



团 体 标 准

T/SCGS 313011—2024

肺癌免疫治疗人工智能辅助决策 软件算法性能测试方法

Algorithm performance test methods for artificial intelligence-assisted
decision-making software for lung cancer immunotherapy

2024-03-11 发布

2024-03-11 实施

中国图学学会 发布
中国标准出版社 出版

目 次

前言	III
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 测试指标	3
4.1 概述	3
4.2 C-Index	3
4.3 Kaplan-Meier 曲线	3
4.4 Log-rank P	3
4.5 混淆矩阵	3
4.6 准确性	3
4.7 敏感度	4
4.8 特异度	4
4.9 阳性预测值	4
4.10 阴性预测值	4
4.11 约登指数	4
4.12 ROC 曲线	4
4.13 组间相关性系数	4
4.14 Kappa 系数	5
5 测试方法	5
5.1 通则	5
5.2 算法应用场景与测试方法	5
5.3 算法质量特性与测试方法	5
6 测试流程	6
6.1 通则	6
6.2 测试前的准备	6
6.3 测试过程中的记录	7
6.4 测试后的整理	7
7 测试要求	7
7.1 测试环境	7
7.2 测试人员	7
7.3 测试参数	7
7.4 测试数据	7

7.5 测试类型.....9

7.6 测试报告.....9

8 测试结果的判定.....10

8.1 通则.....10

8.2 单个指标测试.....10

8.3 整体测试.....10

附录 A（资料性） 常见测试指标的计算公式和使用总结.....11

A.1 C-Index.....11

A.2 Log-rank P11

A.3 混淆矩阵.....11

A.4 准确性.....11

A.5 敏感度.....11

A.6 特异度.....12

A.7 阳性预测值.....12

A.8 阴性预测值.....12

A.9 约登指数.....12

A.10 组间相关性系数.....12

A.11 Kappa 系数.....13

A.12 评估指标使用总结.....13

附录 B（资料性） 测试数据集描述样例.....14

B.1 数据集适用范围.....14

B.2 数据的获取.....14

参考文献.....16

前 言

本文件按照 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定》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。

本文件由北京航空航天大学提出。

本文件由中国图学学会归口。

本文件起草单位：北京航空航天大学、中国科学院自动化研究所、同济大学附属上海市肺科医院、北京大学人民医院、中国医学科学院肿瘤医院深圳医院、广东省人民医院、中国医科大学附属盛京医院、空军军医大学西京医院。

本文件主要起草人：牟玮、何秉羲、杜洋、陈昶、周健、梁颖、姜磊、石喻、康飞、蒋涛、余云浪、方梦捷、曹偲芳、邹锐阳、卢一诺、田捷。

肺癌免疫治疗人工智能辅助决策 软件算法性能测试方法

1 范围

本文件规定了对采用人工智能技术的肺癌免疫治疗人工智能辅助决策软件的测试要求和测试方法。

本文件适用于针对接受肺癌免疫治疗患者的肺癌影像为 CT(Computed Tomography, 计算机断层扫描)和 PET(Positron Emission Tomography, 核医学检查方法)电子发射断层显像的软件测试。

