

· 综述 ·

迁移学习在罕见病研究中的应用与展望

郑雪莹¹, 秦国友¹, 余勇夫^{1, 2}

1. 复旦大学公共卫生学院生物统计教研室 (上海 200032)
2. 国家卫生健康委员会卫生技术评估重点实验室 (复旦大学) (上海 200032)

【摘要】 迁移学习是一种利用源数据的现有知识对相关领域的新任务进行学习的方法。罕见病研究中数据通常较为稀缺, 迁移学习可以有效利用其他相关疾病或领域的数据来提高模型性能和研究效率, 因此有助于研究人员快速识别罕见病的疾病特征和开发潜在治疗方案。目前, 迁移学习在罕见病的系统性特征描述和药物开发等方面取得了显著进展, 且在优化罕见病分类、加速早期诊断以及支持多任务研究等方面展现出潜力。然而, 迁移学习在罕见病研究中的应用也面临诸多挑战。未来, 若能将迁移学习与强化学习、联邦学习、深度学习等多种技术相结合, 有望在罕见病研究领域实现更大的突破。

【关键词】 罕见病; 稀缺数据; 迁移学习

【中图分类号】 TP 181; R 195.1 **【文献标识码】** A

Applications and prospects of transfer learning in rare diseases research

ZHENG Xueying¹, QIN Guoyou¹, YU Yongfu^{1,2}

1. Department of Biostatistics, School of Public Health, Fudan University, Shanghai 200032, China
 2. NHC Key Laboratory of Health Technology Assessment (Fudan University), Shanghai 200032, China
- Corresponding author: YU Yongfu, Email: yu@fudan.edu.cn

【Abstract】 Transfer learning is a method of learning new tasks in related domain using existing knowledge from source data. In rare disease research, data are often limited. Transfer learning can effectively use data from other related diseases or fields to enhance model performance and research efficiency. This approach helps researchers rapidly identify characteristics and develop potential treatments of rare disease. Currently, transfer learning has been applied in the systematic characterization and drug development of rare diseases. It also shows potential in optimizing rare disease classification, accelerating early diagnosis, and supporting multi-task research. However, challenges arise in the application of transfer learning in rare disease research. In the future, if transfer learning can be combined with techniques such as reinforcement learning, federated learning, and deep learning, greater breakthroughs are expected to be achieved in the field of rare diseases.

【Keywords】 Rare diseases; Scarce data; Transfer learning

罕见病已成为全球健康领域的一大挑战。尽管罕见病患者人数相对较少, 但其种类繁多, 总体数量显著, 影响着全球至少 3.5%~5.9% 的人口^[1]。由于每种疾病的患病率极低, 相关领域的

诊断与治疗专业知识非常有限, 研究者对这些疾病的认知不足, 因而相关研究受到诸多限制。事实上, 90% 以上的罕见病尚无获得监管机构批准的治疗方法^[2], 超过一半研究罕见病的随机对照

DOI: 10.12173/j.issn.1005-0698.202412136

通信作者: 余勇夫, 博士, 教授, 博士研究生导师, Email: yu@fudan.edu.cn

<https://ywlbxb.whuznhmedj.com/>

试验未完成，或在试验完成后 4 年中未发表^[3]，这进一步加剧了罕见病领域的临床研究困境。

迁移学习（transfer learning）是一种通过从源领域学习知识来增强目标领域应用的方法，在目标领域训练数据有限的情况下表现出色^[4]。在罕见病研究中，由于样本数据的获取较为困难，迁移学习展现了极大的应用前景。通过在大型公共医疗数据集上预训练模型，并进行参数微调（fine-tuning）或特征迁移，迁移学习能够在罕见病数据集上的模型性能（图 1-A），有效减轻对大规模训练数据的依赖。源数据和目标任务之间的相关性越强，模型的预测效果就越优异。在罕见病研究中应用迁移学习不仅减少了大规模训练集的收集需求和相应工作量，也有望提升罕见病研究和治疗开发的效率。

