·述　评·

重视医学影像人工智能数据库的标准化建设

石镇维，刘再毅

广东省医学科学院 广东省人民医院放射科，广州 510080

通信作者：刘再毅　电话：020-83870125，E-mail：liuzaiyi@gdph.org.cn

【摘要】医学影像被认为是人工智能最具落地潜力的领域之一。然而，人工智能面临着医学影像大数据持续积累所带来的挑战：缺乏高质量数据集、行业标准、管理规范等。因此，构建符合我国国情、法律/法规及科研人员使用习惯的标准化医学影像数据库势在必行。FAIR 数据准则（可查询、可访问、可交互、可再用）有望在数据库构建、数据采集以及医学影像数据描述术语规范化等方面发挥指导作用。期待在国内学者的共同努力下，推动医学影像人工智能标准化数据库的建设。

【关键词】医学影像；人工智能；FAIR 数据准则；标准化数据库

【中图分类号】R-1；R445　【文献标志码】A　【文章编号】1674-9081(2021)05-0599-03

DOI：10.12290/xhyxzz.2021-0507

Attaching Importance to the Standardized Construction of Artificial Intelligence Database of Medical Imaging

SHI Zhenwei, LIU Zaiyi

Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China

Corresponding author: LIU Zaiyi　Tel: 86-20-83870125, E-mail: liuzaiyi@gdph.org.cn

【**Abstract**】Medical imaging is regarded as one of the most potential domains where artificial intelligence can be applied in practice. However, artificial intelligence is facing challenges resulting from continuous growth of data, such as lack of high-quality data, lack of standardization in domain, lack of effective data management and regulation. Therefore, it is necessary to construct a standardized medical imaging database complying with the national condition of China, laws/regulations, and using habits of researchers. FAIR data principle (findable, accessible, interoperable, and reusable) may play a key role in database construction, data acquisition, and regulating descriptions of medical imaging data. Looking forward to boosting the standardized construction of artificial intelligence databases of medical imaging under the combined efforts of national researchers.

【**Key words**】medical imaging; artificial intelligence; FAIR data principle; standardized database

Funding: National Natural Science Foundation of China (81771912, 82102034); National Science Fund for Distinguished Young Scholars (81925023)

Med J PUMCH, 2021,12(5):599-601

基金项目：国家自然科学基金（81771912，82102034）；国家杰出青年科学基金（81925023）

引用本文：石镇维，刘再毅. 重视医学影像人工智能数据库的标准化建设［J］. 协和医学杂志，2021，12（5）：599-601. doi：10.12290/xhyxzz.2021-0507.

近年来，随着医疗条件的不断改善以及医院信息化程度的不断提高，医学影像数据呈现暴发式增长。据《2018 年医疗人工智能技术与应用白皮书》[1]统计，目前我国医疗数据的年增长率约为 30%。互联网数据中心（Internet Data Center，IDC）的统计数据显示，2020 年全球医疗数据量已达到 2010 年的 40 倍，其中医学影像数据（包括 X 线、超声、CT、MRI、PET、病理图像等）约占 80%[2-3]。目前，医疗影像数据具有大规模（volume）、高增速（velocity）、多种类（variety）、高价值（value）和真实准确（veracity）五大特点，符合当代大数据 5V 特征，因此促进了医学影像人工智能（artificial intelligence，AI）的发展[4]。医学影像大数据在为医学影像 AI 带来良好发展前景和机遇的同时，亦面临着数据方面的巨大挑战。

1 医学影像 AI 面临数据方面的挑战

随着全球学者在医学影像 AI 领域的积累，医学图像智能分析与处理算法愈发成熟，医学影像因此也成为 AI 在医疗行业中最有潜力落地的领域。然而，在医学影像 AI 科学研究中，数据是首要难题。首先，目前普遍缺乏高质量的训练数据，虽然国际上有很多高质量的公开数据库，但数据量和多样性依然十分有限，且存在患者人种差异；其次，缺乏行业统一标准，数据采集标准多样，系统误差较大，缺乏对医学图像和疾病征象的统一认识；最后，整个行业缺乏对医疗数据使用标准的判断依据和监管，且由于存在法律和伦理问题，很大一部分医学影像数据未能发挥最大价值，导致医学影像 AI 发展受阻。

目前医学影像 AI 在数据方面的困难阻碍了科研人员对数据的有效使用，包括：无法获取医学影像数据集信息；缺乏对医学影像数据准确的描述信息（如本体[5]）；无法获知数据的真实含义而导致错误使用；无法获知使用者的基本权利和义务等。为克服上述困难，需要政策与科学理论相结合，以推动医学影像 AI 标准化数据库的建立。2016 年《二十国集团领导人杭州峰会公报》第 12 条指出："我们支持采取适当措施促进开放科学，推动在可找寻、可访问、可交互、可再用的原则下，提高获取公共财政资助的研究成果的便利性。"2018 年我国颁布了《科学数据管理办法》，目的在于进一步加强和规范科学数据管理，保障科学数据安全，提高开放共享水平，更好地支撑国家科技创新、经济社会发展和国家安全。但目前，由于相关责任与权利不清晰，导致科研人员（包括医务人员）参与科学数据使用与管理工作的动力不足；而因对于数据隐私安全及其危害缺乏清晰、明确的定义，导致数据公开以及共享困难；此外，医学影像数据具有独特的性质，例如复杂多样、隐私敏感、长尾突发、类型复杂和分散度高等[6]，因此，亟需建立符合医学影像数据特点的使用和管理标准与规范，并在此基础上建立医学影像 AI 标准化数据库，以实现基于标准化医学影像数据促进医疗 AI 的发展。

