·肿瘤研究与诊疗前沿交叉技术·
DOI:10.16262/j.cnki.1000-8217.20250226.007

肿瘤大数据与人工智能在肿瘤诊疗中的应用现状、挑战与未来展望*

肖文铉† 赵 珅† 江一舟**

复旦大学附属肿瘤医院乳腺外科,复旦大学上海医学院肿瘤学系,上海 200032

[摘 要] 肿瘤是全球范围内重大公共卫生问题,尽管现代医学取得了一定进展,但在精准诊疗、风险评估、药物研发等方面仍存在许多未被满足的临床需求。随着大数据技术的发展,肿瘤大数据逐渐成为推动肿瘤领域创新的重要力量。本综述通过描述肿瘤大数据的特征,包括多组学数据(如基因组、转录组、蛋白组、代谢组等)和多模态数据(如临床、影像、病理等),探讨肿瘤大数据与人工智能结合在肿瘤诊疗中的应用现状。同时,分析现阶段面临的主要挑战,如数据标准化程度不足、多模态数据融合的技术瓶颈、模型可解释性有限以及临床验证的缺乏等问题,并初步探讨可能的解决方向。最后,展望未来发展趋势,提出在肿瘤诊疗中构建专用大模型、实现多维数据融合以及设计个体化需求驱动的精准管理策略的重要性。

[关键词] 肿瘤诊疗;肿瘤大数据;人工智能;多组学数据;多模态数据;精准医疗

1 肿瘤诊疗现状

1.1 肿瘤是严重危害人群健康的复杂疾病

肿瘤作为全球主要公共卫生问题之一,其发病率和死亡率仍在持续上升,世界卫生组织报道指出,2022 年全球新增癌症病例近 2 000 万例,死亡病例 970 万例[1]。中国国家癌症中心发布 2024 年全国癌症报告称中国肿瘤负担仍在一直增加,严重影响中国居民的健康、国民经济和社会发展[2]。

肿瘤是一个复杂多变的疾病类型。不同肿瘤在生物学特征、发生发展机制等方面存在显著差异[3]。这种显著的时空异质性,使得肿瘤治疗颇具挑战。近 15% 的患者仍面临治疗耐药及复发转移的挑战[4],现有传统治疗手段仍需更加精确、有的放矢指导患者精准治疗。

1.2 肿瘤诊疗中存在大量未被满足的关键临床需求

目前在肿瘤的早期精准诊断、精准治疗、风险评估及药物研发等四个方面均存在亟待解决的关键临床需求。

1.2.1 精准诊断

当前,许多肿瘤诊断仍依赖于传统影像学检查和组织病理学分析,这些方法往往存在耗时长、精度低等问题,且通常无法全面反映肿瘤的分子特征等信息。未来应通过技术手段确保在肿瘤早期阶段就能准确识别不同类型肿瘤及其分子特征,实现精准诊断。

1.2.2 精准治疗

目前肿瘤传统治疗仍多为根据临床病理信息实

收稿日期:2024-11-25;修回日期:2025-02-24
† 共同第一作者。
* 本文根据国家自然科学基金委员会第 373 期“双清论坛”讨论的内容整理。
** 通信作者,Email: yizhoujiang@fudan.edu.cn
本文受到国家自然科学基金项目(82272822)的资助。

引用格式:肖文铉, 赵珅, 江一舟. 肿瘤大数据与人工智能在肿瘤诊疗中的应用现状、挑战与未来展望. 中国科学基金, 2025, 39(1): 153—161.
Xiao WX, Zhao S, Jiang YZ. Current applications, challenges, and future prospects of tumor big data and artificial intelligence in tumor diagnosis and treatment. Bulletin of National Natural Science Foundation of China, 2025, 39(1): 153-161. (in Chinese)

行放化疗，无法快速、准确鉴别治疗靶点，导致部分患者疗效不佳甚至无效。精准治疗的核心是依据患者临床、分子特征等，对患者进行生物标记物的分析与鉴定，精确寻找治疗靶点予以治疗。精准治疗不仅能显著提高疗效，还能减少不必要的副作用。

1.2.3 风险评估

肿瘤复发转移是治疗中的重要挑战。精准的风险评估有助于医生在治疗过程中实时监测患者的病情，预测复发风险，并做出相应调整。现有的风险评估体系多基于临床病理特征，但由于肿瘤的高度异质性，仍无法充分反映患者的真实风险，缺乏准确可靠的模型评估患者的疗效和预后。

1.2.4 药物研发

肿瘤药物研发面临巨大挑战，尤其是在应对肿瘤耐药和复发转移方面。一方面目前药物研发多基于实验室传统筛选药靶和可能具有药理活性的化合物，研发花费大、耗时长。另一方面市场上大部分肿瘤药物均针对常见靶点，如表皮生长因子受体（Epidermal Growth Factor Receptor，EGFR）等，但其药效常受肿瘤异质性制约。

2 肿瘤大数据

肿瘤研究的快速发展和技术进步导致了海量的多组学和多模态数据的产生，这些数据涵盖了基因组学、转录组学、影像资料和病理图像等多个维度[5]。如何高效整合分析这些庞大且多维的数据，仍是肿瘤研究中的一大挑战。