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。

YY/T 1833.1—2022 人工智能医疗器械 质量要求和评价 第1部分:术语

YY/T 1833.2—2022 人工智能医疗器械 质量要求和评价 第2部分:数据集通用要求

YY/T 1858 人工智能医疗器械 肺部影像辅助分析软件 算法性能测试方法

3 术语和定义

YY/T 1833.1—2022、YY/T 1833.2—2022、YY/T 1858 界定的以及下列术语和定义适用于本文件。

3.1

通过准则 pass criteria

判断一个软件及其算法功能的测试符合预期要求的判断依据。

3.2

成像 imaging

通过医学影像设备对患者医学诊断部位进行采集或扫描,获得肺部肿瘤及其他感兴趣部位的影像学数据。

3.3

成像参数 imaging parameters

采集医学影像时所设置的或所具有的可影响成像结果的各类参数。

注:包括层厚、分辨率、像素间距、重建算法、扫描时间等。

3.4

临床指标 clinical characteristics

通过体格检查、实验室检测、组织活检等手段获得的可提供患者及肿瘤状况相关信息的指标。

注:不包含基于医学影像获得的指标。

3.5

影像学主观征象 subjective signs in radiology

由医生归纳得到的医学影像上患者及肿瘤状况的形态学信息。

3.6

金标准 gold standard

根据当前临床医学界公认的最可靠诊断方法获得的诊断结果。

3.7

内部测试集 internal test set

与训练集患者来源中心相同,并用于模型测试的数据集。

3.8

外部测试集 external test set

与训练集患者来源中心不同,并用于模型测试的数据集。

3.9

总生存期 overall survival

从开始治疗到患者出现任何原因导致的死亡或失访的时间。

注:被称作总生存时间。

3.10

无进展生存期 progression-free survival

从开始治疗到患者出现疾病进展、死亡或失访的时间。

注:疾病进展的定义为:a) 以研究过程中所有测量的靶病灶直径之和的最小值作为参考值,靶病灶的直径之和相对增加至少 20%;b) 直径之和增加的绝对值至少 5 mm;c) 出现一个或多个新病灶。

3.11

完全缓解 complete response

固定随访时间内所有靶病灶消失,全部病理淋巴结(包括靶结节和非靶结节)的直径减少到小于 10 mm。

3.12

部分缓解 partial response

以基线直径总和作为参考值,固定随访时间内靶病灶的直径总和至少减小 30%。

3.13

疾病稳定 stable disease

固定随访时间内,靶病灶的直径总和减小不超过 30%,或相对增加不超过 20%。

3.14

持续临床获益 durable clinical benefit

无进展生存期大于 6 个月。

3.15

假性进展 pseudo progression

第一次评估较基线/最低点相比,靶病灶最长直径之和增加 $\geq 20\%$ 的情况下,若间隔 ≥ 4 周的第二次评估为非进展,即为假性进展。

3.16

超进展 hyper progression

免疫治疗期间,肿瘤生长率增加 2 倍或以上,治疗失败时间 < 2 个月,或 2 个月内肿瘤负荷增加超过 50%。

3.17

主要病理缓解 major pathological response

新辅助免疫治疗后,通过测量切除的原发性肿瘤中残留存活肿瘤的百分比,在原发性肿瘤床中存活的肿瘤细胞不超过 10%。

3.18

病理完全缓解 pathological complete response

新辅助免疫治疗后,肺癌所有切除标本。

注:包括区域淋巴结,苏木精-伊红染色切片完全评估后无任何残余肿瘤细胞。

4 测试指标

4.1 概述

肺癌免疫治疗人工智能辅助决策软件的测试指标通常包括针对预后评估等回归任务(即无进展生存评估、总生存评估)的一致性指数(C-Index)、生存曲线(Kaplan-Meier 曲线)、对数秩检验 P 值(log-rank P),针对疗效预测等分类任务(即完全/部分缓解预测、持续临床获益预测、假性进展预测、超进展预测)的混淆矩阵、准确性、敏感度、特异度、阳性预测值、阴性预测值、约登指数、ROC 曲线(Receiver Operating Characteristic Curve)等,及针对算法质量的测试指标,如组间相关性系数、Kappa 系数等,具体计算公式和使用总结见附录 A。

4.2 C-Index

生存任务中也称作 C 指数,是指所有病人对子中预测结果与实际结果一致的对子所占的比例。其估计了预测结果与实际观察到的结果相一致的概率。C-Index 以百分数表达(数值分数乘以 100),取值范围为 0~100%,数值越大,则该软件对风险的预测能力越好。

4.3 Kaplan-Meier 曲线

主要分析单一因素对生存期的影响,用于估计患者生存率并绘制生存曲线。生存曲线以生存时间为横轴,生存率为纵轴,其呈现出连续阶梯形曲线,用于表明生存时间与生存率之间的关系。生存曲线一般是平滑且水平延伸的,当某个时间点一旦有患者发生终点事件(如死亡),生存曲线就会垂直下降,下降幅度是该时间点上患者发生终点事件个数与上一个时间节点后随访的患者样本量的比。

