的预测表现一致，对于保障公平性至关重要。

机器学习与深度学习模型： Lau 等人强调，在肝移植时预测移植物失败或原发性无功能的能力，对于优化有限供者肝脏资源的利用至关重要。在这方面，机器学习算法展现出了显著的效果 [11]。此外，Liu 等人运用机器学习预测肝移植后的短期生存情况，从而能够及早识别高风险患者[12]。此外，在移植后生存预测方面，Kantidakis et al. 指出，采用梯度提升、随机森林和决策树的机器学习模型优于传统 Cox 模型，从而鲜明地对比了传统统计方法与先进机器学习技术 [2]。Ferrarese 等人还开发了机器学习模型，用于预测肝移植患者的移植前生存情况。这些模型旨在帮助优化器官分配并降低等待名单上的死亡率 [13]。Nitski et al. 证明，Transformer、时间卷积网络、循环神经网络和多层感知机等深度学习模型能够纳入纵向信息，以预测肝移植后的长期结局。此外，他们还将这些模型与逻辑回归进行了比较，重点考察其在不同时间范围内预测导致死亡的移植后并发症的能力 [1]。

预测建模中的公平性：在机器学习 [14] 中，数据收集或模型训练过程中可能会引入偏差。因此，考虑这一点至关重要。

预测过程中的公平性。机器学习中的公平性算法可分为预处理、处理中和后处理 [15]。预处理算法着重于在模型训练过程中消除偏差。Calmon 等人提出了一种凸优化方法，以减少数据预处理中的歧视 [16]。后处理算法根据公平性标准调整训练后的模型预测结果。Pleiss 等人通过校准均等化赔率，最大限度地减少了不同人群之间的误差差异 [17]。 实现公平性的工作大多集中于处理中算法, ，这类算法通过将公平性指标作为优化的正则化项纳入模型训练，从而将公平性融入模型训练过程 [18]。Ding 等人采用知识蒸馏与两步公平机器学习框架，针对肝移植中的移植物失败预测开展研究 [19]。Li 等人开发了一种采用动态加权方案的多任务学习策略，以提升预测模型的公平性。该方法通过在神经网络反向传播过程中动态调整各个预测任务的梯度，克服了不可微公平性指标的限制 [20]。近期提出的一个名为 MDANN、基于对抗学习的框架，通过其框架内的对抗模块有效处理多个敏感特征上的偏差，在生物医学影像数据集上实现了公平预测 [21]。

多任务学习：多任务学习（MTL）是在多个相关任务上同时训练模型，使模型能够通过整合各任务的共性与差异，学习稳健且可迁移的表征。在相关任务上，MTL 往往优于单任务学习，并能同时最小化多个损失函数。然而，平衡学习率颇具挑战，因为梯度冲突可能导致某个任务占据主导地位，从而损害其他任务 [22]。由于各任务训练数据的样本量不同，任务之间的数据分布不均衡，会妨碍 MTL 的性能。焦点损失等技术有助于缓解任务间的类别不平衡问题 [23]。Lee 和 Son 提出了一种基于梯度的元学习方法，通过分离共享层与任务专属层来平衡多任务网络

。该方法通过单步梯度更新共享层，并采用内外循环训练，缓解任务间的梯度不平衡。这种方法提升了泛化能力和效率，并在计算机视觉 MTL 基准测试中取得了最先进的性能 [24]。Navon 等人提出了 Nash-MTL，这是一种多任务学习优化方法，将梯度组合视为任务之间的讨价还价博弈，以求得 纳什讨价还价解，从而平衡各任务的学习率和数据分布 [25]。CoD-MTL 框架采用多任务学习，联合分析肝移植后的死亡原因，重点关注排斥反应和感染 [26]。

3. 方法论

Transformer模型已变得极为流行，并彻底革新了自然语言处理，推动机器翻译和文本摘要取得了重大进展 [27]。TabTransformer是Transformer模型的一种变体，利用自注意力机制捕捉表格数据中特征之间的复杂交互关系 [28]。它通过嵌入技术，高效处理高基数分类变量。在本研究中，我们旨在使用TabTransformer同时预测多种移植后风险。

3.1. TabTransformer 多任务学习模型

我们提出了一种新颖的多任务 TabTransformer 深度学习模型，用于预测肝移植术后的五种风险因素。该模型由列嵌入、层归一化、堆叠的 Transformer 层和多层感知机组成。我们通过为每项风险预测调整输出层，扩展了 TabTransformer 架构以支持多任务学习。

多任务 TabTransformer 模型在联合预测框架中同时处理相关任务，通过在各任务之间共享知识来提升泛化能力和效率，并捕捉复杂的交互关系与依赖关系。

我们的 TabTransformer 多任务公平性实现预测模型的架构如**图 1**所示。

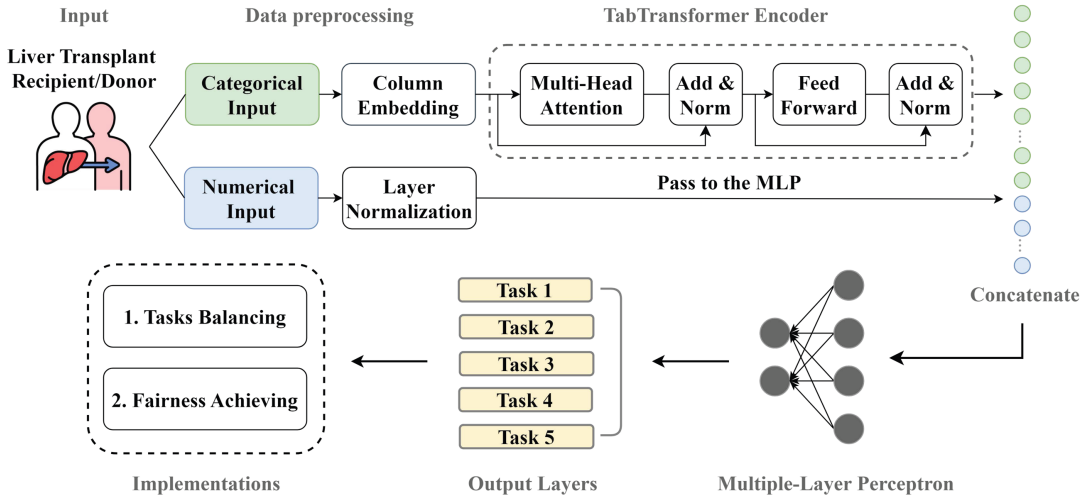


图 1。多风险公平性实现预测模型架构。

令 x_{cat} 表示所有类别特征，即 $x_{cat} = \{x_1, x_2, \ldots, x_p\}$，其中 p表示类别特征的数量。变量 x_{cont} 表示 R^q 中的所有数值特征，其中 q 表示数值特征的数量。在表格数据中，组合特征集 X 是连续特征与类别特征的并集，即 $X = x_{cat} \cup x_{cont}$。我们使用 $D = \{X, y_1, y_2, \ldots, y_M\}$ 来 表示一个数据集，其中每个 y_m（m 取值为 $\{1, \ldots, M\}$）表示第 m^t个任务的输出标签集，而 M 是任务数量。本文中，我们有五个任务，分别对应五种风险。我们针对 $\{y_1, y_2, \ldots, y_M\}$中的每个任务进行二元分类。

3.1.1. 列嵌入与层归一化

我们提出了一种生成类别特征嵌入的方法，称为 $E_\phi(x_{cat})$。$E_\phi(x_{cat})$ 由一组嵌入组成，即 $\{e_{\phi_1}(x_1), \ldots, e_{\phi_p}(x_p)\}$，分别对应每个类别特征。每个特征 x_j，其中 $j \in \{1, \ldots p\}$，都会通过列嵌入转换到一个维度为 d 的参数化度量嵌入空间。具体而言，每个类别列 j

都连接到一个嵌入查找表，记为 e_{ϕ_j}。对于具有 d_j 个类别的特征 j，其嵌入表 $e_{\phi_j}(.)$ 包含 $(d_j + 1)$ 个嵌入，其中额外的一个表示缺失值。

对于神经网络中的每个持续使用变量 $x_{cont} \in R^q$，我们进行层归一化，形式为 $LN_q = \gamma_q \odot N_q + \beta_q$，其中 γ_q 表示可学习的缩放参数，β_q 表示平移参数[29]。

3.1.2. 多层感知机（MLP）作为多任务输出层

参数化嵌入 $E_\phi(x_{cat})$ 作为初始 Transformer 层的输入。第一层的输出随后作为第二个 Transformer 层的输入，这一过程持续进行。通过最终的 Transformer 层，参数化嵌入逐步转换为上下文嵌入，并从其他嵌入中聚合上下文信息。Transformer 层函数序列 f_θ 处理参数化嵌入 $\{e_{\phi_1}(x_1),\ldots, e_{\phi_p}(x_p)\}$，并生成相应的上下文嵌入 $\{\mathrm{h}_1, \ldots, \mathrm{h}_p\}$，其中 $\mathrm{h}_p \in R^d$。

随后，将生成上下文嵌入的分类特征 $\{\mathrm{h}_1, \ldots, \mathrm{h}_p\}$ 与连续特征 x_{cont} 拼接，生成一个维度为 $(d \times p + q)$ 的向量。然后将该向量输入多层感知机（MLP），记作 g_ψ^m，用于每项任务 m 并预测目标 y^m。

在多任务模型中，使用一阶梯度方法最小化损失函数 $L_m(X, y^m)$，以端到端的方式训练 TabTransformer 的所有参数。这些参数包括用于列嵌入的 ϕ、用于 Transformer 层的 θ，以及用于 MLP 层的 ψ。我们实现了五个不同的预测网络，每个网络都采用 MLP 结构。每个 MLP 接收拼接后的向量，并输出概率的 logit。对于这种多任务设置，模型采用硬参数共享策略。