2.1 多组学数据

2.1.1 基因组

肿瘤研究中，基因组学主要关注肿瘤组织中基因突变、拷贝数变异、基因融合等遗传变化[6]。通过全基因组测序或靶向测序，研究人员能够识别肿瘤发生发展驱动基因，以及与治疗耐药、复发转移相关的基因组事件。基因组学为肿瘤的早期诊断提供了可能，也为个性化治疗提供了基础[7]。

2.1.2 转录组

转录组学通过分析特定条件下组织中的全部RNA分子（如mRNA、非编码RNA），揭示细胞在生理或病理状态下的基因表达模式。在肿瘤研究中，传统转录组学技术通过高通量RNA测序鉴定了肿瘤生长、侵袭和转移相关的关键基因表达谱，为探索分子机制和筛选治疗靶点提供了重要依据[8]。然而，基于组织样本的群体水平分析可能掩盖了肿瘤细胞间的异质性。随着技术的发展，单细胞转录组学突破了这一局限。例如，计算框架PERCEPTION利用单细胞分辨率数据预测癌症患者对治疗的个体化反应及耐药性演化过程，为精准医疗提供了新策略[9]。

与此同时，空间转录组学进一步将基因表达与肿瘤微环境的空间特征相结合，在保留组织空间位置信息的同时，实现全基因组水平的基因表达分析。如异构图多模态模型stKeep通过整合空间转录表达数据及分子网络信息，构建了细胞模块、基因模块和细胞通讯模块的三维解析体系，揭示了肿瘤生态系统内细胞亚群的空间分布规律及动态互作网络，为理解肿瘤组织的复杂性提供了全新视角[10]。

2.1.3 蛋白组

蛋白组学研究包括组织蛋白质的表达量、修饰、功能以及相互作用等。肿瘤研究中，蛋白组学可以帮助揭示肿瘤细胞的代谢、信号传导及免疫逃逸等关键生物学过程[11]。由于肿瘤的表型变化与蛋白质功能密切相关，蛋白组学提供了对肿瘤细胞功能的深入理解。借助质谱分析等技术，研究人员可以识别肿瘤组织中的特征性蛋白质标志物，为早期诊断、监测治疗反应以及开发新靶点等提供新的策略[12]。

2.1.4 代谢组

代谢组学主要研究组织小分子代谢物如糖类、脂质、氨基酸等[13]。肿瘤细胞的代谢方式与正常细胞有显著不同，通常表现为代谢重编程，即肿瘤细胞通过改变代谢路径以支持其快速增殖和生存[14]。高通量代谢组学技术通过分析肿瘤组织中的代谢物变化，揭示肿瘤独特的代谢特征[15]。

2.1.5 微生物组

微生物组学的兴起为肿瘤研究提供了新视角。研究表明，肿瘤的发生及治疗反应可能与肠道、皮肤及组织内等微生物群落密切相关[16]。例如，某些微生物及代谢物能通过增强肿瘤免疫反应，提升免疫治疗的效果；反之另一些微生物则可能促进肿瘤免疫逃逸[17,18]。肿瘤微生物组分析能为个性化治疗提供全新策略。

2.1.6 表观组

表观组学主要研究基因表达的调控机制，包括DNA甲基化、组蛋白修饰、非编码RNA等。肿瘤中表观遗传变化如DNA甲基化异常、组蛋白修饰失调等，可能导致肿瘤抑癌基因沉默或癌基因激活，成为早期诊断和预后评估的重要标志[19]。表观组学研究，能够深入理解肿瘤调控的分子机制，并为治

疗提供新的靶点[20]。

2.2　多模态数据

2.2.1　临床数据

临床数据是指患者的病史、体检数据、临床表现、治疗方案及治疗反应等各类与患者健康状况相关的资料。患者的年龄、性别、生活习惯、家族史等因素可能与肿瘤发生风险密切相关，而患者的肿瘤分期、组织类型、是否有转移等临床特征对于治疗方案的选择至关重要[21]。临床数据与其他组学数据的结合将为个性化治疗提供依据，帮助制定最合适的治疗策略。

2.2.2　影像数据

影像数据主要包括超声、磁共振成像等，是肿瘤诊断和治疗监控的重要手段。通过影像学检查，医生可以非侵入性观察肿瘤的大小、位置、形态及其与周围组织的关系，评估肿瘤恶性程度、是否有转移或复发。超声能够实时显示肿瘤的动态变化，尤其适用于肿瘤筛查和评估疗效；磁共振成像则通过更精细的软组织成像，提供更详细的局部信息，特别是在脑部、乳腺和前列腺癌等肿瘤的诊断中尤为重要[22-24]。

