ICS 11.040
CCS C 30

T/ZSMM

浙 江 省 数 理 医 学 学 会 团 体 标 准

T/ZSMM XXXX—XXXX

医学人工智能治理综合评价指南 第 2 部分：安全评价指标

Guideline for comprehensive evaluation of medical artificial intelligence social governance —Part 2: Safety evaluation indicators

（报批稿）

（本草案完成时间：2026 年 2 月 11 日）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

浙江省数理医学学会　　发 布

目　　次

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则》的规定起草。

本文件是《医学人工智能治理综合评价指南》的第2部分。已经发布部分如下：

——T/ZSMM XXXX《医学人工智能治理综合评价指南　第1部分：总则》；

——T/ZSMM XXXX《医学人工智能治理综合评价指南　第2部分：安全评价指标》。

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。

本文件由浙江省数理医学会学会提出并归口。

本文件起草单位：南方医科大学、南方科技大学、深圳市卫生健康委员会、深圳市人民医院、浙江数字内容研究院、深圳市卫生健康发展研究与数据管理中心、中国医学科学院医学信息研究所、阜外华中心血管医院、上海市第六人民医院、南方医科大学第三附属医院、南方医科大学珠江医院、南方医科大学第八附属医院、南方医科大学南方医院赣州医院、南方医科大学中西医结合医院、东莞市石碣医院、广东医科大学附属医院、深圳市第四人民医院、深圳市妇幼保健院、四川大学华西第二医院、香港大学深圳医院、中山市人民医院、珠海市人民医院、佛山市第一人民医院、前海人寿广州总医院、广州市红十字会医院、北京大学深圳医院。

本文件主要起草人：毛燕娜、王冬、姜虹、朱春艳、耿庆山、汤昊宬、丁万夫、郑静、崔书亭、张冬云、李晨程、曹艳林、刘咏梅、许彬彬、陈宝颖、潘鑫、吴超梅、王亚琴、郭洪波、曹蓓、戴辉、杜庆锋、刘仲文、蔡定彬、王诚、李笑天、张少毅、徐小平、黄晓星、郭煜、段光荣、周宏峰、张立贤、赵永胜。

引　言

医学人工智能治理综合评价包括安全评价、风险评价、效用评价、效率评价、效益评价等，由于内容比较多，拟由以下六个部分构成：

——第1部分：总则；

——第2部分：安全评价指标；

——第3部分：风险评价指标；

——第4部分：效用评价指标；

——第5部分：效率评价指标；

——第6部分：效益评价指标。

本文件为——T/ZSMM XXXX《医学人工智能治理综合评价指南　第2部分：安全评价指标》，给出了医学人工智能治理综合评价中安全评价的评价指标体系框架以及评价要素的建议。主要针对当前可预见的典型场景与通用要求进行规定。鉴于医学人工智能治理具有高度复杂性且技术演进迅速，对于本文件尚未覆盖的特殊情形，建议在实施评价时明确适用范围限制或相关免责说明，以保持治理工作的审慎性与适应性。

医学人工智能治理综合评价指南
第 2 部分：安全评价指标

1 范围

本文件给出了医学人工智能治理综合评价中安全评价的评价指标体系框架以及评价要素的建议。

本文件适用于指导人工智能治理安全评价在各类医学场景的社会实验方案创建，反馈医学人工智能技术对社会活动产生的现实或潜在影响。

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 39725—2020 信息安全技术 健康医疗数据安全指南

DB4403/T 634—2025 医学人工智能治理综合评价指标体系

3 术语和定义

3.1

数据用益权 data usufructuary right

数据使用权代理人在转让或跨部门流转数据使用权时，通过签订协议明确利益关系、转让条件及数据复利收益分配规则，以监管数据使用权与收益权合法流转并促进医学人工智能应用开发的权益安排。

3.2

生成内容安全评价 generative content safety evaluation

在采用生成式模型辅助医疗服务供给过程中，对生成内容出现不实、违规或引发安全风险情况的评价。

4 安全评价指标体系内容

4.1 指标体系架构图

安全评价作为医学人工智能治理综合评价的一级指标[来源：T/ZSMM XXXX—XXXX 《医学人工智能治理综合评价指南 第1部分：总则》]，下设两个层级的评价指标，包括二级评价指标3个，三级评价指标22个，见图1。