4.4 Log-rank P

Log-rank 检验是对生存分析中两组数据在生存函数方面是否存在显著性差异的非参数检验方法,它基于两组数据的观测值和期望值进行比较,推断两组时间终点事件之间的差异性。在该检验中, P 值用于判断两组数据间是否具有统计学意义, P 值的阈值为 0.05,若小于 0.05 则具有统计学意义。

通过该检验获取的 Log-rank P 值,可以判断两组数据是否具有显著的差异。

4.5 混淆矩阵

混淆矩阵由软件输出的分类预测值与金标准的分类结果对比结果绘制,各行代表测试数据的真实类别归属,各列代表软件对测试数据的预测结果。

4.6 准确性

准确性指软件分类正确的样本数占总样本数的比例,用于评价软件预测结果的整体正确性。准确性以百分数表达(数值分数乘以 100),取值范围为 0~100%,数值越大,则该软件的分类性能越好。

4.7 敏感度

敏感度指在所有实际正例中,软件能够正确识别的正例的比例,用于评价软件对正例的识别能力。敏感度以百分数表达(数值分数乘以 100),取值范围为 0~100%,数值越大,则该软件对正例的识别能力越好。

4.8 特异度

特异度指在所有实际负例中,软件能够正确识别的负例的比例,用于评价软件对负例的识别能力。特异度以百分数表达(数值分数乘以 100),取值范围为 0~100%,数值越大,则该软件对负例的识别能力越好。

4.9 阳性预测值

阳性预测值指在所有被软件预测为正例的样本中,实际正例的比例,用于评价软件预测为正例时的准确性。阳性预测值以百分数表达(数值分数乘以 100),取值范围为 0~100%,数值越大,则该软件预测为正例时的准确性越大。

4.10 阴性预测值

阴性预测值指在所有被软件预测为负例的样本中,实际负例的比例,用于评价软件预测为负例时的准确性。阴性预测值以百分数表达(数值分数乘以 100),取值范围为 0~100%,数值越大,则该软件预测为负例时的准确性越大。

4.11 约登指数

约登指数综合考虑了软件的敏感度和特异度,是一种对软件综合性能的评价指标。约登指数无单位,取值范围为-1~1,数值越接近 1 表示软件预测性能越好,数值越接近-1 表示软件预测性能越差。

4.12 ROC 曲线

ROC 曲线是一种用于评估软件分类预测性能的图形展示方式。统计软件的预测概率在不同分类阈值下将测试样本分成正类和负类。对于每个阈值,计算真阳性率(即敏感度)和假阳性率(即 1-特异度)。以真阳性率为纵坐标,以假阳性率为横坐标绘制 ROC 曲线,并计算 AUC(Area Under Curve)。ROC 曲线越靠近左上角,表示软件的预测性能越好。AUC 无单位,取值范围为 0~1。当 AUC 为 1 时,表示软件的预测性能完美,即软件能够完美地区分正例和负例;当 AUC 为 0.5 时,表示软件的预测性能与随机预测相当;当 AUC 小于 0.5 时,表示软件的预测性能比随机预测还差,即软件的预测结果与真实标签的关系相反。此外,在预测值多时,可根据目标研究问题,选择使用 macro-ROC 或 micro-ROC。macro-ROC 即将所有类别分别绘制 ROC 曲线后求平均获得;micro-ROC 即将样本的多分类任务通过是否是目标类别划分为二分类问题,并绘制 ROC 曲线。

4.13 组间相关性系数

ICC(Intraclass Correlation Coefficient)利用组间和组内的方差评价组间相关性的强度。ICC 无单位,取值范围为 0~1。当 ICC 为 0 时,表示组间没有相关性;当 ICC 接近 1 时,表示组间具有高度的相关性。

4.14 Kappa 系数

Kappa 系数用于评价软件预测结果与金标准的一致性。Kappa 系数无单位,取值范围为-1~1。数值越接近 1 表示软件预测结果与金标准的一致性越好,数值越接近-1 表示软件预测结果与金标准的一致性越差。当 Kappa 系数为 0 时,表示软件预测结果与金标准之间的一致性与随机一致性相当。