考虑全部五项任务的多任务模型的总体损失函数可定义为各项任务损失之和：

$$Loss(X,Y) = \sum_{m=1}^{M} L_m(X, y^m) \quad (1)$$

其中， $L_m(X, y^m)$ 是每个任务 m defined as$BCE(g_\psi^m(f_\theta \mathrm{L}E_\phi(x_{cat})\mathrm{M}, x_{cont}),$ $y^m)$ 的损失函数。这里，二元交叉熵（ BCE ）是该二元分类问题的损失函数。

3.1.3. TabTransformer 编码器

TabTransformer 模型包含多个 Transformer 层，每层均由一系列组件构成：多头自注意力机制、层归一化以及逐位置前馈网络 [28]。这些组件对输入特征进行处理，并生成精炼后的特征表示，供后续层使用。该机制为每个输入嵌入使用三个参数矩阵： $Key\ (K)$、$Query\ (Q)$ 和 $Value\ (V)$，它们各自处于不同维度。假设有 p 个输入嵌入，对应 p 个类别特征。在这种情况下， Key 和 $Query$ 矩阵可分别表示为 $K \in R^{p\times k}$ 和 $Q \in R^{p\times k}$ ，其中 k 是键向量和查询向量的维度。$Value$ 矩阵可表示为 $V \in R^{p\times v}$ ，其中 v 是值向量的维度。

随后，这些嵌入会输入自注意力机制，该机制会计算每对嵌入之间的注意力权重。计算公式为 $A = softmax((QK^T)/\sqrt{k})$，其中 $A \in R^{p\times p}$ 用于量化注意力权重。这些权重有助于突出显示各输入嵌入的相对重要性。

随后，利用注意力权重对 $Value$ 矩阵进行变换，从而生成原始输入特征的上下文表示。具体操作如下： $Attention\ (K, Q, V) = A \cdot V$。

最后，注意力头的输出（维度为 v）通过一个全连接层映射回原始嵌入维度 d 。这一最终输出是原始输入特征经过变换后的版本，将被馈送到后续层中，用于

预测任务 [28]。

3.2. 任务平衡方法

鉴于各项任务之间存在不平衡，在我们的方法中纳入任务平衡技术至关重要。我们提出了任务优先级排序的概念，并根据模型在特定任务上的预测性能来确定优先级。在以往的工作中，我们采用动态重加权，以促进由性别、种族和年龄等敏感属性定义的子群体之间的公平性。关键在于，根据这些任务在前一训练轮次（$t-1$）上的性能指标，在训练轮次 t 调整任务 m 的加权因子 $w_m(t)$。性能指标可根据要优化的具体评估指标 [20] 灵活选择。在本研究中，训练轮次 t 的修正损失函数是各项任务 m 的损失加权和，

$$Loss_{task-balancing}(t) = \sum_{m=1}^{M} \left(1 - w_m^{metric}(t-1)\right)^{\alpha} L_m(t) \qquad (2)$$

其中，$L_m(t)$ 是任务 m 在训练轮次 t 的损失函数，而 *(*指标*)* 是选定用于优化的评估指标。该方案根据前一训练轮次的性能 $w_m^{metric}(t-1)$ 调整各任务权重的大小，从而实现任务平衡。在每个训练轮次 t，算法会降低在该指标上的表现优于平均水平的任务的权重，并将更多关注转向另一项任务，反之亦然。参数 α 是一个超参数，用于调整权重的大小，进而决定加权强度。该方法旨在模型训练过程中实现任务平衡，以解决数据集所引发的不平衡问题。与以往工作不同，本文创新性地将动态重加权机制用于任务平衡这一新问题。我们并非简单地改造现有方法，而是展示了该方法的适应性，并将其应用拓展到新的领域。

在我们的 AUROC 加权模型中，我们采用动态重加权策略，可根据需要在不同任务之间平衡的特定评估

指标灵活选择评估指标。本研究中特别采用受试者工作特征曲线下面积（AUROC）。随后，我们将根据 AUROC 评估得分制定独特的动态重加权方案。该方案会在每项任务的反向传播过程中对梯度进行重新加权。这种方法可确保在训练过程中更加关注表现较差的任务，从而提升模型在所有任务上的整体平衡性和有效性。我们还可以使用精确率-召回率曲线下面积（AUPRC）、准确率、精确率或召回率等其他指标进行重加权。

3.3. 实现公平性的方法

为促进不同人口统计类别之间的公平性，我们提出了一种新的差异度量 D_m。该度量通过计算最具优势子群体的预测值与所有子群体整体平均预测值之间的差异得出。

敏感属性（例如年龄、性别或种族/族裔）将整个总体划分为互不相交的子群体。例如，子群体可以是男性/女性（当敏感属性为性别时）；白人/黑人/亚裔等（当敏感属性为种族/族裔时）；等等。我们将由敏感属性定义的子群体 表示为集合 $A = \{1, \dots, |A|\}$。计算针对每种敏感属性进行调整，并且针对每个任务 m分别设定。在此情境下，方程中的 i 表示范围 $\{1, \dots, |A|\}$内的任意子群体。D_m 的数学表示为：

$$D_m = \max_{i \in A}(F_i^m) - E_{i \in A}[F_i^m] \tag{3}$$

其中，F_i^m 是我们关注的公平性指标。可根据待优化的具体公平性指标灵活选择该度量。例如，优化人口统计均等时，该指标为正预测率；优化均等化赔率时，该指标为真正例率与假正例率的平均值；等等。

公平性损失定义为所有任务上 D_m 的总和，

$$Loss_{fairness}(t) = \sum_{m=1}^{M} D_m \quad (4)$$

在这一策略中，我们纳入了广泛用于评估预测公平性的公平性指标，包括但不限于人口统计均等和均等化赔率，以减少每个敏感属性下不同子群体之间的差异。在多任务设置中，对于每个任务 $m \in [1, \ldots, M]$，Yi 表示预测结果，Y 表示真实标签。

人口统计均等是由敏感属性定义的所有子群体之间预测正率的平等。

$$\mathbb{P}\left(\hat{Y} = 1 \middle| A = i, Task = m\right) = \mathbb{P}\left(\hat{Y} = 1 \middle| A = j, Task = m\right), \forall\, i, j \in A \quad (5)$$

当人口统计均等是要在子群体之间进行优化的公平性指标时，组 i 上的每项任务公平性指标，即 F_i^m，定义为，

$$F_i^m \triangleq \mathbb{P}\left(\hat{Y} = 1 \middle| A = i, Task = m\right) \quad (6)$$

同样，均等化赔率是另一种广泛使用的公平性标准，要求所有子群体之间的真正例率和假正例率相等，

$$\begin{aligned} &\mathbb{P}(\hat{Y} = 1 | Y = 1, A = i, Task = m) = \mathbb{P}(\hat{Y} = 1 | Y = 1, A = j, Task = m), \forall\, i, j \in A, \\ &\mathbb{P}(\hat{Y} = 1 | Y = 0, A = i, Task = m) = \mathbb{P}(\hat{Y} = 1 | Y = 0, A = j, Task = m), \forall\, i, j \in A. \end{aligned} \quad (7)$$

F_i^m 定义为真正例率和假正例率的平均值，

$$F_i^m \triangleq \frac{\mathbb{P}\left(\hat{Y} = 1 \middle| Y = 1, A = i, Task = m\right) + \mathbb{P}\left(\hat{Y} = 1 \middle| Y = 0, A = i, Task = m\right)}{2} \quad (8)$$

F_i^m 可以在模型训练期间，在一个数据批次（记作 Fi$_i^m$）内通过经验进行估计。

3.4. 公平地平衡任务：统一方法

We 提出我们的端到端算法，以实现两个目标

——同时实现任务平衡并在模型训练期间确保公平性，见**算法1**。 该算法适用于各种任务平衡指标和公平性评估指标。因此，我们用“指标”表示用于平衡任务的通用性能指标，而 $Fi_i^m(t)$ 也可以定义为用于评估不公平性的“公平性指标”。所提出方法的目标是确保预测模型在所有任务上都能实现同等准确的预测，同时兼顾不同子群体（性别、年龄组、种族/族裔）之间的公平性。

该统一方法分两个步骤运行：1) 首先，根据所选的‘指标’执行任务平衡，以缓解任务不平衡。2) 接下来，使用所选的’公平性指标’，在每项任务内的子群体之间促进公平性。通过在初始任务平衡后应用公平性优化，该方法可以缓解每项任务中与敏感属性相关的偏差。

在后续章节中，我们将讨论详细分为两个不同的部分。首先，我们介绍一个仅专注于使用$Loss_{task-balancing}(t)$进行任务平衡的多任务模型。该模型称为“多任务AUROC加权模型”。随后，我们的第二个模型将任务平衡与公平性指标整合到总损失中：$Loss_{task-balancing}(t)+ Loss_{fairness}(t)$. 根据所采用的公平性指标，当选择人口统计均等作为指标时，该模型被指定为“多任务 AUROC加权 + 人口统计均等”，而在应用均等化赔率时，则被指定为“多任务 AUROC 加权 + 均等化赔率”。

3.5 对比研究

我们在 TabTransformer 模型中评估了多种方法，以比较各任务的预测性能和差异。

单任务方法：我们基础性的方法是针对单个风险因素应用二元预测。单个任务的损失函数可以定义为与该任务相关的 BCE 损失，表示为 $L_m(X, y^m)$。

单任务焦点损失 [30]**:** 为了改进常规的单任务学习，我们引入了焦点损失，以解决每个单独任务中的类别不平衡问题。它更加细致地聚焦于区分正类和负类。引入焦点损失的单个任务的损失函数可表示为

$$L_m^{FL}(X, y^m) = -\alpha_{balanced}(1 - prob_{balanced})^{\gamma}\log\left(prob_{balanced}\right) \quad (9)$$

此处，$prob_{balanced} = prob \times y^m + (1 - prob) \times (1 - y^m)$ 表示给定任务调整后的预测概率。$\alpha_{balanced} = \alpha \times y^m + (1 - \alpha) \times (1 - y^m)$ 是平衡后的加权因子。γ 是可调的聚焦参数，其值大于或等于 0。