目前，国内外文献中针对迁移学习在罕见病这一特定领域的研究总结尚不多见。本文首先整理了迁移学习在罕见病研究中的应用实例，介绍几种相关的方法和技术；随后讨论了迁移学习方法在罕见病领域可行的应用场景；最后总结了应用迁移学习方法可能面临的挑战，为进一步开展相关研究提供了参考依据。

1 迁移学习在罕见病相关医学和药学研究中的应用现状

医疗领域广泛采用人工智能技术，利用不同的数据源，如医学图像和电子健康记录来辅助罕见疾病的检测。机器学习（machine learning）方法应用于识别罕见疾病患者、药物发现或重新利用、疾病的分子分型和罕见病患者预后预测等方面已有一段时间^[5]。然而在样本数量非常有限的任务中，仅使用传统的学习技术而缺乏高级学习策略或特定模型设计，可能会导致严重的过拟合。这种过拟合现象会显著降低模型的泛化和预测能力^[6]。

迁移学习因其在数据稀缺或样本数量较少的情况下表现出的优势，被用于研究少数民族的医疗保健不平等问题^[7]和改善数据劣势群体的预测准确性^[8]。近年来，迁移学习在揭示罕见病的系统特征^[9-10]、药物发现^[11]、药代动力学参数预测^[12]和药物敏感性预测^[13]等方面取得了积极进展（表 1）。这些进展得益于多种迁移学习方法的灵活应用，而迁移学习与其他方法的结合进一步提升了研究的效果和准确性。

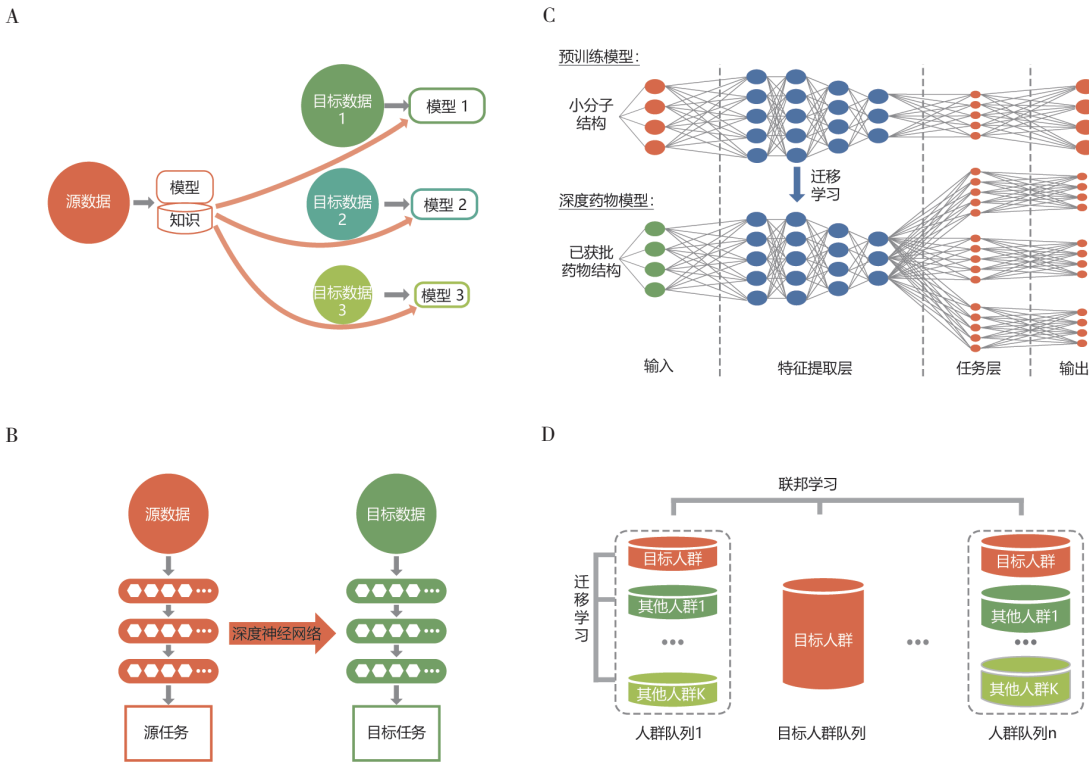


图1 几种迁移学习方法的示意图

Figure 1. Diagram of various transfer learning methods

注：A. 迁移学习的基本思想；B. 深度迁移学习；C. 迁移学习结合多任务学习；D. 联邦迁移学习。

表1 迁移学习在罕见病相关医学和药学领域的应用
Table 1. Application of transfer learning in medicine and pharmacy related to rare diseases