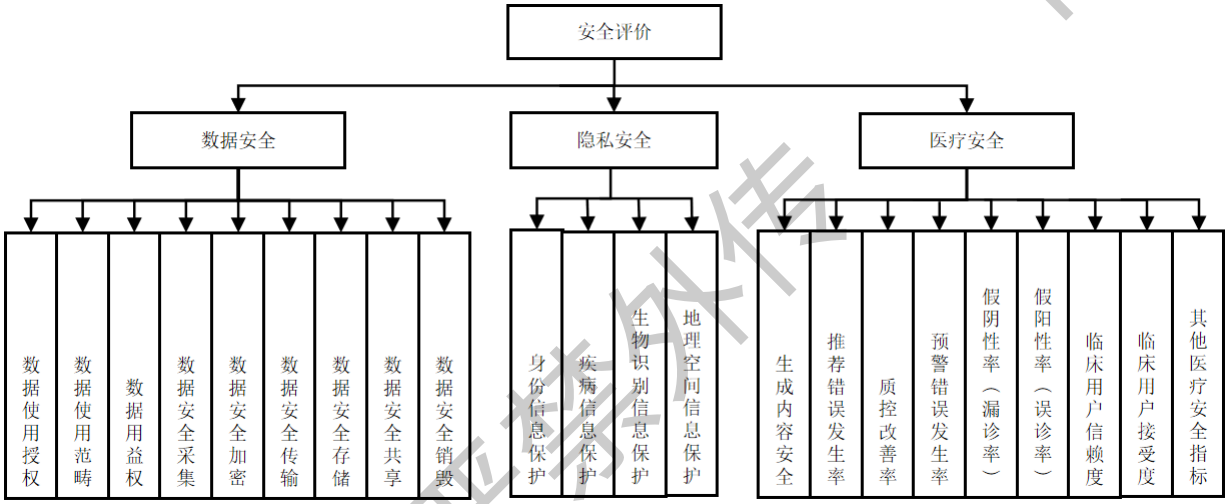

图1　医学人工智能治理综合评价中安全评价指标体系架构图

4.2　二级指标体系的内容

二级指标体系包括以下内容：

——数据安全；

——隐私安全；

——医疗安全。

4.3　三级指标体系的内容

三级指标体系可包括以下内容：

——数据使用授权；

——数据使用范畴；

——数据用益权；

——数据安全采集；

——数据安全加密；

——数据安全传输；

——数据安全存储；

——数据安全共享；

——数据安全销毁；

——身份信息保护；

——疾病信息保护；

——生物识别信息保护；

——地理空间信息保护；

——生成内容安全；

——推荐错误发生率；

——质控改善率；

——预警错误发生率；

——假阴性率（漏诊率）；

——假阳性率（误诊率）；

——临床用户接受度；

——临床用户信赖度；

——其他医疗安全指标。

5　安全评价指标内涵

5.1 二级指标

5.1.1 数据安全

数据安全是对评价对象开发与应用过程中涉及健康医疗数据安全治理的评价指标。重点关注数据的完整性、可用性以及数据处理活动的合法合规性。数据不仅包含健康医疗数据（个人健康医疗数据以及由个人健康医疗数据加工处理之后得到的健康医疗相关电子数据），还涵盖医学人工智能软件全生命周期中产生价值收益的各类数据资产，如原始数据、标注数据、模型参数等衍生数据及相关运营数据。