多任务平衡 [24]**:** 为进行全面比较，我们还实现了 Lee 和 Son 提出的一种任务平衡策略，即单次梯度步更新方法。

首先，该方法计算每个任务的损失。随后，在保持任务专属层不变的情况下，通过单次梯度下降步更新共享层参数。此更新可用以下方程进行数学表示:

$$h_m^{share} \leftarrow h^{share} - \beta\nabla_{h^{share}} L_m\Big(f(h^{share}, h_m^{task})\Big) \quad (10)$$

之后，该方法使用更新后的共享层 $\mathrm{h}_{m,}^{s}{}^{are}$ 和原始任务专属层 h_m^{task}重新计算每个任务的损失。然后，使用所有任务的损失之和进一步优化共享层，公式如下:

$$h^{share} \leftarrow h^{share} - \eta\nabla_{h^{share}} \sum_{m=1}^{M} L_m(f(h_m^{share}, h_m^{task})) \quad (11)$$

其中，β 和 η 是步长。最后，根据以下公式更新任务特定层:

$$h_m^{task} \leftarrow h_m^{task} - \beta\nabla_{h_m^{task}} L_m(f(h^{share}, h_m^{task})) \quad (12)$$

This 方法提供了一种通过平衡多任务学习中的任务来

强调使用基于梯度的技术，分别更新共享层和任务专属层。

算法1. 融入公平性实现的任务平衡

输入 $D = \{X, y_1\ y_2, \ldots, y_M\}$
定义任务 $\{1, \ldots, M\}$
定义敏感属性（年龄、性别、种族/族裔）及其子群体 $\{1, \ldots, |A|\}$
定义用于任务平衡的评估指标，例如 AUROC、AUPRC，准确率、精确率，
和召回率。**定义**公平性实现所采用的公平性指标，例如人口统计均等（DP）、均等化赔率
（EO）**初始化**所有任务 m 的权重 $w_m(t)$
初始化 $\hat{F}_i^m(t)$，适用于所有任务 m 和子群体 i
对于每个训练轮次 t，在总计 _ 个训练轮次中执行
    对于每个批次 b，在批次集合中执行
        对于每个任务 m，在 M 中执行
        // 使用所选评估指标（AUROC 或其他指标）进行任务平衡
            计算 $L_m(t)$
            计算 $w_m^{metric}(t-1))$
        end for
        // 计算所有任务的任务平衡损失总和
        $Loss_{task-balancing}(t) = \sum_{m=1}^{M} \left(1 - w_m^{metric}(t-1)\right)^{\alpha} L_m(t)$
        // 计算每个任务中所选公平性指标对应的各子群体预测正率
每个任务**for** 每个任务 m in M do
            for 每个子群体 i in A do
                计算 $\hat{F}_i^m$ 的训练轮次 t，批次数据为 b
            结束循环
        结束循环
        // 计算特定属性中所有子群体之间的差异
        D_m =maxA$\hat{F}_i^m$B $- E_{i\in A}[\hat{F}_i^m]$, $E_{i\in A}[\hat{F}^m] = 1/|A| \sum_{i=}^{|A|} \hat{F}_i^m{}_1$
            $_i$// 计算所有任务中用于实现公平性的损失总和
        $Loss_{fairness}(t) = \sum_{=} D_{m\cdot/}$,
        // 任务平衡损失与公平性实现损失之和
        $Loss(t) = Loss_{task-balancing}(t) + Loss_{fairness}(t)$
    end for
end for

4. 实验设置

4.1. 数据集

我们使用由美国器官共享联合网络（UNOS）维护的器官获取与移植网络（OPTN）提供的数据，收集了从1987年至2018 [31]期间的160,360名肝移植后患者的大型队列。该数据涵盖广泛的信息，包括移植和等待名单登记、人口统计特征、实验室结果、移植事件以及随访数据。我们在实验中采用了5折交叉验证。数据按60%用于训练、20%用于验证以及20%作为留出测试集进行划分。我们报告了所有折叠上的平均测试性能及标准差，以便全面了解结果。

我们的研究特别关注五种会显著影响肝移植术后结局的风险因素（恶性肿瘤、糖尿病、排斥反应、感染和心血管并发症），如**图 2**所示。条形图展示了数据集中这五种与肝移植术后结局相关的风险因素的患病率。图中清楚显示了分布不均：排斥反应占比高达 23.77%，而占比最低的风险因素仅为 4.10%。这种分布不均给同时预测多项任务带来了相当大的挑战，凸显了在模型开发过程中考虑任务平衡的必要性。

如**图3**所示，各风险因素在肝移植术后病例中的比例随年份变化不一。排斥反应风险的比例明显下降，而糖尿病则呈现显著的峰值模式，在2005年前后达到最高点，之后有所下降。同样，恶性肿瘤病例的比例也呈现较为平缓的峰值趋势。相比之下，感染的比例随时间推移逐渐下降。与此同时，心血管风险在历年肝移植术后病例中的占比相对稳定。

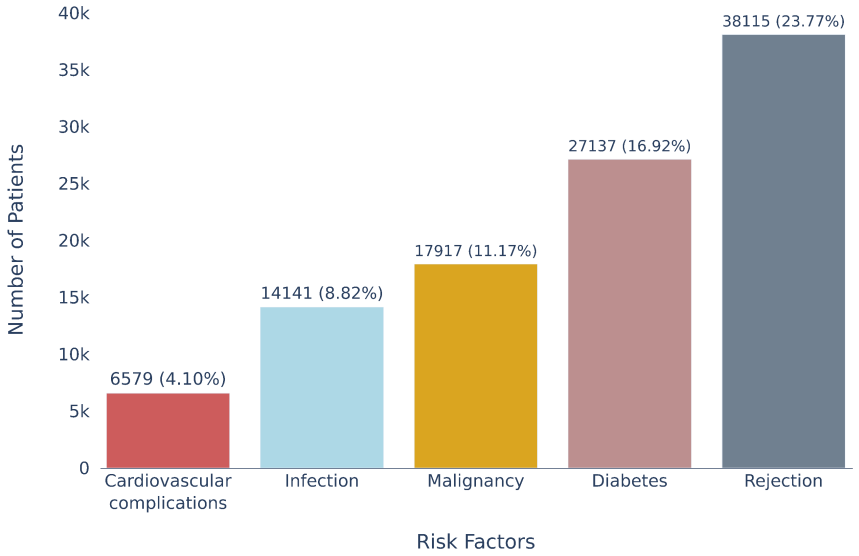


图 2. 肝移植后结局多重风险预测模型的架构。各风险因素对应的患者人数按升序排列，依次为心血管并发症、感染、恶性肿瘤、糖尿病和排斥反应。

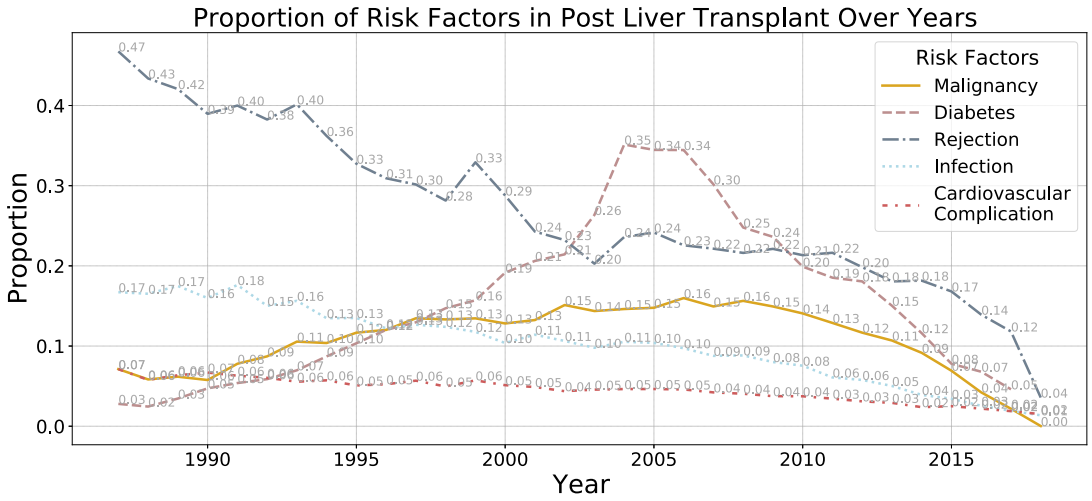


图 3。肝移植病例中各风险因素比例随年份变化的不均衡情况。恶性肿瘤：呈现中等程度的峰值趋势。糖尿病：比例逐渐上升，于 2005 年达到峰值，随后下降。排斥反应：所占比例随年份稳步下降。感染：发病率缓慢、逐渐下降。心血管并发症：在观察期内始终保持稳定。

表 1概述了肝移植数据集中患者的人口统计学信息，包括年龄、性别、族裔等多个方面，

种族和教育程度，这有助于识别肝移植人群中潜在的差异。

表 1。肝移植患者人口统计学信息

总计		160,360（100%）
Age		
	<15	15,135（9.44%）
	15-29	8,208（5.12%）
	30-44	23,056（14.38%）
	45-59	73,704（45.96%）
	60-74	40,005 (24.95%)
	>=75	252 (0.16%)
性别		
	Male	100,624 (62.75%)
	女性	59,736 (37.25%)
族裔/种族		
	白人	115,783 (72.20%)
	黑人	15,180（9.47%）
	亚裔	6,458（4.03%）
	西班牙裔	21,028（13.11%）
	其他	1,911（1.19%）
教育程度		
	None	604 (0.38%)
	小学（0-8年级）	8,129 (5.07%)
	高中（9-12年级）	48,799（30.43%）
	就读大学	26,124 (16.29%)
	副学士学位/学士学位	17,873 (11.15%)
	大学毕业后的研究生学位	7,552 (4.71%)
	不适用 (<5 岁)	9,105 (5.68%)
	未知	42,174 (26.30%)