研究工作	适用数据类型	数据集		任务	主要技术和方法
		源域	目标域		
利用转录组学的迁移学习揭示罕见疾病的系统性特征 ^[9]	转录组数据 (bulk RNA 测序/微阵列)	recount2 多源整合 RNA 测序数据库	ANCA 相关性血管炎相关数据	提取特征分类	在包含多个实验、组织和生物条件的大型公共数据汇编上训练了 PLIER 模型, 然后将该模型用 MultiPLIER 方法迁移到小数据集上
改进对胶质母细胞瘤状态和亚型描述 ^[10]	多组学数据	TCGA	TCGA 抽样形成的小型样本	识别癌症亚型	通过迁移学习进行多组学矩阵分解, 推断小型多组学数据集的潜在因素
药代动力学参数预测 ^[12]	分子指纹 (ECFPs)	DeepChem 生物活性数据集	1 104 种美国 FDA 批准的小分子药物的药代动力学数据集	预测药代动力学参数	采用迁移学习和多任务学习结合改进的数据集分割算法进行药代动力学参数预测, 提高模型的泛化能力
迁移学习方法改善多发性骨髓瘤患者药物敏感性预测 ^[13]	转录组数据 (bulk RNA 测序/微阵列)	总计 1 237 例癌症患者的转录组数据和药物的 IC ₅₀ 值	280 个多发性骨髓瘤患者的转录组数据和药物的 IC ₅₀ 值	预测药物敏感性	通过结合转移学习、特征选择和增强学习策略, 利用来自不同癌症类型的辅助数据提高多发性骨髓瘤患者药物敏感性预测准确性
用深度迁移学习减少生物医学数据不公平 ^[7]	转录组数据 (bulk RNA 测序/微阵列)	TCGA 中的欧洲裔美国人	TCGA 中少数民族裔、the MMRF CoMMpass study	预测疾病结局	通过机器学习模拟揭示种族间模型性能差异, 并通过迁移学习提升劣势族裔模型性能
深度迁移学习为改进机器学习公平性 ^[8]	基因组数据 (GWAS)	dbGaP 数据库中的欧洲人群	dbGaP 中的数据不足人群	预测疾病患病	采用迁移学习策略来改善数据劣势人群疾病发生的临床基因组预测
非传染性疾病的迁移学习方法 ^[14]	表观遗传学数据 (DNA 甲基化)	TCGA	GEO 数据库中的关节炎、哮喘和精神分裂症数据	疾病分类	利用在一类非传染性疾病 (癌症) 上训练的神经网络, 作为非癌症非传染性疾病建模的迁移学习方法的基础

注: ANCA. 抗中性粒细胞胞质抗体 (antineutrophil cytoplasmic antibody); PLIER. 通路水平信息提取器 (pathway-level information extractor); TCGA. 癌症基因组图谱 (the cancer genome atlas); ECFPs. 扩展连接性指纹 (extended-connectivity fingerprints); FDA. 美国食品药品监督管理局 (U.S. Food and Drug Administration); IC₅₀. 半数最大抑制浓度 (half maximal inhibitory concentration); the MMRF CoMMpass study. 多发性骨髓瘤研究基金会将多发性骨髓瘤的临床结果与个人基因特征评估相关联的研究 [Multiple Myeloma Research Foundation (MMRF) Relating Clinical Outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile Study]; GWAS. 全基因组关联研究 (genome-wide association study); dbGaP. 基因型与表型数据库 (The Database of Genotypes and Phenotypes); GEO. 基因表达综合数据库 (Gene Expression Omnibus)。