2 建立医学影像 AI 标准化数据库的重要性

在医学领域，The Cancer Imaging Archive（TCIA）[7]和 The Cancer Genome Atlas（TCGA）[8]是两个被广泛使用的公开数据库。前者包含常见肿瘤的医学影像数据与相应的临床信息；后者则包含肿瘤的病理图像数据与基因信息。TCIA 与 TCGA 对数据审查十分严格，具有数据质量高、对疾病描述准确、数据来源清晰、使用条件规范等特点，为全球医学影像 AI 的发展作出了巨大贡献。使用公开数据集进行医学影像 AI 模型的训练与验证已经成为一种发展趋势。

除此之外，TCIA 为部分影像数据提供了符合 FAIR［findable（可查询），accessible（可访问），interoperable（可交互），reusable（可再用）］数据管理准则的 DICOM-SEGMENTATION 文件，实现对影像标注数据的 FAIR 化与结构化，进而提升了医学影像数据与 AI 技术之间的交互性，更有益于 AI 模型之间的比较与泛化。2016 年国际组织 FORCE11 正式提出了 FAIR 数据科学管理准则，目的在于对数据进行科学管理。FAIR 数据准则详细描述了如何通过科学的方法进行数据管理[9-11]。首先，提升数据的交互性有助于打破数据与 AI 算法之间的交互壁垒，对于机器学习至关重要；其次，FAIR 数据准则着重强调数据结构化，进而提升数据的可再用性。该准则被提出以来，受到科学数据管理领域的广泛认可。在构建医学影像标准化数据库方面，FAIR 数据准则通过对医学影像数据的采集、处理、使用以及管理等方面进行标准化描述，可为医学影像 AI 科研提供标准化数据保障。因此，FAIR 数据准则为长久以来医学影像领域标准化提供了新的机遇。

3 小结与展望

过去 10 年，我国临床所产生的医学影像数据呈

现暴发式增长，但真正规范且可被用于临床科学研究的医学影像数据却极度匮乏，导致很大一部分科学研究仍然依赖于国际医学影像数据库，尤其是一些公开数据库。因此，构建符合我国国情、法律/法规以及科研人员使用习惯的标准化医学影像数据库势在必行。通过建立标准化医学影像 AI 数据库，可提升医学影像数据质量、实现科学数据价值的最大化、促进医学影像 AI 的发展。FAIR 数据准则所倡导的科学使用和管理原则恰好符合上述目标。若基于该准则构建我国医学影像 AI 标准化数据库：首先，需要建立完备的医学影像数据行业标准，并为科研人员提供便捷的标准医学影像数据信息平台和服务；其次，需要清晰描述医学影像数据产生、处理、使用、管理以及发布等各个环节，明确各方的权利、责任以及义务；再次，需要科研人员和医务人员在数据采集、数据库构建及医学影像数据术语描述等方面达成共识（可参照国际标准构建符合中国国情的医学本体语义库）；最后，计算、存储等硬件设备与软件对构建医学影像 AI 数据库同样重要，亟需开发相应的计算机软件以实现医学影像数据的实时转化及存储，并最终实现医学影像 AI 标准化数据库的建设，促进医疗 AI 的发展。

作者贡献： 石镇维负责查阅文献、撰写初稿及文章修订；刘再毅提出修改意见并审校文章。

利益冲突： 无

参 考 文 献

[1] 互联网医疗健康产业联盟. 2018 年医疗人工智能技术与应用白皮书 [EB/OL]. (2018-04-16)[2021-07-30]. http://www.qianjia.com/html/2018-04/16_289594.html.

[2] Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology [J]. Nat Rev Cancer, 2018, 18: 500-510.

[3] Bi WL, Hosny A, Schabath, MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications [J]. CA Cancer J Clin, 2019, 69: 127-157.

[4] Duncan JS, Insana MF, Ayache N. Biomedical imaging and analysis in the age of big data and deep learning [J]. Proc IEEE, 2019, 108: 3-10.

[5] Hartel FW, Coronado S, Dionne R, et al. Modeling a description logic vocabulary for cancer research [J]. J Biomed Inform, 2005, 38: 114-129.

[6] Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises [J]. Proc IEEE, 2021, 109: 820-838.

[7] Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository [J]. J Digit Imaging, 2013, 26: 1045-1057.

[8] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge [J]. Contemp Oncol (Pozn), 2015, 19: A68.

[9] Vesteghem C, Brøndum RF, Sønderkær M, et al. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives [J]. Brief Bioinform, 2020, 21: 936-945.

[10] Wilkinson MD, Dumontier M, Sansone SA, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework [J]. Sci Data, 2019, 6: 174.

[11] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship [J]. Sci Data, 2016, 3: 160018.

（收稿：2021-06-29 录用：2021-07-29 在线：2021-08-19）

（本文编辑：李 娜）