ICS 11.020
CCS C 05

T/GXAS

团　　体　　标　　准

T/GXAS 1332—2026

非小细胞肺癌术后早期复发人工智能预测数据采集与处理规范

Specifications for data collection and processing for artificial intelligence prediction of early postoperative recurrence of non-small cell lung cancer

2026-06-02 发布　　2026-06-08 实施

广西标准化协会　　发 布

目　　次

前　言

本文件参照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由广西人工智能学会提出、归口并宣贯。

本文件起草单位：桂林市第三人民医院、桂林电子科技大学、全州县人民医院、阳朔县医疗保障事业管理中心、桂林市医药招采中心、平果市人民医院、灵山县人民医院、广西壮族自治区桂东人民医院、钦州市中医医院、桂林市妇女儿童医院、桂林市人民医院、中南大学湘雅二医院桂林医院、桂林市社会福利医院、桂林市七星区科技企业发展中心、深圳挚硕计算机技术有限公司、中国人民解放军南部战区总医院、昆明理工大学、广西人工智能学会、广西壮族自治区南溪山医院、杰创智能科技股份有限公司、广西医科大学第二附属医院、桂林市中医医院、灵川县妇幼保健院、长沙东彦文信息科技有限公司、阳朔县人民医院、灵川县人民医院、浙江飞图影像科技有限公司、桂林理工大学、桂林当桂心脑血管中医院有限公司。

本文件主要起草人：朱飞艳、潘细朋、唐春华、唐桂珠、杜静玲、韦艳霞、罗善修、杜明、凌光锋、蓝如师、时永康、谢灵霜、韦捷、潘旻、覃天良、张捷、杨林明、鲁江、陈俊嗣、李旭升、肖静、卞新军、潘国栋、程子龙、查露露、马群英、廖剑平、佘巍巍、陈美宇、孙鸿玉、张海波、何哲、王羲、胡洁、陈晓旺、曾勉、敬嘉盛、朱江平、蒋碧玲、龙慧鑫、朱海艳、佘纯纯、蒋贤鉴、黎之枞、唐乐剑、颜征鸿、吴祖明、倪辉、郑志平、廖相、何叶、文军、赵晓庆、裴毅、吴林峰、徐莹波、陶毅、黄柳宜、李丽娟、黄敏、黄笛、张文军、王韩勇、黄润彪、陈仙鹤、魏晓晴、党正娣、张玲、谢晓兰、董路熙、马志波、张秋玲、黄泽旗。

非小细胞肺癌术后早期复发人工智能预测数据采集与处理规范

1 范围

本文件界定了非小细胞肺癌术后早期复发的术语和定义、缩略语，规定了非小细胞肺癌术后早期复发人工智能预测数据采集与处理的多模态数据要求、数据采集与处理、数据特征提取、数据多模态特征融合、预测及验证模型输出结果等要求。

本文件适用于基于Transformer模型构建的术后早期复发风险预测模型开展的非小细胞肺癌术后早期复发人工智能预测数据采集与处理。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.31 信息技术 词汇 第31部分：人工智能 机器学习

GB/T 39725 信息安全技术 健康医疗数据安全指南

YD/T 6184 面向互联网应用的健康医疗数据应用脱敏技术要求

YY/T 1833.1 人工智能医疗器械 质量要求和评价 第1部分：术语

YY/T 1833.3 人工智能医疗器械 质量要求和评价 第3部分：数据标注通用要求

3 术语和定义

GB/T 5271.31、YY/T 1833.1界定的以及下列术语和定义适用于本文件。

3.1

非小细胞肺癌术后早期复发 early postoperative recurrence of non-small cell lung cancer

非小细胞肺癌根治性切除术后2年内出现的包括局部复发、区域淋巴结转移及远处转移的恶性肿瘤I期、II期复发。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