2.2.3　生化检测数据

生化检测数据包括生物血液、体液和其他样本中的化学物质测量结果，如肿瘤标志物、血糖、肝肾功能、脂类等。在肿瘤的诊断和治疗过程中，生化检测提供了重要的辅助信息。通过监测患者血液标志物的变化如常见肿瘤标志物 CEA（Carcinoembryonic Antigen，癌胚抗原）、AFP（Alpha-Fetoprotein，甲胎蛋白）等，有助于肿瘤的早期发现，或为评估治疗反应和预后预测提供支持[25, 26]。

2.2.4　病理数据

肿瘤病理信息是通过显微镜等手段观察肿瘤组织的形态、结构、分化程度等特征得到的数据。病理信息作为肿瘤诊断的“金标准”，对于肿瘤的分型、分期、预后以及治疗方案的选择具有决定性作用。如通过免疫组化染色、分子病理等技术，可评估肿瘤类型、恶性程度、分子标志物表达、微血管浸润、淋巴结转移等信息[27, 28]。

3　人工智能在肿瘤大数据研究中的应用现状

3.1　人工智能定义及与肿瘤大数据结合的优势

人工智能是指利用计算机模拟人类智能行为以解决问题的技术。近年来，随着计算能力的提升和海量数据的积累，人工智能在各个领域的应用得到了快速发展，尤其在医疗领域，人工智能已展现出巨大的潜力和实际效果。

深度学习是机器学习的子领域之一，其核心在于通过多层非线性变换从数据中自动提取特征。与传统机器学习方法（如线性回归、支持向量机等）依赖人工设计特征不同，深度学习通过深度神经网络（Deep Neural Network，DNN）的架构，将多个层次的中间变量连接在一起，逐层学习数据的抽象表示，从而直接从原始数据中挖掘复杂模式[29]。这种多层结构使得深度学习能够有效处理高维、非线性数据，尤其是在肿瘤大数据这类涉及多源、异构、高维特征的复杂场景中，深度学习的优势更为显著。例如，深度学习通过“反向传播”算法优化网络参数，并结合“堆叠多层”的结构，能够捕捉数据中的高阶交互和复杂关系，从而在分类、预测等任务上表现优于传统统计方法[30]。

3.2　人工智能结合肿瘤大数据解决临床诊疗关键问题

人工智能深度学习已应用在肿瘤早期诊断、分子特征分析、个性化治疗方案制定和治疗疗效预后预测等多方面（图 1）[31, 32]。

3.2.1　精准诊断

精准诊断是肿瘤诊疗的基础，准确识别肿瘤类型和分期对于制定治疗方案至关重要。传统的肿瘤诊断往往依赖于临床症状、影像学检查和病理学分析等，但这些方法耗时长、精度低，且常受到主观判断的影响。近来不少研究证实人工智能结合肿瘤大数据在肿瘤精准诊断领域展示了巨大的潜力。

结合肿瘤病理图像，Barrios 等人[33]利用 CNN 深度学习算法开发 Bladder4Net 网络，用于辅助膀胱癌的诊断。此外，人工智能也可通过对 CT、MRI、X 光等影像数据分析，识别肿瘤位置、大小、形态和恶性程度。例如，Jiang 等人[34]利用人工智能 PMetNet 算法，可实现对 CT 图像分析并判别胃癌是否发生腹膜转移。除了图像数据，人工智能还能整合肿瘤基因组、转录组等多组学数据，为实现肿瘤精准亚分型提供支持。Jin 等人[35]整合了 1 226 例乳腺癌患者的基因组、转录组、蛋白组、代谢组、影像组等多维度数据，提出腔面型乳腺癌分子亚分型，并依靠人工智能模型实现准确预测。

病理诊断大模型的出现为肿瘤精准诊断提供了

更强大的工具。例如，CHIEF 是一种癌症诊断基础模型，通过整合海量的病理图像、基因组数据和临床信息，能够识别多种癌症类型、分析肿瘤基因特征、预测患者生存率，并精确定位肿瘤微环境中的关键区域[36]。

3.2.2 精准治疗

精准治疗是肿瘤诊疗过程中最具挑战性但也是临床最亟需解决的部分。不同患者的肿瘤可能具有不同的分子特征和生物学机制，目前以放化疗为主的传统治疗方式难以快速、准确鉴别治疗靶点，导致部分患者疗效不佳甚至无效。人工智能通过结合分析肿瘤大数据可有效预测治疗靶点，区分治疗响应人群，为精准治疗提供强有力支持。

通过整合病理图像与基因组数据，人工智能可有效预测肿瘤治疗关键靶点，为患者定制个性化治疗方案。Bergstrom 等人[37]结合病理图像开发了人工智能模型 DeepHRD，可实现乳腺癌及卵巢癌患者预测同源重组缺陷，并区分铂类药物治疗疗效。针对肿瘤免疫治疗，斯坦福大学基于图像—文本数据训练的病理学大模型 MUSK 可准确实现肺癌和胃癌的免疫治疗疗效预测[38]。

3.2.3 风险评估

肿瘤的风险评估对于指导患者治疗决策、改善治疗效果及制定后续随访计划具有重要意义。传统的风险评估方法通常依赖于医生对患者临床背景和影像学检查结果的综合判断，缺乏准确便捷且能够有效预测患者的疗效和预后的模型。因此随着肿瘤大数据的积累，人工智能依托大数据可构建更精确的风险评估模型。