5 测试方法

5.1 通则

辅助决策产品的预期场景包括辅助免疫治疗后的生存预测场景和影像学疗效评价预测场景,以及新辅助免疫治疗后的病理学疗效评价预测场景,不包括影像前处理及过程优化。

5.2 算法应用场景与测试方法

5.2.1 生存预测场景

对具有无进展生存预测、总生存预测等回归任务的产品,测试人员应向待测算法输入测试集,输出与参考标准格式兼容的结果,计算一致性指数(C-Index)、Kaplan-Meier 曲线、对数秩检验 P 值(Log-rank P)等测试指标;对具有持续临床获益预测、假性进展预测、超进展预测等分类任务的产品,测试人员应向待测算法输入测试集,输出与参考标准格式兼容的结果,计算混淆矩阵、准确性、敏感度、特异度、阳性预测值、阴性预测值、约登指数、ROC 曲线等指标。

5.2.2 影像学疗效评价预测场景

对具有完全/部分缓解预测等分类任务的产品,测试人员应向待测算法输入测试集,输出与参考标准格式兼容的结果;计算混淆矩阵、准确性、敏感度、特异度、阳性预测值、阴性预测值、约登指数、ROC 曲线等指标。

5.2.3 病理学疗效评价预测场景

对具有病理缓解预测和病理完全缓解预测等分类任务的产品,测试人员应向待测算法输入测试集,输出与参考标准格式兼容的结果;计算混淆矩阵、准确性、敏感度、特异度、阳性预测值、阴性预测值、约登指数、ROC 曲线等指标。

5.3 算法质量特性与测试方法

5.3.1 鲁棒性

肺癌免疫治疗人工智能辅助决策软件的鲁棒性评价应采用亚组分析的方式进行评价,可将测试数据按患者年龄、性别、肿瘤亚型、肿瘤部位、肿瘤分期、影像设备厂家、成像参数等分成不同亚组,分别计算各个亚组上软件的测试指标,通过对比测试指标(包含准确性、AUC 等)在各个亚组上与在测试数据整体上的差异评价软件的鲁棒性。

5.3.2 重复性

肺癌免疫治疗人工智能辅助决策软件的重复性评价应基于随机选出的 30 个~200 个测试数据,建立对照数据集,多个测试人员分别进行测量,考查软件输出在不同测试人员间的稳定性,并使用组间相关性系数进行定量评价。