CT：计算机断层扫描（Computed Tomography）

DICOM：医学数字成像和通信（Digital Imaging and Communications in Medicine）

HE：苏木精-伊红染色法（Hematoxylin and Eosin Staining）

IHC：免疫组化（Immunohistochemistry）

NSCLC：非小细胞肺癌（Non-Small Cell Lung Cancer）

TNM：肿瘤-淋巴结-转移分期（Tumor-Node-Metastasis）

TILs：肿瘤浸润淋巴细胞（Tumor-Infiltrating Lymphocytes）

WSI：全视野数字化切片（Whole Slide Image）

5 多模态数据要求

5.1 CT影像数据

5.1.1 应明确CT扫描的层厚为薄层≤2 mm、层间距为1 mm、管电压为120 kVp、管电流采用光谱CT自适应电流等关键参数。

5.1.2　数据格式应遵循DICOM格式，包含去标识化处理后完整的患者相关元数据。
5.1.3　影像应无明显伪影、噪声干扰，病灶区域清晰可辨。
5.1.4　应记录胸部平扫、动脉期及静脉期双期增强数据，其中双期增强数据应采用投影空间光谱重建，迭代算法强度等级为4级。
5.1.5　所有数据均应经过脱敏处理，脱敏按YD/T 6184的规定执行，移除元数据中的患者姓名、年龄、出生年月、家庭地址、病史记录、诊断描述等信息。

5.2　病理数据

5.2.1　病理切片应规范制作，HE染色、IHC染色切片质量符合病理诊断标准。
5.2.2　WSI扫描应使用不低于20倍物镜的放大倍数，其分辨率应能满足后续AI分析对细胞或组织结构的辨识要求，图像无模糊、污染等问题，扫描分辨率数值宜≥0.5微米/像素。
5.2.3　应包含病理医生对肿瘤区域、腺体、细胞核、TILs等关键结构的标注信息，数据标注应符合YY/T 1833.3的要求。
5.2.4　包括但不限于以下病理学特征信息：
——病理类型（腺癌、鳞状细胞癌、其他类型）；
——分化程度（高分化、中分化、低分化、未分化）；
——脏层胸膜受累情况；
——脉管癌栓；
——神经束膜侵犯；
——切缘状态（肿瘤完全切除、显微镜下残留和肿瘤肉眼残留）。

5.3　临床数据

主要包括：
——基本信息：患者年龄、性别、职业、婚姻、吸烟史及吸烟量、个人恶性肿瘤病史、肿瘤家族史、体质、肝肾功能等；
——治疗相关信息：手术方式、术后病理分期、肿瘤分化程度、是否接受辅助治疗等；
——手术信息：手术及操作名称、手术日期、切除方式（是否楔形/亚肺叶切除）、出入院时间、出院诊断等；
——TNM分期情况：原发肿瘤大小、区域淋巴结转移、远处转移等；
——随访信息：术后复发时间、复发部位、生存状态等，应记录无病生存期、总生存期等核心生存终点，并明确末次随访截止日期或中位随访时间，随访方式及复发评价标准宜作说明。

5.4　数据安全与隐私保护要求

数据安全按GB/T 39725的规定执行，传输敏感文件或数据时，应使用不少于6位数的数字、大小写字母与特殊符号的强加密设置，或通过如虚拟专用网络、安全文件传输协议等加密通道进行传输，实施权限管理。

6　数据采集与处理

6.1　数据采集

数据采集应符合《涉及人的生物医学研究伦理审查办法》的规定，并按以下方法进行采集：
——设计数据采集表格，从医院信息系统中按年度分类导出临床数据；
——按国际疾病分类第10版编码，从疾病和有关健康问题的国际统计分类中进行检索；
——按国际疾病分类第9版编码，从国际疾病分类手术与操作中进行检索；
——从病案首页中，依据国家平台标准字段，筛查包含病理诊断和CT影像诊断的病例，并结合电子病历记录进行筛选；
——从全结构化电子病历或智慧医院平台中，筛选术前胸部平扫或增强CT扫描检查、经术后病理证实为NSCLC、接受根治性手术切除等病例并采集数据，排除图像有伪影、断层等质量不佳现象的数据；