我们的研究揭示了每种敏感属性（性别、年龄和种族/族裔）下各子群体在肝移植后风险因素方面存在差异，如**图4**所示。例如，儿科患者的排斥反应率为39.13%，远高于成人患者的21.97%。与此同时，成人患者其他风险因素的占比更高。例如，成人患者的糖尿病发生率为18.39%，而儿科患者的相应比例为4.43%。

种族和民族进一步揭示了不同子群体之间存在显著差异。例如，白人患者的恶性肿瘤发生率为12.81%，显著高于其他种族和民族子群体中的发生率。就性别而言，除排斥反应外，风险因素的差异似乎比其他方面更加均衡

敏感属性。与其他风险因素相比，排斥反应的差异最大。性别、年龄和种族方面的差异百分比分别为 6.04%、17.16% 和 9.77%。

这些差异凸显了消除不平等并制定针对性干预措施以满足肝移植后护理中不同子群体特定需求的迫切性。

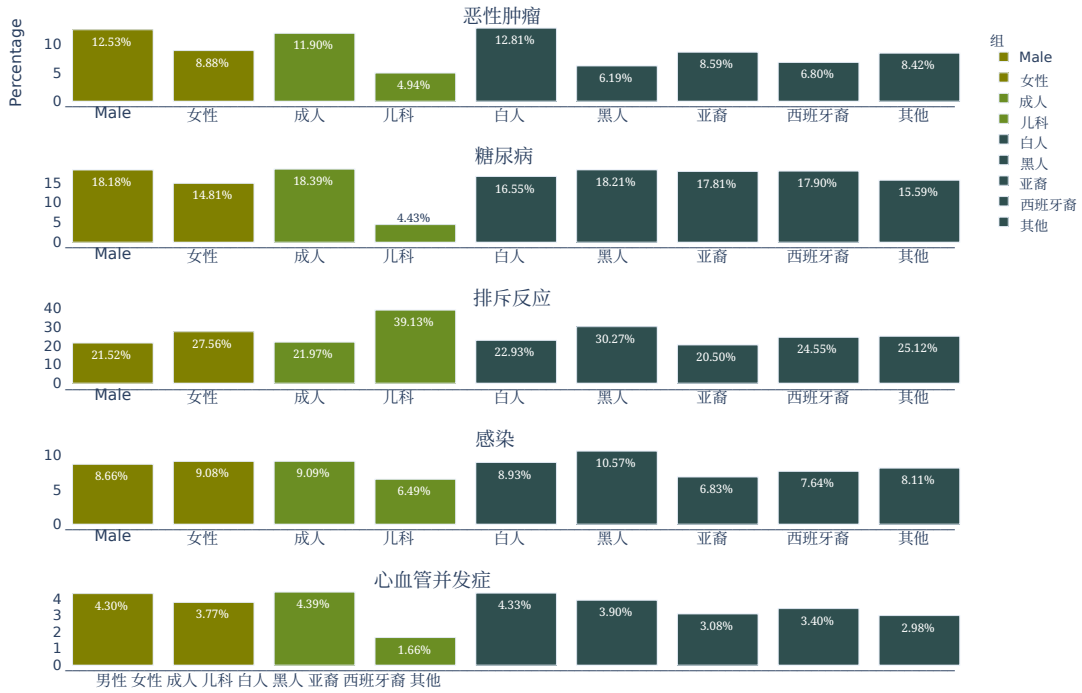


图4。 肝移植后五种风险因素中敏感属性子群体的差异。该图以条形图形式展示了五种风险——恶性肿瘤、糖尿病、排斥反应、感染和心血管并发症——在各个人口统计学子群体中的发生率。敏感属性分为三大类别：性别（男性、女性）、年龄组（成人、儿科）和种族/族裔（白人、黑人、亚裔、西班牙裔和其他）。每一行呈现一种特定风险，并比较其在三种敏感属性子群体中的患病率。

4.2. 特征选择与提取

移植后风险在很大程度上取决于供者与受体之间的匹配质量，而这本质上是一个潜在的复杂生物过程。因此，我们从患者中选取了52项综合特征，并从供者中额外选取了65项特征，以反映生物相容性。具体特征详见

表 S.1 和**表 S.2**。 与主要根据血型、体型、年龄、地理因素和四项血液检测来匹配患者与供者的MELD系统不同，我们的模型可以从历史数据中学习模式，利用丰富的特征优化供者与受体的匹配。[32],

所提出的模型可通过预测移植发生前五种风险的风险评分，为等待名单患者的供者-受体匹配提供指导。器官可用后，该模型可用作患者筛查工具。为此，我们仅选取肝移植前记录的数据特征，包括人口统计学信息和实验室检查结果。移植前特征能够深入反映患者移植前的整体健康状况，这对于理解所选风险因素的预测结果至关重要。

4.3. 评估指标

模型性能根据两个主要目标进行评估：(I) 实现五项任务之间的任务平衡；以及 (II) 确保每个敏感属性的子群体之间的公平性。

针对目标 (I)，我们评估不同任务的性能，并比较采用任务平衡算法前后各任务之间的最大性能差异。用于评估任务平衡的性能指标包括 AUROC 和 AUPRC。针对目标 (II)，我们比较应用公平性实现算法前后经过平衡的多任务模型的性能。考察的公平性指标包括人口统计均等差异和均等化赔率差异。

这些指标的定义如下：Yi：预测；Y：真实标签；A：每个敏感属性中的子群体；m：特定任务。

人口统计均等差异:

$$max_{i,j\in A}\left[P\left(\hat{Y}=1\middle|A=i,Task=m\right)-P\left(\hat{Y}=1\middle|A=j,Task=\ m\right)\right] \quad (13)$$

均等化赔率差异：

$$\begin{aligned} max_{i,j\in A}\big[P\big(\hat{Y}=1\big|Y=1,A=i,Task=m\big) \\ -P\big(\hat{Y}=1\big|Y=1,A=j,Task=m\big)\big] \\ +max_{i,j\in A}\big[P\big(\hat{Y}=1\big|Y=0,A=i,Task=m\big) \\ -P\big(\hat{Y}=1\big|Y=0,A=j,Task=m\big)\big] \end{aligned} \tag{14}$$

5. 结果

5.1. 评估任务平衡的有效性

我们的研究开展了一项消融实验，以评估不同方法在任务平衡方面的有效性及其性能。我们将所提出的多任务AUROC 加权模型与单任务二分类预测、采用焦点损失进行类别平衡的单任务方法、不进行任务平衡的多任务方法，以及采用单次梯度步更新的多任务学习方法进行了比较。如**表2**所示，所提出的模型有效提升了五个任务的任务平衡性能。为确保评估的一致性，所有方法均采用相同的 TabTransformer 架构和超参数。

对心血管并发症（任务 5）的结果分析显示，性能存在显著差异。单任务模型的 AUROC 和 AUPRC 分别为 0.6148（±0.0220）和 0.0598（±0.0039）。相比之下，未采用任务平衡的多任务模型表现更优，数值分别为 0.6572（±0.0084）和 0.0732（±0.0057）。我们的 AUROC 加权模型在任务平衡方面表现出色，准确率也最高，AUROC 为 0.6600（±0.0097），AUPRC 为 0.0753（±0.0081）。表格的最后一行通过记录每种方法在五项任务中 AUROC 或 AUPRC 的最大值与最小值之差，凸显了这些差异。 我们的 AUROC 加权模型最大限度地缩小了任务间差异，AUROC 和 AUPRC 的变化幅度分别为 0.0715 和 0.3150。与单任务模型 0.1172 的 AUROC 差异相比，我们将差异降低了 39%。这一降幅根据公式（初始差异－降低后的差异）/初始差异计算得出。

总之，结果揭示了一项重要发现：多任务学习会自然地利用相关任务之间的共享表示，因此能够超越单任务预测，这一点从较高的得分中可见一斑。此外，与类别平衡相比，任务平衡能够优化各项任务，并更好地应对不同的数据分布。重要的是，该模型实现了理想的平衡——在仅略微牺牲预测准确率的情况下，有效缩小了差异。

表 2。任务平衡模型的预测性能（任务 1：恶性肿瘤，任务 2：糖尿病，任务 3：排斥反应，任务 4：感染，任务 5：心血管并发症）。每项列出了各方法在 5 折交叉验证中的平均 AUROC 和 AUPRC 结果及标准差。最后一行显示了任务平衡模型五个任务之间性能指标的差异。另一方面，跨任务 AUROC 和 AUPRC 差异最小的方法以粗体和星号标出。

	单任务		**单任务（焦点损失）**		**多任务**		**多任务**[24]		**多任务 AUROC 加权**	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Task 1	0.6932 (±0.0068)	0.2103 (±0.0032)	0.6911 (±0.0115)	0.2087 (±0.0094)	0.7046 (±0.0066)	0.2195 (±0.0073)	0.7029 (±0.0073)	0.2149 (±0.0077)	0.7031 (±0.0047)	0.2175 (±0.0059)
Task 2	0.7325 (±0.0053)	0.3342 (±0.0095)	0.7331 (±0.0049)	0.3363 (±0.0092)	0.7398 (±0.0044)	0.3439 (±0.0093)	0.7392 (±0.008)	0.3411 (±0.0135)	0.7315 (±0.0066)	0.3336 (±0.0106)
Task 3	0.6633 (±0.0047)	0.3790 (±0.0062)	0.6664 (±0.0053)	0.3845 (±0.0116)	0.6710 (±0.0026)	0.3958 (±0.0048)	0.6708 (±0.0044)	0.3952 (±0.0071)	0.6683 (±0.0034)	0.3903 (±0.0058)
Task 4	0.7015 (±0.0073)	0.1826 (±0.0059)	0.6984 (±0.0113)	0.1801 (±0.0075)	0.7117 (±0.0082)	0.1923 (±0.0050)	0.7113 (±0.0089)	0.1928 (±0.0080)	0.7115 (±0.0099)	0.1926 (±0.0072)
Task 5	0.6148 (±0.0220)	0.0598 (±0.0039)	0.6162 (±0.0097)	0.0599 (±0.0036)	0.6572 (±0.0084)	0.0732 (±0.0057)	0.6571 (±0.0075)	0.0714 (±0.0040)	0.6600 (±0.0097)	0.0753 (±0.0081)
Diff	0.1176	0.3192	0.1169	0.3246	0.0826	0.3226	0.0821	0.3238	**0.0715***	**0.3150***