1.1 揭示罕见疾病的系统性特征

在罕见病研究领域, 有监督迁移学习通过利用类似疾病的大规模数据集 (带标签) 来提高小规模罕见疾病数据 (可能带有部分标签) 的预测精度。识别罕见疾病新亚型是揭示其系统性特征的关键步骤, 而无监督学习作为常见的数据驱动亚型识别方法, 在罕见病领域却因研究资料匮乏而受限。无监督迁移学习通过整合分子与遗传数据, 为罕见疾病新亚型的识别提供了有效途径。迁移学习方法 MultiPLIER 被用于分析大量基因组数据, 以识别与罕见病相关的生物标记物^[9]。抗中性粒细胞胞浆抗体相关性血管炎是一种罕见的全身性自身免疫病, 影响多种组织。然而, 现有的基因表达检测数据不足以支持无监督机器学习, 且其潜在的分子机制尚不完全明了。为实现有效的知识迁移, 研究假设转录调控、基因表达

数据的模块化性质以及共享的分子机制可以支持无监督迁移学习。借助系统性红斑狼疮全血训练数据集的样本, MultiPLIER 生成了更能捕获生物通路和区分相似分子过程的模型, 识别出两组与疾病活动性和严重程度相关的基因表达模式, 并在此基础上进一步开展差异基因分析和富集分析, 这表明迁移学习在该领域的潜力。此外, 对两个髓母细胞瘤队列使用相同的 MultiPLIER 模型研究能够揭示不同亚型间一致的通路差异, 证明了该方法在不同类型的罕见疾病中的适用性。MultiPLIER 提取的特征具有生物学关联性, 并允许跨不同数据集进行直接比较, 使其在复杂人类罕见疾病的综合分析中表现出价值。

此外, 基于迁移学习的多组学矩阵分解方法 (multi-omics transfer learning, MOTL) 增强了在胶质母细胞瘤小样本集上对癌症状态和亚型的区

分能力^[10]。该方法通过结合从癌症基因组图谱公开的数据集中推断出的潜在因子，对目标数据集进行分解。相较于不使用迁移学习的直接矩阵分解，MOTL 能更有效地识别差异活跃的真实因子并进行基因集富集分析。这些因子反映了癌组织的不同胚源，并强调了免疫和微环境在癌症病理生理及其对治疗反应中的重要性。

这两项研究展示了迁移学习在罕见病与癌症研究中的重要应用。它们不仅提高了生物标记物的识别效果，还增强了对疾病相关生物过程的理解，为针对罕见病和罕见肿瘤的个性化治疗方案提供了坚实的基础。

1.2 药物开发与精准医学

1.2.1 药物发现

迁移学习在 2010 年代被首次应用于药物发现，而随着 2020 年深度迁移学习（deep transfer learning, DTL）方法的提出和应用，该方法的应用开始引起广泛关注。迁移学习在药物发现中的主要应用涵盖三个关键阶段：分子性质和活性预测、分子生成和基于结构的虚拟筛选^[11]。DTL 是将深度学习架构与迁移学习相结合的一种方法（图 1-B），其中最常见的技术之一是微调，这是一种基于参数的迁移学习方法，即将利用深度神经网络中神经元之间连接的学习参数（权重）迁移到目标模型中以更好地解决相关的目标问题。微调使得源模型中的有用特征得以在新任务中重用，从而提高目标模型的性能。特别是在相关性较高的任务间，这种方法能够显著加速训练的收敛来降低计算成本并提升模型表现。

标记数据的稀缺性是人工智能辅助药物发现的主要障碍之一。在药物发现领域，迁移学习可以通过整合来自不同来源（如其他疾病领域）的数据，加速新药的筛选和优化，这使研究者能够更迅速地识别潜在的治疗靶点和候选药物^[15]。