5.1.2 隐私安全

隐私安全是对评价对象开发与应用过程中保护个人敏感信息的处置方式进行评价的评价指标。重点关注个人敏感信息的防泄露、防滥用，以及隐私保护机制的合法合规性。

本文件述及的个人敏感信息主要包括：

a) 个人财产信息：银行账户、鉴别信息(口令)、存款信息（包括资金数量、支付收款记录等）、房产信息、信贷记录、征信信息、交易和消费记录、流水记录等，以及虚拟货币、虚拟交易、游戏类兑换码等虚拟财产信息
b) 个人健康生理信息：个人因生病医治等产生的相关记录，如病症、住院志、医嘱单、检验报告、手术及麻醉记录、护理记录、用药记录、药物食物过敏信息、生育信息、以往病史、诊治情况、家族病史、现病史、传染病史等
c) 个人生物识别信息：如个人基因、指纹、声纹、掌纹、耳廓、虹膜、面部识别特征等；
d) 个人身份信息：身份证、军官证、护照、驾驶证、工作证、社保卡、居住证等；
e) 其他信息：性取向、婚史、宗教信仰、未公开的违法犯罪记录、通信记录和内容、通讯录、好友列表、群组列表、行踪轨迹、网页浏览记录、住宿信息、精准定位信息等。

注：个人敏感信息是指一旦泄露、非法提供或滥用可能危害人身和财产安全，极易导致个人名誉、身心健康受到损害或歧视性待遇等的个人信息。

5.1.3 医疗安全

医疗安全是对评价对象在各类医学场景应用过程中医疗卫生服务质量保障进行评价的评价指标。重点关注医学人工智能辅助临床决策的准确性与可靠性、医师对医学人工智能软件的信赖与接受程度，以及对医疗质量控制的实际改善效果。

5.2 三级指标

5.2.1 数据使用授权

描述数据所有权代理人或代理机构在数据来源符合法律规定的前提下，将数据的使用权合法授予数据使用方。宜涵盖以下内容：

——算法模型所使用数据是否具备个人主体授权、数据所有权代理人或代理机构的合法授权证明；
——算法模型所使用数据的授权证明的授权过程是否合法合规，合法的授权证明包括但不限于伦理审查批件、知情同意书、数据合作协议等。涉及人工智能生成或合成数据，还宜包括其基于的原始数据授权证明及生成行为本身的合规性说明；
——算法模型所使用数据宜建立数据授权管理台账或电子日志说明，宜详细记录数据授权的范围、使用期限、特定用途限制等关键信息，重点监测使用数据是否存在扩大数据使用权限范畴的不合理情况。

注：授权使用的数据量宜满足国家与行业主管部门要求，同时在不改变该数据相关权利与义务的前提下进行使用授权。

5.2.2 数据使用范畴

描述评价对象开发、使用以及测试的过程中所使用的相关数据要符合数据属权国别的政策法规所规定的使用范畴。宜涵盖以下内容：

——算法模型数据使用范畴协议是否界定数据使用的目的、方式、范围及期限；
——算法模型数据使用范畴协议是否符合相关政策法规规范；
——算法模型在数据使用过程中是否留存系统操作日志、API 调用记录等客观证据，以追溯和监管是否存在超范畴使用情况；

——算法模型在数据使用过程中是否采取有效的数据安全保护与监管措施。

5.2.3 数据用益权

描述数据使用权代理人在转让数据使用权后签订数据用益分配的协议或合同，以促进除本评价对象外其它医学人工智能应用的开发，宜确定数据复利的收益分配规则，侧重于监管数据使用方之间开展数据使用权和收益权相互转让的合法性。宜涵盖以下内容：