5.3.3 一致性

测试人员可通过修改测试参数(如输入影像中肿瘤标注结果等)测试软件输出结果的变化程度,用于评价该次测试与原本测试之间的一致性,使用 Kappa 系数进行定量评价。

5.3.4 效率

应记录测试人员在临床典型场景下使用肺癌免疫治疗人工智能辅助决策软件对患者的处理事件,通常以数据导入为起点,以软件输出结果为终点,进行多次计时取平均值。

6 测试流程

6.1 通则

测试流程包括测试前的准备、测试过程中的记录和测试后的整理。测试人员应按测试计划进行测试,在测试过程中记录测试日志,流程图如图 1。

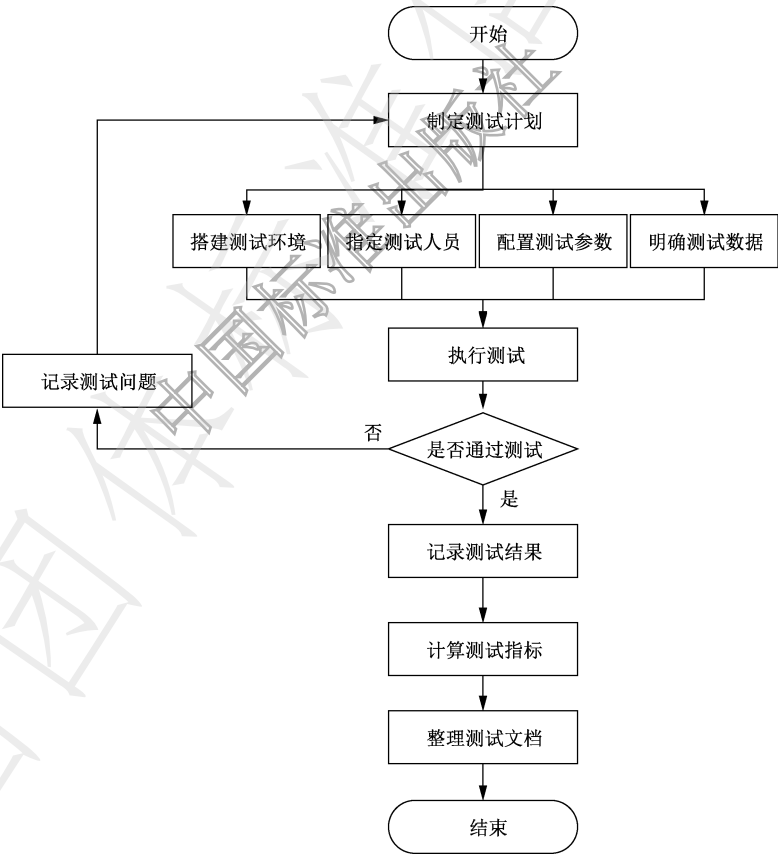


图 1 测试流程图

6.2 测试前的准备

测试前准备满足以下要求：

- a) 测试人员应根据产品预期用途、应用场景确定测试通过准则,编写测试计划,确定测试数据(包含金标准)、测试环境、测试参数、测试人员;
- b) 测试人员应检查待测试的软件是否能对所有测试数据进行处理,并确认输出与输入的测试数

据有唯一对应关系的处理结果。

6.3 测试过程中的记录

测试过程中的记录满足以下要求：

- a) 测试人员应记录肺癌免疫治疗人工智能辅助决策软件对每个测试数据的处理情况,包括处理时间、处理结果,以及软件在运行过程中出现的中间结果和异常提示;
- b) 测试人员应保证记录的软件处理结果的完整性与可追溯性。

6.4 测试后的整理

测试后整理满足以下要求：

- a) 测试人员应将肺癌免疫治疗人工智能辅助决策软件对测试数据的处理结果导出为结构化数据;
- b) 测试人员应基于此结构化数据与测试数据的金标准计算各个测试指标,记录到测试报告中,并对试验结果与产品声称性能指标的符合性做出判定;
- c) 测试人员应整理好完整的测试文档,包括测试计划、测试记录和测试报告。

7 测试要求

7.1 测试环境

在进行肺癌免疫治疗人工智能辅助决策软件性能评价时,满足如下软硬件环境和场地环境要求：

- a) 应采用软件正常运行需要的最低或推荐的软硬件配置及网络环境进行测试,详细记录各类配置,包括:CPU 型号、RAM 型号、显卡型号、操作系统、软件版本、各类支撑软件库的版本号等;
- b) 场地的环境因素不对软件运行产生干扰;
- c) 如存在多个软件环境,且软件环境中规定的运行库/框架等差异对算法性能可能存在影响的,应当在所有存在疑问的环境中分别测试。

7.2 测试人员

测试人员的操作水平应符合软件在临床使用时对使用者的临床医学技能和软件使用技能的普遍要求。

7.3 测试参数

在测试过程中,测试人员应根据测试数据的实际情况按照软件推荐或默认进行参数配置,并对测试时需要配置的参数进行详细记录,包括但不限于模型参数、模型种类和结构、关键公式、预处理方法、后处理方法。

7.4 测试数据

7.4.1 通则

测试数据应严格符合软件在临床使用时对软件输入数据的标准和要求,应包括数据的纳入排除标准、数据的种类要求、影像数据的质量要求、影像数据的格式要求、测试数据样本量数据要求、标注规范和数据传输安全措施。肺部肿瘤测试数据集描述样例见附录 B。

7.4.2 数据的纳入排除标准

根据软件开发时采用的患者及其数据的纳入标准、排除标准,规定软件所适用的具体人群范围。

7.4.3 数据的种类要求

对输入数据的种类要求应规定软件输入的临床数据种类,包括但不限于:医学影像(CT、PET等)、临床指标(血检指标、肿瘤标志物等)、影像学主观征象(肿瘤大小、内部钙化程度、边缘形态等)、影像学半定量指标(SUV_{max}值、长径、短径等)。