1.2.2 药代动力学参数预测

药代动力学评价是药物研发中的关键环节之一，但现有的预测模型预测 4 个主要人体药代动力学参数的准确性仍有限。基于迁移学习和多任务学习的药代动力学参数预测方法 DeepPharm（图 1-C）增强了模型对稀缺训练数据集的泛化能力^[12]。该方法也是一种基于参数的迁移学习方法，其多任务学习模型采用了具有 10 个密集层的前馈神经网络，而迁移学习模型利用预训练模

型的权重作为初始模型权重，并在包含 1 104 种已批准的小分子药物的药代动力学数据集上进行微调。与传统方法相比，DeepPharm 显著提升了模型的预测能力，提供了更为可靠的药物体内动态过程估计。

最近的一项综述汇总了用于药代动力学预测的迁移学习方法^[16]。多项研究开发了基于迁移学习的模型来预测分子的药代动力学性质，包括同构迁移学习（从一种分子表示中迁移知识用于不同的预测任务）、同域异构迁移学习（从不同的分子表示中迁移知识用于单一预测任务），以及跨域异构迁移学习（从不同领域中迁移知识）。基于迁移学习技术的预测药代动力学模型已成为药物设计的重要工具。

1.2.3 药物敏感性预测

在药物敏感性预测方面，迁移学习被应用于预测患者对不同药物的反应，通过利用模型学习现有的治疗反应数据，为临床医生提供个性化治疗建议。在多发性骨髓瘤患者的药物敏感性预测方面，一项研究基于迁移学习方法结合改进的 AdaBoost 算法进一步提高模型性能^[13]。该方法是一种基于实例的迁移学习方法，即从源域中挑选出对目标任务训练有用的实例，从而建立分类精度较高的、可靠的模型。其方法包括以下步骤：首先从目标训练集中筛选出最重要的基因，并基于这些基因构建新的数据集；随后使用改进的 AdaBoost 算法调整样本权重，经过多次迭代来优化模型性能。增强的迁移学习技术取得良好效果的生物学背景在于，癌症治疗过程中常用的联合药物治疗法会提升治疗效果。例如，多西他赛作为一种化疗药物，通常与其他特异性化疗药物联合用于乳腺癌治疗，而硼替佐米与多西他赛联合用于乳腺癌也显示出有效性。因此，预测多发性骨髓瘤患者对硼替佐米的敏感性与预测乳腺癌患者对多西他赛的敏感性有密切关系^[13]。

2 迁移学习在罕见病研究中的应用前景

2.1 优化罕见病的分类

精准医疗的一个目标是对患者进行分类和分层，以提高诊断和医疗水平^[17]。许多罕见病的可用数据量仍然不足，无法依靠现有数据训练神经网络。然而关于非传染性疾病群体，例如癌症患

者,存在大量组学数据和其他相关数据,这为应用迁移学习生成缺乏足够数据的罕见病神经网络模型提供了可能性。一项研究^[14]基于非传染性疾病(癌症)的DNA甲基化数据集,探索了一种名为DiseaseNet的迁移学习方法,用于非癌症非传染性疾病的建模。DiseaseNet在预测关节炎、哮喘和精神分裂症三种非传染性疾病方面表现优异,不仅在分类性能上优于其他模型,还揭示了癌症与非癌症非传染性疾病在表观遗传水平上的共享特征。通过采用类似的方法,可以进一步挖掘非传染性疾病与罕见病在表观遗传水平上的共享特征,从而提高预测效果。

随着更多常见癌症的分子亚群被发现,根据流行病学的定义,这些癌症类型将成为罕见肿瘤,因而罕见癌症患者的数量也越来越多^[18]。通过类似跨疾病的知识迁移,罕见病的研究有望实现更稳健的疾病分类。