——该评价对象的数据使用权代理人向不同机构转让数据使用权时是否签订数据用益分配的协议或合同，宜说明数据使用与权益分配情况、明确利益关系与数据使用权力转让的条件；

——该评价对象的数据使用权代理人在同一机构跨部门使用数据前是否征求数据持有方同意并签订数据用益权转让合同。

5.2.4 数据安全采集

描述评价对象开发、使用以及测试过程中，数据收集行为的合规性及过程安全性。宜涵盖以下内容：

——在使用过程中，针对个人信息主体主动提供个人信息行为，对其采集渠道的安全性（如加密传输）及获取授权同意的合规性说明；

——在使用过程中，针对通过交互或记录行为等自动采集数据的方式，对其采集工具（如 SDK、脚本）的安全性及采集行为的最小必要性合规说明；

——在使用过程中，针对通过共享、转让、搜集公开信息等间接获取个人信息的方式，对其数据来源的合法性审核机制及数据交接过程的安全性说明；

——宜明确数据采集的安全责任边界，如果产品或服务的提供者提供工具供个人信息主体使用，且未对产生的个人信息进行访问、回传或存储的，则不属于本文件所规制的采集行为。

5.2.5 数据安全加密

描述评价对象在开发、使用以及测试过程中进行数据传输、运算或存储时使用加密技术或隐私计算技术进行处理的技术说明。宜涵盖以下内容：

——使用过程中数据存储使用的加密技术说明；

——使用过程中数据传输使用的加密技术说明；

——使用过程中数据运算的隐私计算技术说明。

5.2.6 数据安全传输

描述评价对象在开发、使用、测试过程中涉及医疗健康数据从一个实体发送至另一个实体的过程安全性，侧重评价数据传输中断、篡改、伪造及窃取等安全风险。宜涵盖以下内容：

——开发过程的安全管理，包括保障数据传输工具的安全性、可用性及稳定性的评价。宜开展必要的源码安全审计、三方组件安全评审、渗透测试及支持库漏洞查找，并评估传输链路质量；

——建立对非网络直接传输渠道包括但不限于 U 盘、移动硬盘、光盘等移动存储介质以及纸质报告的管控与审计机制；

——采用防火墙、入侵检测等安全技术或设备，确保数据传输网络安全性；

——不同网络区域或者安全域之间宜进行安全隔离和访问控制；

——终端宜采取准入控制、终端鉴别等技术措施，防止非法或未授权终端接入内部网络；

——对通信双方进行身份确认，确保数据传输双方是可信任的；

——采用数字签名、时间戳等方式，确保数据传输的抗抵赖性；

——采用密码技术或非密码技术等方式，确保数据完整性；

——选用安全的密码算法，可使用国密序列 SM3、SM4 等，不应使用如 MD5、DES-CBC、SHA1 等不安全的算法。

5.2.7 数据安全存储

描述评价对象在应用中将医疗健康数据进行持久化保存的过程，包括但不限于采用磁盘、磁带、云存储服务、网络存储设备等载体存储数据。评价数据存储过程中可能存在的数据泄漏、篡改、丢失、不可用等安全风险。宜涵盖以下内容：

——根据数据安全级别、重要性、量级、使用频率等因素，将数据分域分级存储。其中，数据分级依据 GB/T 39725—2020 中的 6.2 有关规定划分；
——定期对防火墙、入侵检测等安全技术或设备进行风险评价，不同网络区域或者安全域之间应进行安全隔离和访问控制；
——隔离存储脱敏后的数据及其用于还原数据的恢复文件，严格审批使用恢复原始数据的技术，并留存相关审批及操作记录；
——采取密码技术、权限控制等技术措施保证 3 级及以上数据存储的完整性；
——建立数据容灾备份机制，根据数据分级情况采取本地备份或异地备份措施，并定期验证备份数据的可用性。

5.2.8 数据安全共享

描述评价对象在开发、使用以及测试过程中对数据集进行完全公开共享、受控公开共享以及领地公开共享的情况，评价数据共享在不同开放模式下流通受控程度。宜涵盖以下内容：
——所用数据集通过互联网直接公开发布的数据量的情况说明；
——所用数据集通过数据使用协议对数据的使用进行约束的数据量的情况说明；
——所用数据集在物理或者虚拟的领地范围内共享，数据不能流出到领地范围外的数据量的情况说明。

5.2.9 数据安全销毁

描述评价对象在开发、使用以及测试后，依据《个人信息保护法》的相关规定，在实现日常业务功能所涉及的系统中去除个人信息的行为，使其保持不可被检索与访问的状态，且确保销毁操作不可逆。侧重于评价个人信息数据销毁技术实现的有效性、生命周期管理的规范性、法律法规的符合度以及销毁过程的可追溯与审计。宜涵盖以下内容：
——个人信息数据销毁的技术说明；
——个人信息数据销毁的周期说明；
——个人信息数据销毁的合规说明；
——个人信息数据销毁的记录或审计日志。

