北京市人工智能医疗器械
生产质量管理规范检查指南
（2024版）

人工智能医疗器械是指基于“医疗器械数据”，采用人工智能技术实现其预期医疗用途的医疗器械，包括第二类、第三类人工智能独立软件和含有人工智能软件组件的医疗器械(包括体外诊断医疗器械)。

本指南结合人工智能医疗器械特点，依据《医疗器械生产质量管理规范》《医疗器械生产质量管理规范附录独立软件》《医疗器械生产质量管理规范独立软件现场检查指导原则》等文件，明确了对人工智能医疗器械的生产质量管理体系要求。旨在帮助北京市医疗器械生产监管人员增强对人工智能医疗器械产品生产监管的认知，指导和规范全市医疗器械生产监管人员对人工智能医疗器械注册人、受托生产企业（以下简称“生产企业”）的监督检查工作。同时，为相关生产企业开展生产管理活动提供参考和依据。

本指南中引用的国家相关法律、法规、规章、标准、检查指南等版本发生变化时，要以当时执行的最新版为准。随着法规、强制性标准体系的不断完善以及科技能力、认知水平的不断发展，必要时，北京市药品监督管理局将重新研究修订，以确保本

指南持续符合要求。

一、机构和人员

人工智能医疗器械具有对数据和算法依赖性高的特点，数据处理和算法相关的机构和人员是关键性要素，生产企业应定义相关人员角色、明确职责和资质要求，可参照 IEEE Std 2801-2022 医学人工智能数据集质量管理推荐标准。

（一）数据处理人员：生产企业应确立一名数据管理代表全面负责数据处理相关工作。数据处理人员一般包括首席研究员、数据管理员、采集人员、初级标注人员、审核人员、仲裁人员等，标注过程中处理同一例数据时，初级标注、审核、仲裁人员之间不可相互兼任；应明确人员的职责、任职资质、选拔、培训、考核要求，如职称、工作年限、工作经验等；应有任命、培训（如培训材料、培训方案）及考核记录（如方法、频次、指标、通过准则、一致性）等；对于生产企业外部聘用人员的还应记录聘用、主要执业机构、培训等信息，明确其资质是否满足岗位要求。

（二）算法相关人员：生产企业应建立具有算法研发能力且稳定的算法团队，确立一名团队负责人全面负责算法相关工作。算法团队人员一般包括：算法研究人员、算法工程人员、算法测试人员、算法部署人员，上述人员中同一算法测试人员不可与其他角色兼任。若采用外部委托研发或者直接使用开源项目进行算法的研发，则必须有算法测试人员和部署人员，并对算法相关的质量负责。

二、厂房与设施

人工智能医疗器械的数据集是进行开发的要素，对采用自建数据集的产品，数据预处理、清洗、标注等操作的场所为真实场所或模拟场所，真实的场所应明确环境条件（如空间、照明、温度、湿度、气压等），如有特殊要求应保持相关记录，采用模拟场所情形可根据产品实际情况调整模拟程度，需详述调整理由并予以记录；对采用公开数据集、第三方数据集，则需对数据集开发方所能提供范围内的场地信息进行记录。

三、设备

生产企业应结合自身产品的实际情况，在产品生存周期过程提供充分、适宜、有效的软硬件设备、开发测试工具、网络资源以及病毒防护、数据备份与恢复等保障措施。

（一）数据集构建所用设备

1.数据采集设备管理应明确兼容性和采集特征等要求，设备的兼容性记录应包括采集设备的名称、规格型号、制造商、性能指标，若无需考虑兼容性要求应详述理由并予以记录。采集特征需明确采集设备的采集方式（如常规成像、增强成像）、采集协议（如MRI成像序列）、采集参数（如CT加载电压、加载电流、加载时间、层厚）、采集精度（如分辨率、采样率）。数据采集若使用历史数据，需列明采集设备及采集特征要求，并开展数据采集质量评估工作。

2.企业应配备执行数据集相关任务需要的资源，如访问、读

取数据、预览、检索等任务需要的软件、硬件、网络配置。测试集应配备封闭管理需要的软件、硬件、网络配置，明确管理要求。

3.数据标注应明确标注软件（包含自动标注软件）的要求，明确标注软件的名称、规格型号、完整版本、制造商、运行环境、软件确认等要求并予以记录。

4.若需使用特定的外部设备获取附加的信息（如病理结果、检验结果、数据模态转化、多模态配准、体积测量、三维打印等），设备的规格型号、计量信息（如需计量）等应确认要求并予以记录。