2.2 加速罕见病的早期诊断

罕见病的诊断面临很大挑战,尤其是在早期阶段。在实际情况中,数据持有者往往仅拥有极为有限的与罕见疾病相关的患者数据。罕见病数据的分布可能受到人口统计和地理信息等因素的影响。针对这种情况,联邦学习(federated learning)方法允许研究人员仅共享模型参数更新,从而在保护患者隐私的同时,根据来自大量罕见疾病患者的数据开发更精准的模型^[19]。例如,在一项针对罕见的胶质母细胞瘤的研究^[20]中,联邦学习汇集来自六大洲71个地点的数据(样本量6 314例)。该模型通过共享来自分散数据集的模型参数进行训练,降低了对数据共享的需求,并增加了地理上不同合作者的参与,从而扩大了用于训练罕见病模型的数据规模和多样性。另一项研究^[21]则开发了动态联邦元学习方法,以改进罕见病的预测。

然而针对罕见病诊断和检测的联邦学习研究仍然有限。联邦迁移学习(federated transfer learning)^[22](图1-D)结合了迁移学习和联邦学习的特点,能够在多个参与方(如人群队列)之间实现模型和知识的迁移,而无需直接共享参与者的数据。通过联邦迁移学习,可以有效整合来自不同来源的罕见病患者数据,包括基因组学数据、影像学数据、电子健康记录和患者报告结局。这种方法不仅保护了数据隐私,同时借助跨疾病

的知识迁移,更有望加速罕见病的早期诊断。

2.3 多任务和多领域的罕见病研究

假设每个数据持有者和分析者都能够掌握所有形式的罕见病患者数据并不现实,因此,开发能够有效处理不同机构之间多任务场景下的罕见病数据的迁移学习框架显得尤为重要。多任务学习旨在同时解决多个任务,并学习出一个对所有研究任务表现良好的共享模型^[12]。这种方法可以利用训练集中相似结构分子的特征信息,并将其应用于其他任务;通过对多任务神经网络的共享层进行训练,可以更好地获得用于预测任务的共享特征表示。由于罕见病的分子分型通常表现出重叠性和异质性,多任务迁移学习有望推动对罕见疾病的深入理解与精确治疗,助力罕见病的早期诊断与个性化治疗。

3 迁移学习在罕见病研究中的挑战

3.1 量化不同任务之间的相关性

迁移学习的性能往往依赖于源任务和目标任务的相关性,任务的相关性通常比数据的大小更为重要。在源域(如普通疾病)与目标域(如罕见病)之间,可能存在显著的特征差异,这对模型的迁移学习能力构成了挑战。例如,药代动力学预测涉及复杂的生物过程和相互作用,在保持这些模型的完整性的同时结合迁移学习需要仔细考虑源数据和目标数据的一致性。

将分子生物学和化学知识量化成为任务之间的相关性对于药物发现中的迁移学习至关重要。在任务相关性难以明确界定的情况下,基于特征的方法可以帮助在潜在空间中发现不同域之间的相似性,从而提升模型的泛化性能。此外,有时单一源域可能不足以有效改进某种罕见病的目标任务。因此,从多个具有良好可转移性的源域进行迁移学习可以带来显著的优势。这种多源域迁移策略已经被证明在减少负转移的风险上具有一定效果,能够帮助模型更全面地学习潜在特征并提高模型在目标任务上的表现^[15]。

3.2 构建指标评估迁移学习的性能

由于迁移学习场景中的数据规模较小,过拟合问题需要特别关注。在罕见疾病的数据集中,简单依赖准确性作为性能衡量标准可能并不理想。如果模型将所有样本都识别为非罕见病样本,它仍然可能获得较高的准确率,但这并不能反映

其实际识别罕见病样本的能力。因此,构建处理类不平衡的评价指标显得尤为重要,例如加权损失函数^[5]。在模型训练阶段,可以采用加权损失函数,使模型在优化过程中对罕见病样本给予更高的关注,从而改善模型对少数类样本的识别能力。通过引入这些更加精准的度量标准,可以有效提升罕见疾病模型的性能评估,确保其在实际应用中具备更好的识别能力。

3.3 处理不准确的罕见病标签

罕见病的标签可能不够精确或存在不一致性,这会严重影响模型的学习效果^[23]。因此,需要与临床专家合作共同注释数据从而生成更加稳健的数据集。但由于医生缺乏对这些疾病的全面了解,导致罕见病患者的误诊很常见^[24]。