通过整合患者的临床信息、组学测序和病理影像等多维度数据，人工智能可以辅助预测肿瘤患者个体预后。Arbour 等[39]基于非小细胞肺癌患者影像资料、文本报告构建深度学习模型，预测其对免疫治疗 PD-1 阻断剂的疗效及无进展生存期。Jee 等[40]基于自然语言处理（Natural Language Processing，NLP）技术对来自真实世界的数据（如患者的临床信息、生活习惯、治疗历史等）进行有效整合，显著提高癌症预后预测的准确性和能力。

3.2.4 药物研发

药物研发是肿瘤治疗中的重要环节，通过寻找可以影响肿瘤进程的靶标，设计调节靶标活性的化合物，以在安全范围内产生治疗效果。传统的药物研发过程周期长、成本高、成功率低，且常规靶点药效常受肿瘤异质性制约。人工智能凭借其极强的特征提取能力和泛化能力，极大加速了药物研发的过程，有效预测新药靶，已成功应用于分子设计、虚拟筛选、活性评估和合成预测等药物开发领域[41]。

在大数据的背景下，人工智能从基因组学、代谢组学和蛋白质组学等多组学数据分析中有效预测潜在药靶。诺贝尔奖获得者 Hassabis 等人[42]提出化合物结构预测框架 AlphaFold3，高准确性预测蛋白质与其他各种生物分子相互作用，以及预测翻译后修饰对这些分子系统的结构影响。Zhang 等人[43]系统总结了目前人工智能在药物研发中的应用、挑战及未来发展，详细列举了人工智能在临床试验和真实世界实践中参与药物研发的具体应用。

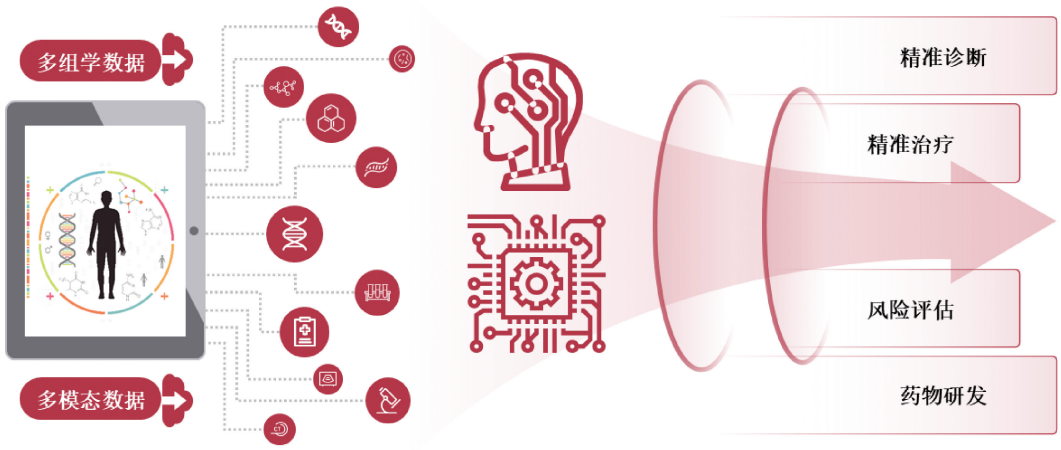

图 1 人工智能结合肿瘤大数据解决临床诊疗关键问题

Fig. 1 Artificial Intelligence Combined with Tumor Big Data to Address Key Issues in Tumor Diagnosis and Treatment

4　基于肿瘤大数据开发人工智能诊疗模型的瓶颈问题及解决路径

实现肿瘤大数据到人工智能诊疗模型的转化通常分为三个阶段：首先，针对特定临床问题（如早期诊断、预后预测或治疗反应评估），系统收集多源异构数据（基因组、影像、病理及临床数据），并进行清洗、标准化和特征提取；其次，基于深度学习技术（如卷积神经网络、图神经网络或 Transformer），开发预测或分类模型，并通过交叉验证、迁移学习等方法优化性能；最后，将模型嵌入临床工作流程（如影像诊断或电子病历系统），进行前瞻性验证，评估其准确性、鲁棒性和临床效用，并结合医生反馈持续优化。然而，这一过程中仍面临诸多瓶颈，限制了人工智能的临床应用效果和推广速度。以下总结其瓶颈问题及可能的解决路径。

4.1　缺乏统一的格式标准和结构化信息提取工具

目前肿瘤领域的大数据来源广泛，不同数据格式存在显著差异，同种数据格式在不同医学中心同样可能存在较大差异。基于上述原因，多中心之间数据的整合共享变得十分困难，不仅增加了数据处理的复杂性，也影响了人工智能模型的训练和应用效果。