7.4.4 影像数据的质量要求

软件对输入影像数据的质量要求,在图像端参数宜包括信噪比、对比度、空间分辨率、是否增强扫描等;在采集端参数宜包括重构核、管电压电流、重建算法、噪声等效计数率等。

7.4.5 影像数据的格式要求

软件输入的文件类型包括但不限于 Dicom、Nrrd、Nii 等。

7.4.6 测试数据样本量

7.4.6.1 测试人员应根据产品预期用途和临床应用场景,在保证研究具有一定可靠性条件下,对测试数据的最小样本量进行限定,确保测试具备科学性和经济性。对于预后评估任务模型,即预测目标是总生存、无进展生存等生存预测,样本量 N 应不少于按公式(1)计算的结果:

$$N = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta} \right)^2}{\log^2(b) p_1 p_2 d} \quad \dots\dots\dots (1)$$

式中:

$Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}$ ——标准正态分布的分位数;

b ——两组的危险比;

p_1, p_2 ——高风险组和低风险组的危险率;

d ——观察到规定时间的比率。

7.4.6.2 对于疗效预测模型,即预测目标是否发生客观缓解、持续临床获益、超进展、假性进展等分类预测,可采用灵敏度计算单次测试中阳性样本的样本量,用特异度计算单次测试中阴性样本的样本量, N_2 计算公式(2)如下:

$$N_2 = \frac{Z_{1-\frac{\alpha}{2}}^2 P(1-P)}{\Delta^2} \quad \dots\dots\dots (2)$$

式中:

N_2 ——单次测试中阳性样本(发生客观缓解、持续临床获益、超进展、假性进展)/阴性样本(未发生上述事件)的样本量;

$Z_{1-\frac{\alpha}{2}}$ ——标准正态分布的分位数;

P ——灵敏度或特异度的预期值;

Δ —— P 的允许误差大小,一般取 P 的 95% 置信区间宽度的二分之一,常用取值为 0.05~0.10。

7.4.7 数据标注规范

7.4.7.1 标注参考依据:胸部 CT 肺结节数据标注与质量控制专家共识(2018 年)。

7.4.7.2 标注流程:标注流程以多轮次分组交叉进行,主要包含肺部肿瘤的检出、边界分割和审核。每一批标注任务由标注组长(8 年以上临床经验)带领两名标注医师承担,并由仲裁专家(15 年以上临床经验)进行审核,分为 3 个主要环节:

- a) 检出环节:3 名标注医师背靠背独立标注,然后用计算机自动判断检出的一致性,以所有人标注结果的并集作为结果;
- b) 边界分割环节:在检出完成之后,肿瘤的边界分割由 1 名标注医师执行,由标注组长进行审核;
- c) 审核环节:由标注组长和仲裁专家各自独立对检出和边界分割结果进行审核与修改,纠正漏诊、误诊和误判。如果遇到疑难问题,仲裁专家可以组织进行集体讨论。

7.4.8 数据传输安全措施

测试过程中应保护患者隐私,对输入数据进行脱敏处理,隐去患者的姓名、地址、联系方式、ID 号等信息。

7.5 测试类型

7.5.1 通则

在进行肺癌免疫治疗人工智能辅助决策软件性能评价时,测试人员应明确该次测试的测试类型是回顾性还是前瞻性。

7.5.2 回顾性测试

回顾性测试是从以往数据库中入组的患者组成内部测试集和外部测试集评价软件性能。

7.5.3 前瞻性测试

前瞻性测试应在中国临床试验注册中心(Chinese Clinical Trial Registry, ChiCTR)、美国临床试验数据库(ClinicalTrials.gov)等世界卫生组织临床试验注册平台(International Clinical Trials Registry Platform, ICTRP)一级注册机构上注册登记,并严格地从临床中连续纳入符合要求的患者。

7.6 测试报告

测试报告对测试结果进行客观、定量描述,内容应至少包含:

- a) 软件环境;
- b) 硬件环境;
- c) 测试集描述;
- d) 测试类型描述;
- e) 算法性能指标的符合性分析,包含性能指标的定义、测试通过准则;
- f) 算法错误分析。

8 测试结果的判定

8.1 通则

测试人员应在测试前对单个测试指标与软件整体确定通过准则。

8.2 单个测试指标

对于单个测试指标,测试人员应确定其需要达到的标称值及允差,并描述允差与测试数据样本量之间的关系。

8.3 整体测试

对于软件整体,测试人员应根据其技术特性、预期用途和使用场景,确定各个测试指标的权重,以及加权求和的总体分数需要达到的阈值。最终,通过计算总体分数判定软件是否通过测试。

附录 A

(资料性)

常见测试指标的计算公式和使用总结

A.1 C-Index

定义可配对样本对数为 N , 预测正确的样本对数为 K , 按公式(A.1)计算。

$$C\text{-Index} = \frac{K}{N} \quad \dots\dots\dots (A.1)$$

A.2 Log-rank P

通过公式(A.2)计算统计量, 进而获得 P 值。

$$P = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \quad \dots\dots\dots (A.2)$$

式中:

O_A ——组 A 的所有观察数之和;

E_A ——组 A 的所有期望数之和;

O_B ——组 B 的所有观察数之和;

E_B ——组 B 的所有期望数之和。

A.3 混淆矩阵

表 A.1 提供了二分类混淆矩阵的示例。

表 A.1 二分类混淆矩阵示例

金标准	软件输出为阳性	软件输出为阴性
阳性	TP	FN
阴性	FP	TN
注: TP 为真阳性, FP 为假阳性; TN 为真阴性, FN 为假阴性。		

A.4 准确性

通常按式(A.3)计算。

$$Acc = \frac{\sum_{i=1}^n N_{i,i}}{\sum_{j=1}^n \sum_{l=1}^n N_{j,l}} \times 100\% \quad \dots\dots\dots (A.3)$$

式中:

Acc —— 准确性;

$N_{i,i}$ —— 实际归属第 i 类的样本被预测为第 i 类的样本个数, 单位为个;

$N_{j,l}$ —— 实际归属第 j 类的样本被预测为第 l 类的样本个数, 单位为个。

A.5 敏感度

通常按式(A.4)计算。

$$\text{Sen} = \frac{TP}{TP + FN} \times 100\% \quad \dots\dots\dots (A.4)$$

式中：
Sen —— 敏感度；
TP —— 真阳性样本的个数,单位为个；
FN —— 假阴性样本的个数,单位为个。

A.6 特异度

通常按式(A.5)计算。

$$\text{Spe} = \frac{TN}{FP + TN} \times 100\% \quad \dots\dots\dots (A.5)$$

式中：
Spe —— 特异度；
TN —— 真阴性样本的个数,单位为个；
FP —— 假阳性样本的个数,单位为个。

A.7 阳性预测值

通常按式(A.6)计算。

$$\text{PPV} = \frac{TP}{TP + FP} \times 100\% \quad \dots\dots\dots (A.6)$$

式中：
PPV —— 阳性预测值；
TP —— 真阳性样本的个数,单位为个；
FP —— 假阳性样本的个数,单位为个。

A.8 阴性预测值

通常按式(A.7)计算。

$$\text{NPV} = \frac{TN}{FN + TN} \times 100\% \quad \dots\dots\dots (A.7)$$

式中：
NPV —— 阴性预测值；
TN —— 真阴性样本的个数,单位为个；
FN —— 假阴性样本的个数,单位为个。

A.9 约登指数

通常按式(A.8)计算。

$$Y = \text{Sen} + \text{Spe} - 1 \quad \dots\dots\dots (A.8)$$

式中：
Y —— 约登指数；
Sen —— 敏感度；
Spe —— 特异度。

A.10 组间相关性系数

通常按式(A.9)计算。

$$ICC = \frac{MS_r - MS_e}{MS_r + (c - 1)MS_e} \quad \dots\dots\dots (A.9)$$

式中：

ICC ——组间相关性系数；

MS_r ——输出在不同测试人员间的方差；

MS_e ——输出在测试人员间的方差；

c ——测试人员数量,单位为个。

A.11 Kappa 系数

通常按式(A.10)及式(A.11)计算。

$$K = \frac{p_o - p_e}{1 - p_e} \quad \dots\dots\dots (A.10)$$

$$p_e = \frac{\sum_{i=1}^n \left(\sum_{j=1}^n N_{i,j} \times \sum_{j=1}^n N_{j,i} \right)}{\left(\sum_{a=1}^n \sum_{b=1}^n N_{a,b} \right)^2} \quad \dots\dots\dots (A.11)$$