在此基础上,考虑将迁移学习与集成学习(ensemble learning)相结合,以提高模型对不一致标签的稳健性。集成学习是指通过综合多个基础模型的预测结果,以提高整体预测的准确性和稳健性。由于不同的模型在面对标签不一致时可能会做出不同的预测,集成方法能够有效降低错误标签对最终决策的影响,从而提升预测的准确性与可靠性。在这种背景下,将迁移学习与集成学习相结合,可以进一步增强模型在面对不一致标签时的稳健性,通过迁移已有知识并结合多个模型的力量,提升整体性能。此外,还可以考虑将强化学习(reinforcement learning)^[25]与迁移学习相结合。强化学习通过与环境的交互来学习制定策略,旨在动态决策问题中最大化累积的奖励信号。将强化学习与迁移学习相结合,可以通过构建适当的奖励机制,有助于强化学习算法更快地理解目标任务的状态空间,使模型在面对不精确标签时,可以更好地关注稳健的特征学习和决策制定。在复杂的决策环境中(例如罕见病标签不准确),迁移学习结合强化学习通过试错的方法逐步改善模型的性能,从而应对标签不准确的挑战。

3.4 构建支持迁移学习的罕见疾病数据集

组学分析会产生超高维数据集。然而,典型的罕见病数据集通常样本数量有限,导致特征空间的维度远大于样本量,这为构建高度可泛化的模型增添了难度^[26]。针对罕见病数据缺乏的问题,一种有效策略是整合不同类型的数据,并结合现有知识深入挖掘。在合并数据集时,需要考虑特

定于各数据集的差异,并采用适当策略将原本较小的数据集扩充整合成更大的复合数据集。因此,高效的数据整合方法在迁移学习应用于罕见病研究中尤为重要。

3.5 探索罕见病的发病机制

目前罕见疾病研究中,迁移学习方法用于探索疾病生物机制的研究尚显不足。例如,开发具有模型可解释性的方法可以识别与潜在疾病机制相关的重要特征。对新开发模型进行稳健的误差分析可以帮助研究者理解哪些特征在多大程度上影响模型性能。这种分析不仅提升了解释能力,也有助于深入探讨某些特征对潜在疾病机制的贡献。

4 结语

本文总结了罕见疾病研究中采用的迁移学习及其相关方法,概述了当前利用迁移学习进行罕见疾病研究的前沿进展,并探讨了未来研究的方向。迁移学习相关技术的应用能够有效地从相关大型源数据中提取知识,为罕见病的研究提供了一种有效的解决方案。期待越来越多的研究人员投入这一重要的研究领域,以推动罕见病的科学理解 and 治疗方法的发展。

利益冲突声明: 作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

- 1 周奇,李沁原,刘雅莉,等.罕见病指南的制订:现状、挑战与机遇[J].协和医学杂志,2023,14(3):621-628.[Zhou Q, Li QY, Liu YL, et al. The development of guidelines for rare diseases: past, present and future[J]. Medical Journal of Peking Union Medical College Hospital, 2023, 14(3): 621-628.] DOI: 10.12290/xhyxzz.2022-0360.
- 2 Denton N, Mulberg AE, Molloy M, et al. Sharing is caring: a call for a new era of rare disease research and development[J]. Orphanet J Rare Dis, 2022, 17(1): 389. DOI: 10.1186/s13023-022-02529-w.
- 3 Rees CA, Pica N, Monuteaux MC, et al. Noncompletion and nonpublication of trials studying rare diseases: a cross-sectional analysis[J]. PLoS Med, 2019, 16(11): e1002966. DOI: 10.1371/journal.pmed.1002966.
- 4 潘璐璐,余勇夫,秦国友.迁移学习简介及其在医学研究领域中的应用[J].复旦学报(医学版),2024,51(6):1016-1020.[Pan LL, Yu YF, Qin GY. Introduction and application of transfer learning in medical research[J]. Fudan University Journal