——邀请肿瘤科、放射科、病理科、呼吸科、胸外科等副主任医师以上相关临床专家对筛选出的病例进行复核确认；
——确认临床数据合格后，以 DICOM 原格式导出对应的影像及病理图片，并以病案号或住院号命名，不应使用患者姓名，并使用不少于 6 位数的数字、大小写字母与特殊符号的强加密设置。

6.2 数据预处理

按以下要求进行数据预处理：
——剔除影像模糊、切片破损、重复及逻辑错误的数据；
——对 CT 影像数据进行重采样、调窗以及灰度值归一化等处理，对病理图像进行扫描倍率标准化、染色归一化等预处理操作；
——采用几何变换、颜色变换、仿射变换等数据增强技术扩充数据集，增强过程应保持病灶特征的完整性；
——对患者个人身份信息进行去标识化处理，所有数据图片不应包含患者姓名、身份证号、联系方式，地址等个人信息。

7 数据特征提取

7.1 CT 影像

7.1.1 采用 nn-UNet 模型，自动分割 CT 影像中的肿瘤病灶，分割结果应能支持肿瘤病灶边界、体积、形态及空间位置等信息的提取，并制定校正方法和教程视频，必要时由副主任以上医师进行人工复核或校正。
7.1.2 应在完成病灶分割后，提取并量化肿瘤病灶的影像组学特征，涵盖几何形态特征、灰度统计特征、纹理特征、多尺度或变换域高阶特征，及由深度学习模型表征的高阶语义特征。

7.2 病理图像

7.2.1 采用 Hover-Net 模型，对 WSI 中的肿瘤组织、腺体、细胞核、间质区域及 TILs 等关键结构进行自动识别，模型识别结果存在不确定性或边界不清的区域，应由病理专业人员进行复核或校正。
7.2.2 应提取肿瘤上皮区和间质区 TILs 密度、肿瘤细胞与 TILs 的空间距离、细胞间互作关系、多尺度纹理特征、TILs 亚类型密度等病理特征，及由深度学习模型表征的高阶语义特征。

7.3 临床特征筛选

采用统计学方法开展单因素分析与多因素分析，对候选临床变量进行系统评估：
——单因素分析中连续变量采用独立样本 t 检验，分类变量采用 χ^2 检验，将 P 值＜0.05 的变量纳入后续多因素分析；
——多因素分析则采用二元逻辑回归，筛选 NSCLC 术后早期复发的显著相关因素。

8 数据多模态特征融合

8.1 根据研究目的和数据条件，采用基于交叉注意力的低秩多模态特征融合方法，整合 CT 影像特征、病理图像特征和临床特征，形成用于术后复发风险预测的综合特征表示。
8.2 CT 影像、病理图像和临床数据应来源于同一患者或同一临床事件。
8.3 采用堆栈降噪自编码网络对融合后的多维特征进行深度关联与重构，对不同模态特征进行规范化或标准化处理。
8.4 处理不同模态数据缺失、不完整或质量不一致的问题，记录对缺失数据的处理方式，并在模型验证报告中说明。
8.5 采用能反映不同模态特征贡献差异的融合策略，通过特征选择、降维或正则化等方法控制高维特征带来的过拟合风险。
8.6 记录特征融合的输入模态、融合层级、融合策略、特征筛选方法和最终纳入模型的特征集合。

9 预测及验证模型输出结果

9.1 模型预测输出

包括但不限于以下内容：

——术后复发风险概率值；

——风险等级分类：宜根据预设临床阈值将概率值转换为风险等级，以变量的不同取值依次作为候选截断点，将样本分组并进行生存分析，选择 P 值最小的分割点作为最优截断值，从而将患者划分为低危、中危、高危；