式中：

K ——Kappa 系数；

p_o ——预测准确率；

p_e ——偶然一致性；

Acc ——该次测试结果相对于原本测试结果的准确性；

$N_{i,j}$ ——实际归属第*i*类的样本被预测为第*j*类的样本个数,单位为个；

$N_{j,i}$ ——实际归属第*j*类的样本被预测为第*i*类的样本个数,单位为个；

$N_{a,b}$ ——实际归属第*a*类的样本被预测为第*b*类的样本个数,单位为个。

A.12 评估指标使用总结

详见表 A.2。

表 A.2 评估指标使用总结

公式编号	指标	使用场景
A.1	C-Index	生存预测评估
A.2	Log-rank <i>P</i>	生存预测评估
A.4	准确性	疗效预测评估
A.5	敏感度	疗效预测评估
A.6	特异度	疗效预测评估
A.7	阳性预测值	疗效预测评估
A.8	阴性预测值	疗效预测评估
A.9	约登指数	分类模型截断值选择
A.10	组间相关系数	结果一致性评估
A.11	Kappa 系数	结果一致性评估

附 录 B
(资料性)
测试数据集描述样例

B.1 数据集适用范围

数据集适用于能对肺部肿瘤 CT、PET 等图像进行分析的软件产品,这些产品预期用途为肺部肿瘤辅助检出、分类等。

B.2 数据的获取

数据的获取通过如下方式。

- a) 患者人群:数据来源于某家医院,年龄 18 岁以上,男性占×%,女性占×%。
- b) 采集场所:某家三甲医院,某家二级医院。
- c) 影像设备:主流厂家的 CT 设备/PET 设备等。
- d) 数据格式:Dicom、Nrrd、Nii。
- e) 采集人员:临床医生。
- f) 伦理:肺部肿瘤测试数据集应获得医院伦理委员会批准,患者的隐私保护应当满足法规的要求,详见表 B.1。

表 B.1 测试数据集

年龄	40 岁~60 岁	××例
	61 岁~80 岁	××例
	>80 岁	××例
性别	男	××例
	女	××例
肿瘤种类	肺腺癌	××例
	肺鳞癌	××例
	小细胞癌	××例
	其他类型	××例
影像设备机型	××公司××型号	××例
	××公司××型号	××例
部位	右肺上叶	××例
	右肺中叶	××例
	右肺下叶	××例
	左肺上叶	××例
	左肺下叶	××例
吸烟史	有	××例
	无	××例
肿瘤分期	T1	××例

表 B.1 测试数据集（续）

肿瘤分期	T2	××例
	T3	××例
	T4	××例
ECOG PS	0	××例
	1	××例
	2	××例
	3	××例
	4	××例

参 考 文 献

- [1] Eisenhauer E A, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)[J]. European journal of cancer, 2009, 45(2): 228-247.
- [2] Seymour L, et al. RECIST working group. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncology, 2017, 18(3):e143-e152.
- [3] Wei Mu, Lei Jiang, Yu Shi, et al. Non-invasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of PET/CT images, Journal for ImmunoTherapy of Cancer, 9 (6), e002118, 2021.
- [4] Wei Mu, Ilke Tunali, Jhanelle E. Gray, et al. Radiomics of 18F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy, European journal of nuclear medicine and molecular imaging, 2020,47 (5), 1168-1182.
- [5] Wei Mu, Ying Liang, Lawrence O Hall, et al. 18F-PET/CT habitat radiomics predicts outcome of cervical cancer patients treated with chemoradiotherapy, Radiology: Artificial Intelligence, 2(6): e190218, 2020.
-