5.2. 评估公平实现算法的有效性

使用我们的多任务 AUROC 加权模型实施任务平衡方法后，我们进一步加入了公平性实现技术。我们对多任务 AUROC 加权模型及其公平性增强版本进行了比较分析。具体而言，我们采用了所提出的公平实现算法（使用人口统计均等或均等化赔率）。比较中使用人口统计均等差异和均等化赔率差异作为关键评估指标，详见**表 3**。

在我们对多任务AUROC加权模型进行的实验中，我们发现不同性别、年龄和种族群体之间的排斥反应风险（任务 3）存在显著差异。值得注意的是，排斥反应风险表现出显著的人口统计均等差异（DPD），为0.1642（±0.0051），以及均等化赔率差异（EOD）为

0.3058 (±0.0099)，尤其突出了成人与儿科子群体之间的差异。然而，将公平性指标——人口统计均等和均等化赔率——引入模型后，缓解了这些差异。采用多任务AUROC 加权+ 人口统计均等模型后，排斥反应（任务 3）的DPD 降至0.0039 (±0.0009)，EOD 降至 0.0222 (±0.0018)。另一方面，实施多任务AUROC加权 + 均等化赔率后，同一任务的DPD 为 0.0107 (±0.0016)，EOD为0.0134(±0.0038)。此外，对于性别、年龄和种族之间固有差异极小的风险，我们的公平性增强模型仍然减少了差异。采用所提出的公平性方法处理后，感染（任务4）和心血管并发症（任务 5）的差异均大幅减少，展现出稳健的差异最小化效果。

所有五项风险的预测概率累积分布函数（CDF）如**图 5**所示，涵盖每个敏感属性下的各种子群体。CDF 可全面展示这些子群体内预测风险的分布情况。例如，对于排斥反应风险（任务 3），多任务 AUROC 加权模型的 CDF 显示，在每个风险水平下，成人与儿科患者的比例存在差异。在预测风险为 0.25 时，约 70% 的成人患者处于该水平或低于该水平，而儿科患者中仅有 10% 达到这一比例。相比之下，对于所提出的公平性模型，CDF 更为接近地对齐——两个子群体中约 60% 的患者被预测为排斥反应风险为 0.25 或更低。这种对齐表明不同子群体之间的预测更加公平。同样，所提出的模型在其他敏感属性下也始终使不同子群体之间的 CDF 曲线保持一致，展现出其实现公平性的稳健性。为量化子群体 CDF 之间的差距，我们标注了人口统计均等差异（DPD）值，其中越接近 0 的值表示不同子群体之间的预测越公平。所提出模型较低的 DPD 值验证了其在不同人口统计群体中提供公平风险预测的有效性。

总体而言，所提出的实现公平性的方法显著减少了每项任务中不同性别、年龄和种族/族裔子群体在人口统计均等和均等化赔率方面的差异。这一成果证明了我们所提出的解决方案在实现不同子群体之间的公平性方面的有效性。

表 3。我们比较了多任务 AUROC 加权模型与两种公平性增强模型的公平性表现：多任务 AUROC 加权 + 人口统计均等和多任务 AUROC- 加权 + 均等化赔率。 性别、年龄组和种族/族裔之间的公平性比较。采用人口统计均等差异（DPD）和均等化赔率差异（EOD）作为评估指标。每个数值描述了不同子群体（男性与女性；成人与儿科；白人/黑人/亚裔/西班牙裔/其他）在某项指标上的最大差异。我们突出标示了公平性实现模型相较于多任务 AUROC 加权模型表现有所提升的情况。

		多任务 AUROC 加权		多任务 AUROC-加权 + 人口统计均等		多任务 AUROC 加权均等化赔率 +	
		DPD	**EOD**	**DPD**	**EOD**	**DPD**	**EOD**
Age	**Task 1**	0.0903 (±0.0040)	0.2142 (±0.0110)	**0.0033 (±0.0008)**	**0.0234 (±0.0015)**	**0.0009 (±0.0006)**	**0.0172 (±0.0041)**
	Task 2	0.1172 (±0.0078)	0.2486 (±0.0166)	**0.0017 (±0.0008)**	**0.0162 (±0.0026)**	**0.0031 (±0.0007)**	**0.0108 (±0.0025)**
	Task 3	0.1642 (±0.0051)	0.3058 (±0.0099)	**0.0039 (±0.0009)**	**0.0222 (±0.0018)**	**0.0107 (±0.0016)**	**0.0134 (±0.0038)**
	Task 4	0.0243 (±0.0025)	0.0476 (±0.0106)	**0.0022 (±0.0010)**	**0.0117 (±0.0039)**	**0.0012 (±0.0011)**	**0.0090 (±0.0054)**
	Task 5	0.0891 (±0.0062)	0.1975 (±0.0164)	**0.0041 (±0.0012)**	**0.0125 (±0.0064)**	**0.0010 (±0.0005)**	**0.0122 (±0.0072)**
性别	**Task 1**	0.0537 (±0.0030)	0.1131 (±0.0061)	**0.0009 (±0.0006)**	**0.0062 (±0.0021)**	**0.0035 (±0.0010)**	**0.0031 (±0.0014)**
	Task 2	0.0348 (±0.0015)	0.0659 (±0.0042)	**0.0010 (±0.0004)**	**0.0089 (±0.0030)**	**0.0029 (±0.0011)**	**0.0050 (±0.0029)**
	Task 3	0.0577 (±0.0032)	0.1136 (±0.0076)	**0.0015 (±0.0012)**	**0.0028 (±0.0014)**	**0.0047 (±0.0013)**	**0.0044 (±0.0031)**
	Task 4	0.0119 (±0.0016)	0.0222 (±0.0074)	**0.0007 (±0.0004)**	**0.0055 (±0.0039)**	**0.0013 (±0.0009)**	**0.0053 (±0.0041)**
	Task 5	0.0159 (±0.0030)	0.0333 (±0.0075)	**0.0008 (±0.0008)**	**0.0039 (±0.0038)**	**0.0011 (±0.0008)**	**0.0037 (±0.0022)**
种族/族裔	**Task 1**	0.0899 (±0.0065)	0.1885 (±0.012)	**0.0085 (±0.0034)**	**0.0184 (±0.0051)**	**0.0089 (±0.004)**	**0.0200 (±0.0080)**
	Task 2	0.0230 (±0.0051)	0.0488 (±0.0102)	**0.0083 (±0.0029)**	**0.0158 (±0.0061)**	**0.0065 (±0.0027)**	**0.0146 (±0.0051)**
	Task 3	0.0717 (±0.0043)	0.1422 (±0.0119)	**0.0095 (±0.0028)**	**0.0183 (±0.0058)**	**0.0113 (±0.0027)**	**0.0240 (±0.0080)**
	Task 4	0.0518 (±0.0046)	0.1049 (±0.0134)	**0.0098 (±0.0029)**	**0.0184 (±0.006)**	**0.0114 (±0.0034)**	**0.0241 (±0.0127)**
	Task 5	0.0384 (±0.0073)	0.0855 (±0.0115)	**0.0091 (±0.0028)**	**0.0216 (±0.0052)**	**0.0077 (±0.0037)**	**0.0270 (±0.0096)**

5.3. 评估特征重要性

完成任务平衡后，我们希望评估所提出的多任务 AUROC 加权模型在预测方案中可能引入的不公平性。为此，我们分析了模型中的特征重要性。我们采用基于置换的特征重要性评估技术，在患者之间打乱特征值

，并考察相应的模型性能变化。借助这一技术，我们能够识别出对模型性能有显著影响的特征。

对每种肝移植后风险因素的预测影响最大的五项特征见**图6a**。这些主要特征对 AUROC 性能的相对影响更大。例如，最近一次计算的体重指数结果 (END_BMI_CALC) 对预测糖尿病风险至关重要，而移植时受者腹水 (ASCITES_TX) 有助于预测排斥反应。值得注意的是，所提出的多任务 AUROC 加权模型在预测时高度依赖年龄组、种族/族裔和性别等人口统计学特征，而非生物标志物。例如，年龄组 (AGE_GROUP) 特征对预测心血管并发症最为重要。种族/族裔 (ETHACT) 是恶性肿瘤和排斥反应的首要特征。这表明，如果不施加公平性约束，预测模型可能会依赖不恰当的特征进行预测，从而导致不公平的结果。

在**图6b**和 **图6c**中，我们采用了 AUROC加权任务平衡方法，分别结合人口统计均等 (DP) 或均等化赔率 (EO)。这两种模型旨在减轻年龄组特征对模型性能的影响。我们观察到，年龄组特征不再出现在任何风险因素最重要的五项特征中，这证明了公平性干预措施的有效性。

此外，我们还将这些相同的实现公平性的方法应用于另外两个敏感属性：性别和种族/族裔。**图 S.1** 和 **图S.2**概述了这些干预措施对性别和种族特征在影响各种风险因素时的重要性的影响，进一步彰显了我们致力于提升多个敏感属性上的模型公平性的努力。