要解决这一问题，首先需要制定统一的肿瘤大数据格式标准，包括定义数据存储的标准格式、命名规范以及数据交换协议，确保各类数据能够在不同系统之间顺利传递和共享。此外，针对非结构化数据（如病理报告、医生笔记、影像数据等），需要开发高效的结构化信息提取工具，能够自动识别和提取其中的关键信息，并将其转化为统一的结构化数据形式，以便进一步分析和应用。实现肿瘤大数据标准化不仅能提高数据质量，还能促进跨机构、跨平台的数据共享，推动肿瘤大数据的协同创新，为人工智能诊疗模型的开发和优化提供更加可靠的数据基础。

4.2　多模态数据融合缺乏高效手段实现数据对齐与解决数据缺失

多模态性是临床肿瘤大数据的显著特征，各模态数据从不同角度描述肿瘤特征，提供互补信息。然而，不同模态数据在维度、表示方式、数量和尺度上存在差异，如何对齐和融合这些数据以提升模型性能仍是技术难题。此外，数据缺失问题（如部分患者缺乏影像学或基因检测数据）进一步增加了多模态融合的复杂性。尽管人工智能在单一模态数据处理中取得显著进展，但在多模态数据融合和对齐方面仍面临挑战。

现有研究在这一领域取得一定进展。例如，HECTOR 模型通过整合肿瘤组织染色图像、图像预测的分子类型和癌症分期，成功预测子宫内膜癌复发[28]。然而，多模态数据融合的效率和精度仍需进一步提升，以应对临床肿瘤大数据中的复杂性和多样性。未来研究需进一步突破以下两个技术瓶颈：（1）不同模态数据（如影像学、基因组学、病理学等）在特征空间和表示方式上存在显著差异。例如，影像数据以空间信息为主，而基因组数据则以序列信息为主。如何设计高效的特征提取和映射方法，实现跨模态数据的对齐与统一表示，是当前的主要挑战之一。现有方法在处理高维异构数据时，往往难以捕捉模态间的深层关联，导致信息融合不充分。（2）临床肿瘤数据常因患者个体差异或检测条件限制而存在模态缺失问题（如部分患者缺乏影像学或基因检测数据），如何在数据不完整的情况下实现鲁棒的多模态融合，避免信息丢失或引入偏差。现有方法多依赖于插值或生成模型补全缺失数据，但这些方法可能引入噪声或过度假设，影响模型的可靠性和泛化能力。

4.3　人工智能模型缺乏可解释性

人工智能因其复杂性和“黑箱”特性，面临可解释性不足的问题，尤其在肿瘤诊疗中，医生不仅需要模型的预测结果，还需理解其决策依据。缺乏可解释性会降低医生和患者对模型的信任，影响临床应用。解决这一问题的关键在于提升人工智能模型的可解释性。例如，为应对肺腺癌组织病理学分级挑战所开发的 ANORAK 模型将复杂生长模式转化为数字图像特征，实现了肺腺癌生长模式的自动识别和量化[44]。SHapley Additive exPlanations（SHAP）作为乳腺癌研究中广泛使用的模型无关可解释人工智能（Explainable AI，XAI）技术，能够解释模型预测结果、诊断生物标志物及预后分析；Grad-CAM 等模型特定 XAI 方法在图像处理任务中同样表现出色[45]。

然而，尽管 XAI 方法在临床中具有潜力，其结果的可靠性和有效性仍需进一步验证。未来研究需进一步突破以下技术瓶颈：（1）可解释性与模型性能的平衡。在提升模型可解释性的同时，如何保持或提升其预测性能仍是一个关键挑战，现有 XAI 方法可能因简化模型结构或引入额外约束而降低模型精度。（2）XAI 方法的临床验证与标准化。当前 XAI

方法在临床中的应用缺乏统一标准和验证框架，导致其结果的可靠性和普适性存疑。未来需建立标准化评估体系，确保 XAI 方法在真实临床场景中的有效性和可重复性。突破上述瓶颈将推动人工智能在肿瘤诊疗中的更广泛应用，增强临床医生和患者对 AI 系统的信任。

4.4 模型的临床泛化性验证不足

当前人工智能诊疗模型在特定研究环境中表现优异，但由于训练数据多来自单一机构或数据集，其在不同医疗机构和人群中的泛化能力较弱，导致应用效果不稳定，尤其在跨地区、种族或治疗背景时表现不佳。如果从不同中心获取数据训练模型，则可能面临实际医疗环境中的资源限制和数据隐私问题。

联邦学习为解决这些问题提供了新思路。例如，FLPedBrain 平台通过联邦学习提升肿瘤分类和分割性能，为术前诊断、治疗规划和疾病监测提供支持，同时保护患者数据隐私[46]。训练好的模型还需要在跨地区、种族或治疗背景下检验泛化效能，跨机构、多中心的临床试验是检验模型泛化性的关键。通过在多个医疗机构和不同人群中进行大规模验证，可确保模型的稳定性和适用性。结合迁移学习等方法，利用不同数据源的共享信息，可进一步提升模型在不同环境下的适应性。广泛的验证是确保模型临床可靠性和一致性的基础，为其实际应用提供坚实保障。