——风险分层验证：绘制患者首次复发或转移的 Kaplan-Meier 曲线，并比较组间差异，以验证模型对肿瘤复发风险的分层预测效能，若 $P<0.05$，则可证实模型对肿瘤复发风险具有稳定、可靠的分层预测价值；

——可解释性信息（可选）：提供 CT 影像、病理图像、临床特征等各模态特征对预测结果的贡献度排序等。

9.2 输出数据格式

模型预测输出应采用结构化数据格式，包括但不限于以下字段：

——患者唯一标识；

——预测生成时间；

——模型版本号；

——输入数据版本标识；

——预测风险概率值及对应的风险等级；

——模型置信区间或预测不确定性估计。

9.3 模型验证结果输出

模型验证结果应独立于预测输出，以表格或文本形式报告包括但不限于以下核心性能指标：

——时间依赖的 ROC 曲线下面积及其 95%置信区间；

——一致性指数及其 95%置信区间；

——对于不少于 3 个独立中心的外部验证，应分别输出每个中心的验证指标，并汇总平均值及标准差；

——验证结果输出文件应同时注明训练集、内部验证集、外部验证集的样本量、阳性事件数、缺失数据比例及数据采集时间范围。

9.4 模型输出记录

应建立并保存模型输出记录，记录内容包括但不限于：

——模型训练及验证所用数据集的描述；

——最终纳入模型的输入特征完整列表及各模态特征融合策略；

——模型超参数及训练终止条件；

——每次预测输出对应的模型版本、运行环境及输出生成日期。

参 考 文 献

[1] DB14/T 3268-2025 DICOM医学影像数据脱敏规范

[2] 国家卫生健康委，教育部，科技部，国家中医药局. 关于印发涉及人的生命科学和医学研究伦理审查办法的通知（国卫科教发〔2023〕4号）

[3] DICOM Standard, Parts 1-18, PS3.1-3.18-2025, National Electrical Manufacturers Association (NEMA), 2025.

[4] Wang Y, et al. Multi-scale pathology image texture signature is a prognostic factor for resectable lung adenocarcinoma: a multi-center, retrospective study[J].Journal of Translational Medicine, 2022, 20(1):1-17.

[5] Han C, et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels[J]. Medical Image Analysis, 2022, 80: 102487.

[6] Chen RJ, et al. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis[J]. IEEE Transactions on Medical Imaging, 2020, 41: 757-770.

[7] Pan X, et al. SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations[J]. Medical Image Analysis 2023, 88: 102867.

[8] Pan X, et al. LesionMix data enhancement and entropy minimization for semi-supervised lesion segmentation of lung cancer. Applied Soft Computing 2024, 167: 112244.

[9] Liu Z, et al. Label-efficient transformer-based framework with self-supervised strategies for heterogeneous lung tumor segmentation[J]. Expert Systems with Applications, 2025, 269: 126364.

[10] Bian X, et al. Federated cross-source learning for lung nodule segmentation with data characteristic-aware weight optimization[J]. Pattern Recognition, 2026, 172: 112396.

[11] Li W, et al. Data-efficient federated semi-supervised learning framework via pseudo supervision refinement strategy for lung tumor segmentation[J]. Biomedical Signal Processing and Control, 2025, 107: 107793.

[12] Peng H, et al. CNN and transformer feature adaptive fusion for image segmentation with imbalanced weight information[J]. Signal, Image and Video Processing, 2026, 20: 216.

[13] Gillies RJ, et al. Radiomics: Images are more than pictures, they are data[J]. Radiology 2016, 278: 563-77.

[14] Bi WL, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications[J]. CA: A Cancer Journal for Clinicians, 2019, 69: 127-57.

[15] Hosny A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study[J]. PLoS Medicine, 2018, 15: 1-25.

中华人民共和国团体标准

非小细胞肺癌术后早期复发人工

智能预测数据采集与处理规范

T/GXAS 1332—2026

广西标准化协会统一印制