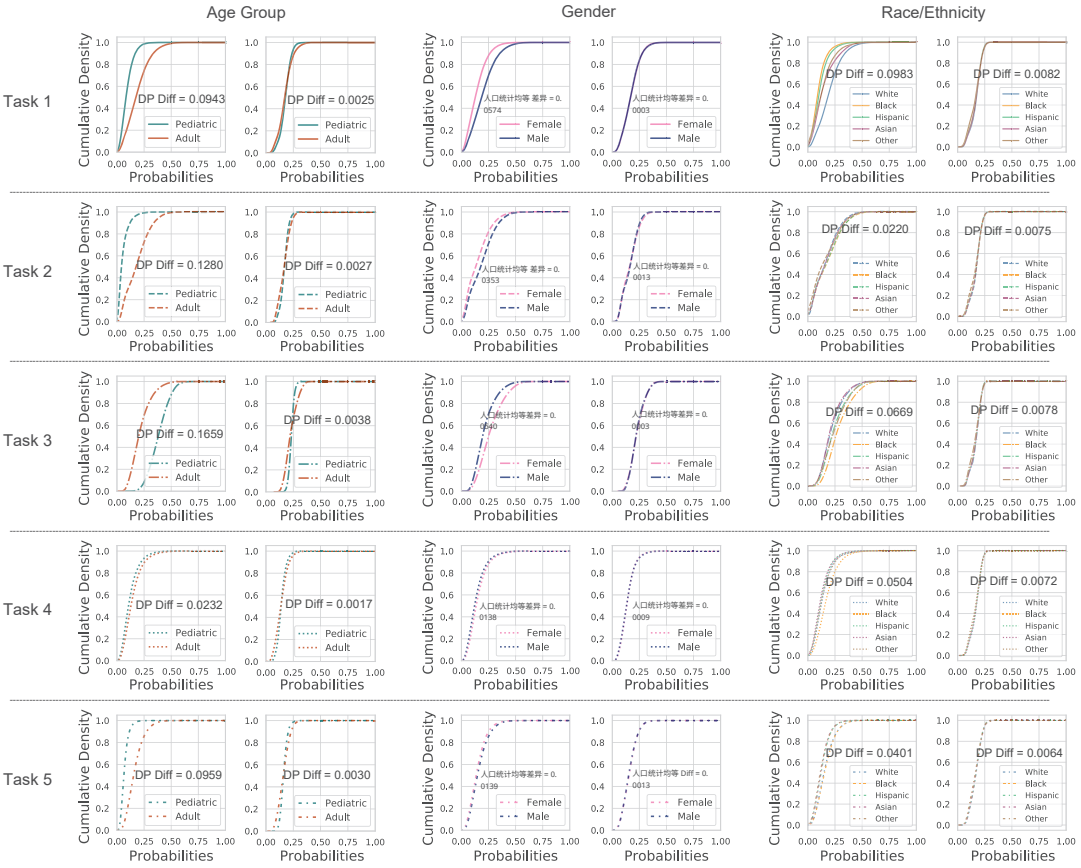


图 5. 使用多任务 AUROC 加权模型和实现公平性的模型，对不同敏感属性下预测概率的累积分布函数（CDF）进行比较。该图展示了三种敏感属性（年龄组、性别和种族/族裔）下五种风险的预测概率累积分布函数：恶性肿瘤（任务 1）、糖尿病（任务 2）、排斥反应（任务 3）、感染（任务 4）和心血管并发症（任务 5）。左侧展示了多任务 AUROC 加权模型在各敏感属性设定下生成的累积分布函数。右侧展示了所提出的实现公平性的算法得出的累积分布函数。在每种设定中，各条曲线分别代表相应敏感属性的具体子群体（例如，性别中的男性与女性、年龄组中的成人与儿科，以及种族/族裔中的白人、黑人、西班牙裔和亚裔等）。

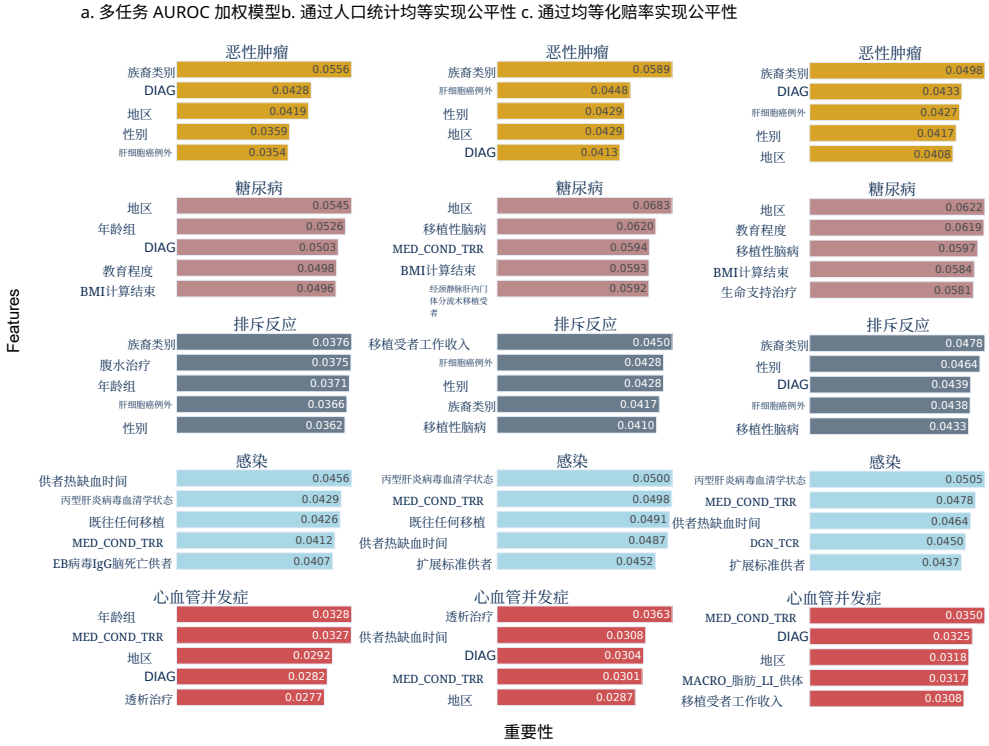


图 6。不同模型中肝移植后风险因素的前 5 项特征重要性比较分析。通过置换特征重要性确定的每个肝移植后风险因素的前 5 项特征重要性，呈现在三种不同设置中：a) 采用任务平衡的多任务 AUROC 加权。b) 采用 AUROC 加权任务平衡方法和实现公平性的方法的模型，以促进人口统计均等（DP），具体旨在减轻不同年龄组之间的偏差。c) 采用AUROC 加权任务平衡方法和针对均等化赔率的实现公平性的方法的模型，该模型同样关注年龄组之间的偏差。风险因素（从上到下）依次为恶性肿瘤、糖尿病、排斥反应、感染和心血管并发症。子图 a、b 和 c 中的每个图分别代表上述一种设置。

6. 结论

在本研究中，我们开发了一种采用任务平衡的多任务学习模型，以联合预测多个临床上重要的肝移植后风险因素。我们的基于 Transformer的神经网络模型实现了准确的多任务预测，并在各项风险因素之间取得了均衡的性能。此外，我们基于所提出的任务平衡模型，提出了一种实现公平性的算法，以确保不同人口亚群之间具有相同的预测性能。

我们提出的平衡实现和公平实现算法增强了对肝移植后风险因素的预测，有望帮助优化现有肝移植匹配系统中的供体-患者匹配。我们的方法也具有灵活性，并适用于其他领域。评估指标和公平性指标的选择可根据任务的具体需求和特征进行调整。因此，我们的工作展示了多任务学习结合任务平衡和公平度量如何提升对多个关键临床风险因素的预测能力。

7. 讨论

总之，我们的多任务学习模型有效结合了任务平衡与公平性实现算法，为预测肝移植后的多种风险因素提供了一个前景良好的框架。尽管该模型在平衡各任务的表现和减轻人口统计学差异方面表现稳健，但其效能仍受到一些局限因素的影响。

首先，所提出的方法侧重于模型偏差，具体而言，是预测模型在训练和推理过程中产生的偏差。通过设计新的训练策略和损失函数，可以显著降低模型偏差。所得模型的性能和公平性可能对这些超参数较为敏感，因此需要投入额外精力进行微调，以在预测准确性与模型公平性之间实现最佳权衡。

此外，与旨在实现公平性的算法类似，任务平衡算法往往会牺牲模型的整体准确率，以确保模型在多个任务上的表现公平。这给设计优化良好的损失函数和优化技术带来了挑战，因为它们需要在三个目标之间找到最佳折中：预测准确性、任务平衡，以及不同子群体之间的公平性。

本研究分析了1987年至2018年的数据，凸显了医疗实践和患者人口统计学随时间发生的复杂变化。尽管我们采用了

针对缺失值和离群值的链式方程多重插补（MICE）[33], 可能无法完全捕捉时间动态。虽然我们的主要关注点是肝移植中的多风险预测和公平性，但仍值得承认时间动态及其潜在影响。未来研究应进一步探讨过去几十年来肝移植患者人口统计特征的变化。

总之，我们提出了一种创新的模型和算法，以应对上述挑战。我们的研究动机源于对肝移植领域一项关键问题的观察：该领域亟需准确的预测和公平性。除了移植成功或失败这一二元结果外，多个常被忽视的移植后风险因素也同样重要。这些因素共同影响移植的整体成功，进而影响患者的生存时间和生活质量。因此，我们的研究首次尝试开发一种模型，能够以相当的准确率预测多种风险，同时确保不同子群体之间的公平性。

致谢

XJ 是 CPRIT 癌症研究学者（RR180012）。他还获得了 Christopher Sarofim 家族讲席教授职位、UT Stars 奖、UTHealth 创业基金，以及美国国立卫生研究院（NIH）项目资助（编号：R01AG066749、R01AG066749-03S1、R01LM013712、U24LM013755 和 U01TR002062）和美国国家科学基金会（NSF）项目资助（编号：2124789）的部分支持。本研究还获得了卫生资源与服务管理局合同 HHSH250-2019-00001C 的部分支持。本文内容仅由作者负责，不一定反映卫生与公众服务部的观点或政策；提及商品名称、商业产品或组织，也不代表美国政府对其予以认可。