4.5 模型临床诊疗应用场景适配性差，临床干预与全周期管理精准性不足

当前人工智能模型与临床工作流程的结合仍存在显著障碍，主要受限于实验室—临床环境脱节、数据更新滞后等问题，导致其难以满足医生实际需求并融入诊疗决策链。此外，肿瘤诊疗涉及全病程管理，人工智能工具需在不同阶段（筛查、诊断、治疗、随访）提供全程支持，并适应患者个体化需求。

提升临床适配性需从以下两方面突破：（1）构建闭环迭代系统。利用人工智能模型通过新数据持续优化的特性，建立“数据—预测—反馈”闭环系统，将临床实时数据（如影像、检验结果）与医生决策反馈纳入模型迭代训练，形成动态更新的诊疗支持工具[47]。（2）全病程智能化整合。开发标准化接口对接电子病历系统，整合患者历史数据（诊断记录、治疗史）与实时监测信息，结合临床指南和医疗资源，动态生成个体化诊疗方案。同时，通过即时数据分析为医生提供实时决策建议（如药物剂量调整、并发症预警），实现诊疗效率与精准性的双重提升。

5 展 望

随着人工智能技术和肿瘤大数据的飞速发展，肿瘤诊疗模式正在发生深刻变革，人工智能与肿瘤大数据的深度融合将推动精准医学和个性化治疗的全面普及。以下对未来实现肿瘤智能诊疗方向进行展望。

5.1 围绕特定癌种的专用大模型构建和临床验证

当前，医学领域的基础大模型研究已广泛开展，例如哈佛医学院等团队开发的 CONCH 模型[48]和 UNI 模型[49]，展现了跨模态数据处理与通用任务泛化的潜力。然而，这些通用模型在肿瘤临床场景中仍面临显著局限，其核心挑战在于缺乏对肿瘤生物学特性、多模态数据关联性及临床路径的深度理解。尽管通用模型具备任务灵活性，但在肿瘤异质性解析、个性化治疗生成等复杂场景中，常因领域知识不足而难以满足临床需求。

肿瘤个性化诊疗需整合多组学、多模态数据（如基因组、影像、病理和临床表型），其复杂性要求模型基于特定癌种知识进行训练。专用大模型通过聚焦特定癌种（如肺癌、乳腺癌）的数据特征与临床逻辑，可精准解析肿瘤生物学行为、识别驱动基因并预测治疗响应。例如皮肤病学领域，专用大模型 MONET 通过整合皮肤病理图像、患者病史和分子标记，实现对皮肤病学医学概念自动标注，辅助临床医生精准诊断[50]。

通过临床考验仍是专用大模型能否广泛应用的关键。未来需要通过多中心、跨地区的大规模临床试验，检验专用大模型在真实临床环境中的有效性和稳定性。通过不断反馈和优化，确保专用大模型能够真正实现跨医疗机构、跨患者群体的泛化性和应用可靠性。

5.2 多维临床诊疗数据赋能的肿瘤智慧诊疗策略

多模态数据的深度整合与应用是肿瘤智慧诊疗的核心。传统的单一数据源难以全面揭示肿瘤的复杂生物学特征，而通过多组学、多模态数据的有机结合，能够从分子、细胞到组织多层次解析肿瘤的发生发展机制。

多模态生成技术为肿瘤智慧诊疗提供了新的工具。例如，MINIM 模型作为医学图像—文本生成模型，能够基于文本描述生成多种器官和成像模态的医学图像（如眼科 OCT、胸部 CT 等），为临床教学、诊断辅助和科研探索提供支持[51]。这类技术不仅

能够弥补数据缺失问题，还可通过生成高质量的多模态数据，进一步丰富肿瘤特征的表征能力，为智慧诊疗提供更全面的数据基础。

跨学科合作是推动肿瘤智慧诊疗发展的关键。医学、生物学、计算机科学等多领域的协同创新，能够加速多模态数据的整合与应用。例如，生物信息学家可开发高效的数据分析算法，临床医生提供领域知识与验证场景，工程师则优化模型部署与临床集成。通过跨学科协作，构建从数据采集、模型开发到临床验证的完整闭环，最终实现肿瘤诊疗的智能化与精准化。

5.3　个体化需求驱动的肿瘤患者人机融合全病程精准诊疗管理

肿瘤诊疗不是一个单向指导的过程，而是一个跨越多治疗阶段的全病程管理过程。患者的需求是动态的，不仅仅局限于肿瘤的治疗效果，还包括早期预防、精准诊断及治疗过程中可能出现的副作用管理、心理状态支持等。人工智能应与临床工作更紧密结合，实现个体化需求驱动的全病程精准诊疗管理。

然而，在实际医疗环境中，个体化精准管理策略的实施面临资源限制和数据隐私问题的挑战。医疗资源的有限性，尤其是在基层医疗机构，可能限制人工智能技术的广泛应用。因此，需要根据医疗机构的资源水平，设计分层管理的方案，确保资源的高效利用。同时，肿瘤患者的医疗数据涉及高度敏感的个人信息，必须严格遵守相关法律法规，采用隐私计算技术（如联邦学习）确保数据安全。通过政策支持、技术普及和多机构协作，可以在保护患者隐私的前提下，推动人工智能在肿瘤全病程管理中的有效应用。

参　考　文　献

[1] World Health Organization. Cancer Today. [2025-02-24]. https://gco.iarc.fr/today/en.