5.2.10 身份信息保护

描述社会身份信息，如身份证号、医保识别号、职业、单位、家庭住址等，根据《个人信息保护法》，需对以上信息进行保护，评价特定数据类型（身份信息）的安全合规与防护能力。宜涵盖以下内容：
——在收集身份信息前，宜通过合同协议等方式，明确双方在数据安全方面的责任与义务；
——在收集身份信息前，宜通过合同协议等方式，明确数据采集范围、频次、类型、用途等；
——在收集身份信息前，应取得个人信息主体的明确同意或书面授权；
——宜对身份信息数据访问权限和实际访问控制情况进行审计，宜每半年 1 次对访问权限规则和已授权清单进行复核，及时清理已失效的账号和授权；
——身份信息数据不宜导出，确需导出宜使用加密、脱敏等技术手段防止数据泄漏，同时宜经卫生行政机构高级管理层批准，并配套数据跟踪溯源机制；
——通过数据溯源方法从算法模型中反向推导出训练数据中的个人身份信息的情况同样适用以上的规定。

5.2.11 疾病信息保护

描述患者疾病诊疗相关信息，需对该信息进行保护，评价特定数据类型（疾病信息）的安全合规与防护能力，侧重点在于评价该类数据被滥用分析与非法交易的风险。宜涵盖以下内容：
——在收集疾病信息前，宜通过合同协议等方式，明确双方在数据安全方面的责任与义务；
——在收集疾病信息前，宜通过合同协议等方式，明确数据采集范围、频次、类型、用途等；
——在收集疾病信息前，宜制定数据供应方约束机制，审查供应方的数据来源合法性证明，确保数据来源清晰、授权链条完整，防止接收非法获取的医疗数据；

——在收集疾病信息前，应事前开展个人信息保护影响评估。针对数据处理活动，检验其合法合规程度，重点判断其对患者名誉、健康、保险权益等造成损害的潜在风险，并评价保护措施的有效性；
——宜对疾病信息数据访问权限和实际访问控制情况进行审计，宜每半年 1 次对访问权限规则和已授权清单进行复核，及时清理已失效的账号和授权；
——疾病信息数据不宜导出，确需导出宜使用加密、脱敏等技术手段防止数据泄漏，同时宜经卫生行政机构高级管理层批准，并配套数据跟踪溯源机制。

5.2.12 生物识别信息保护

描述个人基因、指纹、声纹、掌纹、耳廓、虹膜、面部识别特征等，需对以上信息进行保护，评价特定数据类型（生物识别信息）的安全合规与防护能力，侧重点在于评价该类数据被用于伪造身份的风险。宜涵盖以下内容：
——在收集生物识别信息前，应向个人信息主体告知收集的必要性，并取得单独同意；在非必要场景下，应提供替代性的身份验证方式。
——在收集生物识别信息前，宜通过合同协议等方式，明确数据采集范围、频次、类型、用途等，严格限制在授权范围内使用；
——在收集生物识别信息前，宜制定数据供应方约束机制，并事前开展数据安全影响评价，针对数据处理活动，检验其合法合规程度，判断其对相关方合法权益造成损害的各种风险，以及评价相关保护措施有效性的过程；
——生物识别信息不宜存储原始图像或样本，宜仅存储摘要或特征值，同时生物识别信息应与个人身份信息（如姓名、ID）分开存储；
——宜对生物识别信息数据访问权限和实际访问控制情况进行审计，宜每半年 1 次对访问权限规则和已授权清单进行复核，及时清理已失效的账号和授权；
——生物识别信息数据原则上禁止导出，确需转移或导出的，应通过国家网信部门组织的安全评估，并采用加密通道传输。