参考文献

[1] O. Nitski, A. Azhie, F.A. Qazi-Arisar, X. Wang, S. Ma, L. Lilly, K.D. Watt, J. Levitsky, S.K. Asrani, D.S. Lee, B.B. Rubin, M. Bhat, B. Wang，肝移植受者长期死亡风险分层：在纵向数据上实时应用深度学习算法，Lancet Digit Health. 3 (2021) e295–e305.[2] G. Kantidakis, H. Putter, C. Lancia, J. de Boer, A.E. Braat, M. Fiocco，自肝移植以来的生存预测模型——Cox 模型与机器学习技术的比较，BMC Med. Res. Methodol. 20 (2020) 277.[3] P.R. John, P.J. Thuluvath，糖尿病患者肝移植的结局：一项病例对照研究，Hepatology. 34 (2001) 889–895.[4] Y. Zhang, Q. Yang，多任务学习综述，IEEE Trans. Knowl. Data Eng. 34 (2022) 5586–5609.[5] B. Parent, A.L. Caplan，公平就是公平：我们必须重新分配用于移植的肝脏，BMC Med. Ethics. 18 (2017) 26.[6] S.E. Gentry, A.B. Massie, S.W. Cheek, K.L. Lentine, E.H. Chow, C.E. Wickliffe, N. Dzebashvili, P.R. Salvalaggio, M.A. Schnitzler, D.A. Axelrod, D.L. Segev，通过重新划分区域来解决肝移植中的地域差异，Am. J. Transplant. 13 (2013) 2052–2058.[7] S. Akshat, S.E. Gentry, S. Raghavan，公平分配肝移植的异质供者圈，Health Care Manag. Sci. (2022). https://doi.org/10.1007/s10729-022-09602-7.[8] P. Dureja, M.R. Lucey，终末期肝病模型时代之后的肝移植差异：我们已经实现目标了吗？#，Hepatology. 50 (2009) 981.[9] A.K. Mathur, D.E. Schaubel, Q. Gong, M.K. Guidinger, R.M. Merion，美国肝移植率中的性别差异，Am. J. Transplant. 11 (2011) 1435–1443.[10] D. Bertsimas, T. Papalexopoulos, N. Trichakis, Y. Wang, R. Hirose, P.A. Vagefi,

平衡肝移植获取机会中的效率与公平性：器官分配政策评估的权衡曲线，《移植》104（2020）981–987。[11] L.Lau、Y. Kankanige、B. Rubinstein、R. Jones、C. Christophi、V.Muralidharan、J. Bailey，机器学习算法预测肝移植后的移植物失败，《移植》101（2017）e125–e132。[12] C.-L. Liu、R.-S.Soong、W.-C. Lee、G.-W.Jiang、Y.-C. Lin，使用机器学习预测肝移植后的短期生存，《科学报告》10（2020）。[13] A. Ferrarese、G. Sartori、G. Orrù、A.C. Frigo、F.Pelizzaro、P.Burra、M.Senzolo，肝移植中的机器学习：解决某些未解决问题的工具？《国际移植》34（2021）398–411。[14] K.N. Vokinger、S. Feuerriegel、A.S.Kesselheim，减轻机器学习在医学中的偏差，《医学通讯》1（2021）25。[15] N.Mehrabi、F.Morstatter、N. Saxena、K.Lerman、A. Galstyan，机器学习中的偏差与公平性综述，《ACM 计算调查》54（2021）1–35。[16] F.P. Calmon、D. Wei、K.N. Ramamurthy、K.R.Varshney，优化数据预处理以防止歧视，ArXiv [Stat.ML]。（2017）。https://proceedings.neurips.cc/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html（访问日期：2022年12月12日）。[17] G. Pleiss、M.Raghavan、F.Wu、J.Kleinberg、K.Q. Weinberger，论公平性与校准，《神经信息处理系统进展》30（2017）。
https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html。[18] M. Wan、D. Zha、N. Liu、N. Zou，机器学习公平性的处理内建模技术：综述，《ACM知识发现与数据挖掘汇刊》17（2023）1–27。[19] S. Ding、R. Tang、D. Zha、N. Zou、K. Zhang、X. Jiang、X. Hu，公平地预测

肝移植中的移植物失败与器官分配，AMIA年会论文集 2022 (2022) 415–424.[20] C. Li, S. Ding, N. Zou,X. Hu, X. Jiang, K. Zhang，通过动态重加权实现医疗预测建模中的公平性的多任务学习，生物医学信息学杂志 143(2023) 104399.[21] B.Li, X.Shi,H. Gao, X. Jinag, K. Zhang,A.O. Harmanci,B. Malin，通过优化多个域对抗网络增强疾病预测中的公平性，BioRxiv。(2023)。https://doi.org/10.1101/2023.08.04.551906.[22] T. Yu, S. Kumar,A.Gupta,S. Levine,K.Hausman,C. Finn，面向多任务学习的梯度手术，神经信息处理系统进展 33 (2020)5824–5836.[23] P.T.D. Thinh, H.M. Hung, H.-J.Yang, S.-H. Kim,G.-S. Lee，使用改进的多任务学习技术进行不完整标签下的情感识别，ArXiv [Cs.CV]。 (2021)。http://arxiv.org/abs/2107.04192.[24] S. Lee, Y. Son，使用单次梯度步更新进行任务平衡的多任务学习，神经计算。467 (2022) 442–453.[25] A.Navon, A.Shamsian, I. Achituve, H.Maron, K. Kawaguchi,G.Chechik,E. Fetaya，多任务学习作为讨价还价博弈，ArXiv [Cs.LG]。 (2022)。http://arxiv.org/abs/2202.01017.[26] S. Ding, Q. Tan, C.-Y. Chang, N. Zou, K. Zhang,N.R. Hoot, X.Jiang,X.Hu，用于移植后死亡原因分析的多任务学习：肝移植案例研究，ArXiv [Cs.LG]。 (2023)。http://arxiv.org/abs/2304.00012.[27] D.Khurana,A. Koli,K. Khatter, S.Singh，自然语言处理：技术现状、当前趋势与挑战，多媒体工具与应用 82(2023) 3713– 3744.[28] X. Huang, A.Khetan, M. Cvitkovic,Z.Karnin, TabTransformer：使用上下文嵌入进行表格数据建模，ArXiv [Cs.LG]。(2020)。 http://arxiv.org/abs/2012.06678.[29] J.L.Ba,J.R.Kiros, G.E.Hinton，层归一化，ArXiv [Stat.ML]。 (2016)。

http://arxiv.org/abs/1607.06450.[30] T.-Y. Lin、P. Goyal、R. Girshick、K. He、P. Dollár, 《密集目标检测的焦点损失》，ArXiv [Cs.CV]。（2017）。http://arxiv.org/abs/1708.02002.[31] OPTN：器官获取与移植网络 - OPTN, （无日期）。http://optn.transplant.hrsa.gov（访问于2023年6月14日）。[32] MELD计算器 - OPTN, （无日期）。
h
tt
ps://optn.transplant.hrsa.gov/data/allocation-calculators/meld-calculator/（访问于2023年10月9日）。[33] M.J. Azur、E.A. Stuart、C. Frangakis、P.J. Leaf, 《链式方程多重插补：它是什么以及如何运作？》，《国际精神病学研究方法杂志》20（2011）40–49。

补充材料

表 S.1. 患者特征的详细描述和类别

特征	描述
ABO	登记时的受者血型
年龄_组	受者年龄组 A=成人 P=儿科
白蛋白_移植	移植时的受者血清白蛋白
AMIS	A位点错配级别
人工智能_LL_TCR	列入名单时的人工肝受体
腹水_治疗	移植时受者腹水
BACT_PERIT_TCR	登记时受体发生自发性细菌性腹膜炎
BMIS	B位点不匹配程度
CMV_IGM	移植时受体CMV IgM检测结果
CMV_状态	受体移植时的巨细胞病毒状态
肌酐_TX	移植时受体血清肌酐
等待天数_CHRON	肝脏等待名单上的天数
DGN_TCR	列入名单时的主要诊断
DGN2_TCR	列入名单时的次要诊断
DIAG	受体主要诊断
DIAL_TX	移植前一周透析?
DRMIS	DR位点不匹配程度
教育程度	受体登记时的最高教育程度
ENCEPH_TX	移植时脑病
END_体重指数_CALC	计算的移除时/当前时间候选体重指数
族裔类别	受体族裔类别
族裔	受体民族
EXC_肝细胞癌	相对于肝细胞癌的例外类型：Hbl、肝细胞癌、非肝细胞癌：Hbl=肝母细胞瘤

最终 _MELD_PELD_LAB_评分	最近一次等待名单上的 MELD/PELD 实验室评分，或移除时的评分（如已移除，可能与分配评分/Endstat 不同）
最终血清钠_ _	最近一次等待名单上的血清钠，或移除时的血清钠（如已移除）
FUNC_STAT_TCR	登记时受体的功能状态
性别	受者性别
乙型肝炎E抗体 旧版 _	移植时受者的乙型肝炎E抗体血清学检测（1999年10月25日前数据）
乙型肝炎病毒核心 抗体_	受体乙型肝炎核心抗体
乙型肝炎病毒核 酸检测_	移植乙型肝炎病毒核酸检 测结果
乙型肝炎病毒表面抗原_ _	受体乙型肝炎表面抗原
乙型肝炎病毒表面抗体总量 _ _	移植时受体的乙型肝炎病毒表面抗体总量
肝细胞癌诊断_	受体是否曾被诊断为肝细胞癌?
HCV血清学状态_	受体肝脏 C 状态
HGT CM CALC_ _	计算的受体身高（厘米）
人类免疫缺陷病毒血清状态_	移植时受体的人类免疫缺陷病毒血清学状态
HLAMIS	HLA不匹配程度
INR TX_	移植时受体国际标准化比值 (INR)
LIFE SUP TRR_ _	移植时的移植前受体生命支持
列出终末期肝病模 型_	患者是否在 MELD/PELD 实施前列入名单?
医疗状况 TRR_ _	受者移植前医疗状况
门静脉 TCR_ _	受者登记时的门静脉血栓形成史
门静脉 TRR_ _	受者移植时的门静脉血栓形成
既往腹部手术 TCR_ _ _	受者登记前既往上腹部手术
受者既往上腹部手术_ _ _	受者移植前既往上腹部手术
受者既往任何器官移植 _ _	计算得出的任何器官类型既往移植
PRVTXDIF	受者既往移植与当前移植之间的天数
地区	Wl UNOS/OPTN 地区（列入名单/接受移植）
总胆红素 TX_	移植时受体总胆红素

经颈静脉肝内门体分流术	_TCR	受体经颈静脉肝内门体分流术 登记时的分流术（经颈静脉肝内门体分流术）
呼吸机	_TCR	登记时依靠生命支持的受体——呼吸机（1=是，0=否）
WORK_收入	_TCR	登记时是否有收入来源?