[2] Han B, Zheng R, Zeng H, et al. Cancer incidence and mortality in China, 2022. Journal of the National Cancer Center, 2013, 4: 47—53.

[3] Lee D, Park Y, Kim S. Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches. Briefings in Bioinformatics, 2021, 22(3): bbaa188.

[4] Shi XL, Wang XY, Yao WT, et al. Mechanism insights and therapeutic intervention of tumor metastasis: latest developments and perspectives. Signal Transduction and Targeted Therapy, 2024, 9: 192.

[5] Jiang P, Sinha S, Aldape K, et al. Big data in basic and translational cancer research. Nature Reviews Cancer, 2022, 22(11): 625—639.

[6] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature, 2009, 458(7239): 719—724.

[7] Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine. Nature Reviews Clinical Oncology, 2018, 15(6): 353—365.

[8] Jobanputra V, Wrzeszczynski KO, Buttner R, et al. Clinical interpretation of whole-genome and whole-transcriptome sequencing for precision oncology. Seminars in Cancer Biology, 2022, 84: 23—31.

[9] Sinha S, Vegesna R, Mukherjee S, et al. PERCEPTION predicts patient response and resistance to treatment using single-cell transcriptomics of their tumors. Nature Cancer, 2024, 5(6): 938—952.

[10] Zuo CM, Xia JJ, Chen LN. Dissecting tumor microenvironment from spatially resolved transcriptomics data by heterogeneous graph learning. Nature Communications, 2024, 15: 5057.

[11] Müller JB, Geyer PE, Colaço AR, et al. The proteome landscape of the Kingdoms of life. Nature, 2020, 582(7813): 592—596.

[12] Poulos RC, Hains PG, Shah R, et al. Strategies to enable large-scale proteomics for reproducible research. Nature Communications, 2020, 11: 3793.

[13] Martínez-Reyes I, Chandel NS. Cancer metabolism: looking forward. Nature Reviews Cancer, 2021, 21(10): 669—680.

[14] Schmidt DR, Patel R, Kirsch DG, et al. Metabolomics in cancer research and emerging applications in clinical oncology. CA: A Cancer Journal for Clinicians, 2021, 71(4): 333—358.

[15] Voss K, Hong HS, Bader JE, et al. A guide to interrogating immunometabolism. Nature Reviews Immunology, 2021, 21(10): 637—652.

[16] Cammarota G, Ianiro G, Ahern A, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nature Reviews Gastroenterology & Hepatology, 2020, 17(10): 635—648.

[17] Park JS, Gazzaniga FS, Wu M, et al. Targeting PD-L2-RGMb overcomes microbiome-related immunotherapy resistance. Nature, 2023, 617(7960): 377—385.

[18] Wang H, Rong XY, Zhao G, et al. The microbial metabolite trimethylamine N-oxide promotes antitumor immunity in triple-negative breast cancer. Cell Metabolism, 2022, 34(4): 581—594. e8.

[19] Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics, 2019, 20(4): 207—220.

[20] Gangoso E, Southgate B, Bradley L, et al. Glioblastomas acquire myeloid-affiliated transcriptional programs via epigenetic immunoediting to elicit immune evasion. Cell, 2021, 184(9): 2454—2470. e26.

[21] Liu RS, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature, 2021, 592(7855): 629—633.

[22] Panayides AS, Amini A, Filipovic ND, et al. AI in medical imaging informatics: current challenges and future directions. IEEE Journal of Biomedical and Health Informatics, 2020, 24(7): 1837—1857.

[23] Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nature Reviews Cancer, 2018, 18(8): 500—510.

[24] Saltz J, Almeida J, Gao Y, et al. Towards generation, management, and exploration of combined radiomics and pathomics datasets for cancer research. AMIA Joint Summits on Translational Science Proceedings AMIA Joint Summits on Translational Science, 2017, 2017: 85—94.

[25] Cai GX, Cai MY, Feng ZQ, et al. A multilocus blood-based assay targeting circulating tumor DNA methylation enables early detection and early relapse prediction of colorectal cancer. Gastroenterology, 2021, 161(6): 2053—2056. e2.

[26] Yang JS, Xu RY, Wang CC, et al. Early screening and diagnosis strategies of pancreatic cancer: a comprehensive review. Cancer Communications, 2021, 41(12): 1257—1274.

[27] van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nature Medicine, 2021, 27(5): 775—784.

[28] Volinsky-Fremond S, Horeweg N, Andani S, et al. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. Nature Medicine, 2024, 30(7): 1962—1973.

[29] Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine. Nature Biotechnology, 2018, 36(9): 829—838.

[30] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436—444.

[31] Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. Cancer Cell, 2021, 39(7): 916—927.

[32] Bhinder B, Gilvary C, Madhukar NS, et al. Artificial intelligence in cancer research and precision medicine. Cancer Discovery, 2021, 11(4): 900—915.