5.2.13 地理空间信息保护

描述GPS定位地理信息、居住地址、工作单位地址、出生地等信息，需对该信息进行保护，评价特定数据类型（地理空间信息）的安全合规与防护能力，侧重点在于评价该类数据被用于重识别和行为画像的风险。宜涵盖以下内容：
——在收集地理空间信息前，宜通过协议明确数据采集的精度等级与触发频率。评价其是否严格遵循业务最小必要原则，在非导航等必要场景下，宜采用模糊化、加噪等手段降低定位精度；
——在收集地理空间信息前，明确双方责任时，应重点强调防止因实时定位泄露导致的人身安全风险的防护义务，并向用户提供显著的“仅在使用期间允许”或“关闭后台定位”的授权选项；
——在收集地理空间信息前，宜制定数据供应方约束机制，并事前开展数据安全影响评价。重点评估“时空数据重识别风险”，即检验通过长期轨迹数据反向推断用户真实身份的可能性，以及基于位置的行为画像是否对用户造成歧视性定价或诱导性营销；
——宜对地理空间信息数据访问权限和实际访问控制情况进行审计，重点审计实时定位查询与历史轨迹下载的高风险操作。宜每半年 1 次对访问权限规则和已授权清单进行复核，确保无非必要人员拥有查看用户实时行踪的权限；
——地理空间信息数据不宜长期完整留存，宜采取轨迹分段或位置与身份分离存储的措施。确需导出或汇聚大量高精度地理信息时，应评估是否构成互联网地图服务或测绘成果，并严格遵循国家地理信息安全及地图管理相关规定进行审批。

5.2.14 生成内容安全

描述在评价对象采用生成式技术辅助医疗服务供给过程中，生成式内容出现内容不实、违规或引发安全风险的情况，侧重于评价因生成式内容误导临床诊疗决策的风险。宜涵盖以下内容：

——评价因生成内容违反法律法规或医学伦理（如推荐非法疗法、建议不符合伦理的诊疗操作等），导致合规性与伦理对齐失效，从而诱导医疗服务背离法理要求及生命伦理准则的情况；

——评价因生成内容违背医学常识、伪造诊疗数据或存在逻辑错误（如虚构不存在的药物剂量或治疗方案），误导医务人员采纳虚假信息从而导致用药事故或错误临床决策的情况；

——评价因安全性护栏机制缺失，在面对诱导性攻击（如诱导开具精神类管制药物）或超出预设医疗服务范围的指令时未能有效拒绝，导致输出危害性建议，进而引发医疗服务越界与滥用的情况。

5.2.15 推荐错误发生率

描述在评价对象介入医疗服务供给过程，根据辅助决策系统推荐错误诊断、错误治疗处置、错误用药等情况的百分比，侧重于评价因算法模型的决策逻辑缺陷或特征提取偏差，导致输出的诊疗建议不符合医学临床指南、金标准或患者实际病情的情况。宜涵盖以下内容：

——该评价对象介入后，因算法误判临床特征或决策逻辑偏差，导致对患者疾病诊断结论错误的情况；

——该评价对象介入后，推荐的治疗处置方案不符合临床适应症、禁忌症要求或偏离现行诊疗规范的情况。

5.2.16 质控改善率

描述评价对象介入医疗服务供给过程后，医疗文书规范性与临床决策安全性的实际改善程度进行评价的评价指标，侧重于评价对象在关键诊疗节点上，对误诊、漏诊及不合理医疗行为的实时识别与有效阻断能力。宜涵盖以下内容：

——该评价对象介入后，某临床学科科室病历质量的提升情况，重点评价其对病历书写中逻辑不一致、数据缺失及时序矛盾等缺陷的修正能力，反映其减少因文书不规范导致的医疗纠纷隐患的情况；

——该评价对象介入后，某临床学科科室实质性医疗失误的规避情况，重点评价其通过实时预警机制，对关键决策失误（如误诊、漏诊）及未遂事件的成功拦截率，反映其降低实际医疗不良事件发生的情况。

5.2.17 预警错误发生率

描述在评价对象介入患者健康监测与管理决策过程，患者实际健康状况平稳，但根据辅助决策系统被判为需医疗处置状况的百分比，侧重于评价因系统频繁发出无效警报，导致患者或医务人员预警疲劳及医疗资源无效占用的情况。宜涵盖以下内容：