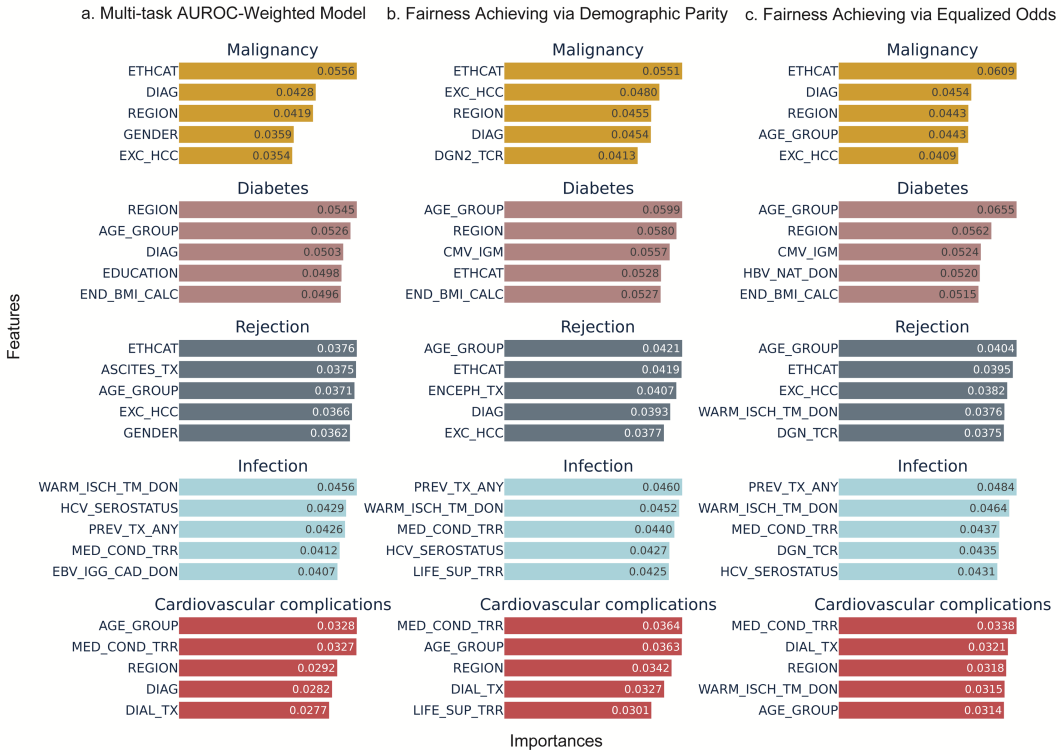
表 S.2. 供者特征的详细描述与分类

特征	描述
ABO_DON	供者血型
ABO_MAT	供者与受者 ABO 匹配程度
AGE_DON	供者年龄（岁）
酒精 _大量 _DON	Ddr 重度饮酒情况（是/否/未知）
抗高血压药物_DON	已故供者在阻断前24小时内使用抗高血压药物阻断
精氨酸 _DON	已故供体-是否给予供者精氨酸加压素交叉阻断前24小时内?
血液 _INF_DON	已故供体血液作为感染源
BUN_DON	已故供体终末期血尿素氮
心脏骤停 _神经系统	脑死亡后心脏骤停的已故供体
CDC_RISK_HIV_DON	已故供体——按公共卫生服务标准，供者是否存在风险传播血源性疾病的因素?
CMV_DON	供者血清学抗CMV（针对活体供者，UNet前数据仅)
COD_CAD_DON	已故供体-死亡原因
持续使用_CIG_DON	已故供体-既往香烟史：>20包年+近6个月使用
持续使用_可卡因 _DON	已故供体-可卡因使用病史+近6个月使用
持续使用_OTH_DRUG_DON	已故供体-既往其他药物史+近期6个月使用
受控 _DON	受控非心脏停搏逝世供体（是/否）
肌酐 _DON	逝世供体终末实验室肌酐
去氨加压素 _DON	逝世供体合成抗利尿激素（去氨加压素）
死亡 _CIRCUM _DON	逝世供体死亡情况
死亡 _MECH_DON	已故供体-死亡机制
DIABDUR _DON	已故供体-糖尿病持续时间
糖尿病 _DON	已故供体-糖尿病病史（是，否）

DIET_DON	高血压经饮食控制的逝世供体
供体标准最大年龄_ _	供体标准 - 可接受的供体最大年龄（周岁） - 本地
DONCRIT 最大里程_ _	供体标准 - 最大行驶英里数 器官或器官获取团队将行驶
供体标准最大体重_ _	供体标准 - 可接受的供体最大体重（千克） - 本地
DONCRIT MIN WGT_ _	供体标准 - 可接受的供体最小体重（千克） - 本地
EBV_IGG_CAD_DON	已故供体 Epstein-Barr 病毒免疫球蛋白G检测结果
EBV_IGM_CAD_DON	已故供体 Epstein-Barr 病毒 IgM 检测结果
ECD_供者	扩展的供者每肾分配定义 1=是
HBSAB _DON	已故供体乙肝表面抗体检测结果
HBV_CORE_DON	供者乙肝核心抗体
HBV_NAT_DON	供者乙肝核酸检测结果:
红细胞压积 _DON	供者红细胞压积:
HEP_C_ANTI_DON	已故供体-丙型肝炎病毒抗体检测结果
HGT_CM_DON_CALC	计算所得供者身高（厘米）
HIST_癌症 _DON	已故供体-癌症病史（是/否）
HIST_CIG_DON	已故供体-过去吸烟史：>20包年
病史 可卡因 供体_ _	已故供体既往可卡因使用史
病史 高血压 供体_ _	已故供体-高血压病史
病史 胰岛素 依赖 供体_ _ _	已故供体-胰岛素依赖性糖尿病 (Y,N)
供者其他药物病史_ _ _	已故供体——既往使用其他药物史
供者心肌梗死病史_ _	DDR既往心肌梗死史
供者胰岛素_	已故供体——阻断前24小时内是否给供者使用过胰岛素?
LI_活检	遗体供体肝脏活检
宏观 _FAT_LI_DON	已故供体宏观脂肪%

微量 _FAT_LI_DON	已故供者微量脂肪百分比
NON_HRT_DON	已故非心跳供者
PH_DON	供者：血液酸碱度
蛋白质 –尿液	已故供者尿液中的蛋白质
PT_利尿剂 _DON	已故供体-利尿剂 B/N 脑死亡 W/In 24 小时内 采购
PT_类固醇 _DON	已故供体-类固醇：脑死亡后 24 小时内 采购
PT_T3_DON	已故供体-三碘甲状腺原氨酸-三碘甲状腺原氨酸（T3） B/N 脑死亡 W/In 器官获取后24小时
PT_T4_DON	已故供体-甲状腺素-T4 B/N 脑死亡 W/In 24 小时 采购相关的
PULM_INF_DON	已故供体-感染肺部来源
复苏 _DUR	已故供体心搏骤停后复苏 脑死亡
SGOT_DON	逝者供者终末期谷草转氨酶
SGPT_DON	已故供体终末期谷丙转氨酶
SKIN_癌症 _DON	已故供体获取时皮肤癌 (是/否)
总胆红素 _DON	已故供体终末期总胆红素
输血 _TERM_DON	Ddr：此期间（终末期）的输血次数 住院：
尿液 _INF_DON	已故供体感染尿液来源
血管扩张剂 _DON	已故供体—交叉阻断前 24 小时内使用血管扩张剂
VDRL_DON	已故供体—RPR-VDRL 检测结果
WARM_ISCH_TM_DON	已故供体—非心跳供体估计温暖时间 缺血时间（分钟）

图 S.1。 不同模型下肝移植后风险因素前 5 项特征重要性的比较分析。通过置换特征重要性确定的每个肝移植后风险因素前 5 项特征重要性，呈现在以下三种不同设置中：a) 未采用实现公平性的方法的多任务 AUROC 加权。b) 采用 AUROC 加权任务平衡方法和实现公平性的方法以促进人口统计均等（DP）的模型，具体旨在减轻性别偏差。c) 采用 AUROC 加权任务平衡方法和实现公平性的方法以实现均等化赔率的模型，该模型同样关注不同性别之间的偏差。风险因素（从上到下）依次为恶性肿瘤、糖尿病、排斥反应、感染和心血管并发症。子图 a、b 和 c 中的每个图分别代表上述一种设置。



图S.2。不同模型下肝移植后风险因素前5项特征重要性的比较分析。通过置换特征重要性确定的每个肝移植后风险因素前5项特征重要性，展示于三种不同设置中：a) 未采用公平性实现方法的多任务 AUROC 加权。b) 采用 AUROC 加权任务平衡方法和实现人口统计均等（DP）公平性的方法的模型，旨在减轻种族/族裔偏差。c) 采用 AUROC 加权任务平衡方法和针对均等化赔率的公平性实现方法的模型，该模型同样关注种族/族裔之间的偏差。风险因素（从上到下）依次为恶性肿瘤、糖尿病、排斥反应、感染和心血管并发症。子图 a、b 和 c 中的每个图分别代表上述设置中的一种。