[33] Barrios W, Abdollahi B, Goyal M, et al. Bladder cancer prognosis using deep neural networks and histopathology images. Journal of Pathology Informatics, 2022, 13: 100135.

[34] Jiang YM, Liang XK, Wang W, et al. Noninvasive prediction of occult peritoneal metastasis in gastric cancer using deep learning. JAMA Network Open, 2021, 4(1): e2032269.

[35] Jin X, Zhou YF, Ma D, et al. Molecular classification of hormone receptor-positive HER2-negative breast cancer. Nature Genetics, 2023, 55(10): 1696—1708.

[36] Wang XY, Zhao JH, Marostica E, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature, 2024, 634(8035): 970—978.

[37] Bergstrom EN, Abbasi A, Díaz-Gay M, et al. Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. Journal of Clinical Oncology, 2024, 42(30): 3550—3560.

[38] Xiang JX, Wang XY, Zhang XM, et al. A vision-language foundation model for precision oncology. Nature, 2025, 638: 769—778.

[39] Arbour KC, Luu AT, Luo J, et al. Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade. Cancer Discovery, 2021, 11(1): 59—67.

[40] Jee J, Fong C, Pichotta K, et al. Automated real-world data integration improves cancer outcome prediction. Nature, 2024, 636(8043): 728—736.

[41] Chen HM, Engkvist O, Wang YH, et al. The rise of deep learning in drug discovery. Drug Discovery Today, 2018, 23(6): 1241—1250.

[42] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630(8016): 493—500.

[43] Zhang K, Yang X, Wang YF, et al. Artificial intelligence in drug development. Nature Medicine, 2025, 31(1): 45—59.

[44] Pan X, AbdulJabbar K, Coelho-Lima J, et al. The artificial intelligence-based model ANORAK improves histopathological grading of lung adenocarcinoma. Nature Cancer, 2024, 5(2): 347—363.

[45] Ghasemi A, Hashtarkhani S, Schwartz DL, et al. Explainable artificial intelligence in breast cancer detection and risk prediction: a systematic scoping review. Cancer Innovation, 2024, 3(5): e136.

[46] Lee EH, Han M, Wright J, et al. An international study presenting a federated learning AI platform for pediatric brain tumors. Nature Communications, 2024, 15: 7615.

[47] Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 2019, 20(5): e262—e273.

[48] Lu MY, Chen BW, Williamson DFK, et al. A visual-language foundation model for computational pathology. Nature Medicine, 2024, 30(3): 863—874.

[49] Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 2024, 30(3): 850—862.

[50] Kim C, Gadgil SU, DeGrave AJ, et al. Transparent medical image AI via an image-text foundation model grounded in medical literature. Nature Medicine, 2024, 30(4): 1154—1165.

[51] Wang JZ, Wang K, Yu YF, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. Nature Medicine, 2025, 31: 609—617.

Current Applications, Challenges, and Future Prospects of Tumor Big Data and Artificial Intelligence in Tumor Diagnosis and Treatment

Wenxuan Xiao† Shen Zhao† Yizhou Jiang*

Department of Breast Surgery, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China

Abstract Tumor remains a significant global public health challenge. Despite advancements in modern medicine, there are still numerous unmet clinical needs in areas such as precision diagnosis and treatment, risk assessment, and drug development. With the evolution of big data technologies, tumor big data has gradually emerged as a crucial force driving innovation in the field of oncology. This review outlines the characteristics of tumor big data, including multi-omics data (such as genomics, transcriptomics, proteomics, and metabolomics) and multi-modal data (such as clinical, imaging, and pathological data), and explores the current applications of integrating tumor big data with artificial intelligence in tumor diagnosis and treatment. Additionally, we analyze the main challenges faced at this stage, such as insufficient data standardization, technical bottlenecks in multi-modal data integration, limited model interpretability, and a lack of clinical validation, while also discussing potential solutions. Finally, we look ahead to future trends, emphasizing the importance of constructing specialized large models for tumor diagnosis and treatment, achieving multi-dimensional data fusion, and designing personalized, demand-driven precision management strategies.

Keywords tumor diagnosis and treatment; tumor big data; artificial intelligence; multi-omics data; multi-modal data; precision medicine

江一舟　复旦大学附属肿瘤医院乳腺外科主任医师、研究员、博士研究生导师。研究方向为乳腺癌的分子分型和精准治疗。主持国家自然科学基金青年科学基金项目(A 类)、国家重点研发计划课题等。入选上海市优秀学术带头人、上海市浦江人才等计划，获得达摩院青橙奖、上海市科技进步奖一等奖(2/10)等奖励荣誉。

肖文铉　复旦大学附属肿瘤医院乳腺外科肿瘤学博士研究生。主要从事乳腺癌基础与临床转化研究。

赵　珅　复旦大学附属肿瘤医院主治医师，研究方向为肿瘤大数据与人工智能。

(责任编辑　陈　鹤　张　强)

† Contributed equally as co-first authors.

* Corresponding Author, Email: yizhoujiang@fudan.edu.cn