——该评价对象介入后，患者健康监测与管理过程中出现无效干预或错误启动诊疗流程的情况；

——该评价对象介入后，患者疾病诊疗处置错误预判的情况。

5.2.18 假阴性率（漏诊率）

描述在评价对象介入疾病诊断辅助决策过程，患者实际患病（金标准阳性），但根据辅助决策系统被判为无病（预测阴性）的百分比，即临床漏诊率（或第Ⅱ类错误），侧重于评价因算法未能识别病灶而导致患者丧失最佳治疗窗口、病情恶化甚至死亡的直接伤害。宜涵盖以下内容：

——金标准宜采用病理学结果或多位高级职称专家的双盲仲裁结果；

——假阴性人数与金标准阳性的比率，并宜提供相应的统计学置信区间；

——在医学人工智能软件介入疾病诊断辅助决策下，某科室某类疾病出现假阴性率的情况。

5.2.19 假阳性率（误诊率）

描述在评价对象介入疾病诊断辅助决策过程，患者实际无病（金标准阴性），但根据辅助决策系统被判为患病（预测阳性）的百分比，即临床误诊率（或第Ⅰ类错误），侧重于评价因算法过度预警而导致患者遭受不必要的侵入性检查、治疗副作用及医疗资源浪费的次生伤害。宜涵盖以下内容：

——金标准宜采用病理学结果或多位高级职称专家的双盲仲裁结果；

——假阳性人数与金标准阴性人数的比率，并宜提供相应的统计学置信区间；

——在医学人工智能软件介入疾病诊断决策过程下，某科室某类疾病出现假阳性率的情况。

5.2.20 临床用户信赖度

描述评价对象介入医疗服务供给过程后，医师基于主观体验（如感知有用性、感知易用性）对该评价对象产生的心理信任倾向与使用意愿，侧重于评价因算法厌恶和认知负荷带来安全隐患的情况。宜涵盖以下内容：

——该评价对象介入后，不同年资、不同临床学科医师对其在识别医疗风险、辅助查漏补缺方面的感知有用性评价，重点关注是否因感知价值低而导致安全辅助功能被弃用；

——该评价对象介入后，不同年资、不同临床学科医师对其交互界面与操作流程的感知易用性评价，重点关注是否因操作繁琐或提示不清增加认知负荷，从而引发医源性操作失误；

——该评价对象介入后，不同年资、不同临床学科医师的主观信任感及依赖倾向评价，重点关注是否存在盲目信任（缺乏复核）或算法厌恶（非理性拒绝）的心理预期。

5.2.21 临床用户接受度

描述评价对象介入医疗服务供给过程后，医师在实际诊疗决策行为中对评价对象推荐意见的客观执行情况，主要体现为实际采纳率及对推荐意见的修改率/否决率，侧重于评价因盲目信任与人机协同失效直接伤害的情况。宜涵盖以下内容：

——该评价对象介入后，不同年资、不同临床学科医师在规定适应证范围内的实际开启率与使用活跃度，重点关注是否存在非适应证滥用带来的误导风险，或在关键风险场景下的废弃不用导致防线缺失的情况；

——该评价对象介入后，不同年资、不同临床学科医师对推荐意见的实际采纳率分布，重点关注极高采纳率（如接近 100%）背后可能掩盖的自动化偏差的情况；

——该评价对象介入后，不同年资、不同临床学科医师对推荐意见的修改率及否决合理性，重点关注因修改率过高反映出的算法低特异性与警报疲劳风险，以及医师对高风险预警信号的异常忽略的情况。

5.2.22 其他医疗安全指标

描述除本文件给出的医疗安全评价指标外，根据特定医疗应用场景、具体评价对象特性或特殊临床安全要求需增设的补充性评价指标。侧重于评价医学人工智能在特定细分领域或特殊技术情境下可能引发的非通用性医疗安全风险。宜涵盖以下内容：

——针对涉及硬件控制或植入操作的医学人工智能（如手术机器人、智能假体），评价其与特定医疗器械、手术材料的物理兼容性、操作精度及故障引发的物理伤害风险；

——针对涉及药物推荐或治疗规划的医学人工智能，评价其在复杂用药组合下的药物相互作用预警能力，以及对禁忌症、过敏史的逻辑校验能力；

——针对服务对象为儿童、孕产妇、老年人或精神障碍患者等特殊群体的医学人工智能，评价其算法模型在特定生理参数范围、认知交互模式等方面的适配性与安全性。

参 考 文 献

[1] GB/T 20001.8—2023 标准起草规则 第8部分：评价标准
[2] GB/T 35273—2020 信息安全技术 个人信息安全规范
[3] GB/T 37373—2019 智能交通 数据安全服务
[4] GB/T 39725—2020 信息安全技术 健康医疗数据安全指南
[5] GB/T 41867—2022 信息技术 人工智能 术语
[6] GB/T 44811—2024 物联网 数据质量评价方法
[7] GB/T 45288.2—2025 人工智能 大模型 第2部分：评测指标与方法
[8] GB/T 45288.3—2025 人工智能 大模型 第3部分:服务能力成熟度评估
[9] GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求
[10] GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范
[11] GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范
[12] GB/T 45438—2025 网络安全技术 人工智能生成合成内容标识方法
[13] GB/T 46071—2025 数据安全技术 数据安全和个人信息保护社会责任指南
[14] GB/T 45577—2025 数据安全技术 数据安全风险评估方法
[15] GB/T 45574—2025 数据安全技术 敏感个人信息处理安全要求
[16] GB/T 46347—2025 人工智能 风险管理能力评估
[17] JR/T 0197—2020 金融数据安全 数据安全分级指南
[18] JR/T 0221—2021 人工智能算法金融应用评价规范
[19] JR/T 0223—2021 金融数据安全 数据生命周期安全规范
[20] JR/T 0287—2023 人工智能算法金融应用信息披露指南
[21] MZ/T 165—2020 居民家庭经济状况核对 数据安全管理要求
[22] NY/T 4261—2022 农业大数据安全管理指南
[23] YD/T 3801—2020 电信网和互联网数据安全风险评估实施方法
[24] YD/T 3865—2021 工业互联网数据安全保护要求
[25] YD/T 3956—2021 电信网和互联网数据安全评估规范
[26] YD/T 4043—2022 基于人工智能的多中心医疗数据协同分析平台参考架构
[27] YD/T 4960—2024 移动智能终端可信人工智能安全指南
[28] YY/T 1833（所有部分） 人工智能医疗器械 质量要求和评价
[29] YY/T 1858—2022 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
[30] DB11/T 2251—2024 信息安全 人工智能数据安全通用要求
[31] DB37/T 4845—2025 人工智能技术应用伦理风险的治理要求
[32] DB52/T 1726—2023 糖尿病视网膜病变人工智能筛查应用规范
[33] DB4403/T 634—2025 医学人工智能治理综合评价指标体系
[34] European Union.General Data Protection Regulation[Z].Geneva:EU,2018
[35] World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models[R]. Geneva: World Health Organization, 2024.
[36] 全国人民代表大会常务委员会.中华人民共和国个人信息保护法：主席令〔2021〕91号.2021年
[37] 全国人民代表大会常务委员会.中华人民共和国数据安全法：主席令〔2021〕84号.2021年.
[38] 国家卫生健康委员会规划与信息司，国家卫生健康委员会统计信息中心.全国医院信息化建设标准与规范（试行）：国卫办规划发〔2018〕4号，2018年
[39] 国家卫生健康委,国家中医药管理局.全国基层医疗卫生机构信息化建设标准与规范（试行）：国卫规划函〔2019〕87号.2019年
[40] 国家互联网信息办公室，中华人民共和国国家发展和改革委员会，中华人民共和国教育部，中华人民共和国科学技术部，中华人民共和国工业和信息化部，中华人民共和国公安部.生成式人工智能服务管理暂行办法：国家广播电视总局令第15号.2023年

［41］ 国家卫生健康委办公厅.关于印发医疗机构临床决策支持系统应用管理规范（试行）：国卫办医政函〔2023〕268号.2023年

［42］ 深圳市第七届人民代表大会常务委员会.深圳经济特区数据条例：深圳市第七届人民代表大会常务委员会公告（第十号）.2022年