5.数据整理所用软件工具（含脚本）均需明确名称、规格型号、完整版本、制造商、运行环境，并进行软件确认。

（二）算法研发所用设备

1.应明确定义并记录进行算法训练、算法测试、算法部署所用到服务器算力的典型配置（如GPU型号和数量、CPU型号和数量、内存大小、网络带宽等）。

2.应明确定义并记录算法训练所用的操作系统、开发环境（如编程语言及版本、集成开发环境及版本、web服务及版本、支持软件及版本等）、算法架构（如基础计算包、GPU指令集、集成开发环境插件等）。

3.应明确定义并记录算法测试所用的操作系统、开发环境、算法框架、基础服务等，若与其他外部设备进行配合或作为软件组件集成到其他医疗器械中，明确说明外部设备和器械的规格型

号。

4.应明确定义并记录算法部署运行的操作系统、开发环境、算法框架、基础服务、虚拟机、应用容器引擎等。

四、设计开发

生产企业应结合质量管理体系要求，建立人工智能医疗器械生存周期过程，开展与软件安全性级别相匹配的产品质量保证工作，确定需求分析、数据收集、算法设计、验证与确认、部署运行、更新控制等活动要求，将风险管理、可追溯分析（需包含算法和数据）贯穿于生存周期全程，形成记录。

（一）需求分析

1.需求分析应当以用户需求与风险为导向，结合产品的预期用途、使用场景、核心功能，综合分析法律、法规、规章、标准、用户、产品、功能、性能、接口、用户界面、网络安全、警示提示等需求，明确风险管理、可追溯性分析、数据收集、算法性能、使用限制、现成软件（现成算法）的验证与确认等活动要求，形成相应文件。

2.数据收集应当确保数据来源的合规性、充分性和多样性，数据分布的科学性和合理性，数据质控的充分性、有效性和准确性。

3.算法性能需结合医疗实际和产品定位，明确假阴性与假阳性、重复性与再现性、鲁棒性/健壮性、实时性等性能指标的适用性及其要求，并兼顾不同性能指标的制约关系。

4.使用限制需考虑产品禁用、慎用等场景，准确表述产品使用场景，提供必要警示提示信息。

（二）数据收集

1.数据采集

数据采集应当明确采集设备、采集过程、数据脱敏等质控要求，并建立数据采集操作规范。若使用历史数据，应当结合样本规模、采集难度等影响因素合理选择数据采集方式，明确数据筛选标准并对采集的数据进行质量评估。

采集的数据应进行数据脱敏以保护患者隐私，数据脱敏需明确脱敏的类型（静态、动态）、规则、方法以及脱敏内容的确定依据。如使用历史数据，企业接收的数据应为脱敏后的数据，不得有敏感数据流入企业。

2.数据整理

数据整理应基于原始数据库明确数据清洗、数据预处理的质控要求。数据清洗应当明确清洗的规则、方法、结果，数据预处理应当明确处理的方法（如滤波、增强、重采样、尺寸裁剪、均一化等）、结果。数据经整理后形成基础数据库，需明确样本类型、样本量、样本分布等信息。

3.数据标注

数据标注应当明确标注资源管理、标注过程质控、标注质量评估等要求，并建立数据标注操作规范。标注过程质控包括人员职责（如人员资质、人员数量、职责分工）、标注规则（如临床

指南、专家共识、专家评议、文献分析）、标注流程（如标注对象、标注形式、标注轮次、标注步骤、结果审核）、分歧处理（如仲裁人员、仲裁方式）、可追溯性（如数据、操作）等要求。数据经标注后形成标注数据库，样本类型可分为数据块（如图像区域、数据片段）、单一数据（由多个数据块组成）、数据序列（由多个单一数据组成）。标注数据库的样本量、样本分布等要求及风险考量与基础数据库相同。

数据标注若使用自动标注软件，结果不得直接使用，应由标注人员审核后方可使用。

4.数据集构建

基于标注数据库构建训练集、调优集（若有）、测试集，应当明确训练集、调优集、测试集的划分方法、划分依据、数据分配比例。训练集原则上需保证样本分布具有均衡性，测试集、调优集原则上需保证样本分布符合真实情况，训练集、调优集、测试集的样本应两两无交集并通过查重予以验证。

数据扩增（若有）应当明确扩增的对象、范围、方式（离线、在线）、方法（如翻转、旋转、镜像、平移、缩放、滤波、生成对抗网络等）、倍数，在线扩增亦需予以记录，扩增需考虑数据偏倚的影响及风险。原则上不得对测试集进行数据扩增，对抗测试除外。

数据经扩增后应当形成扩增数据库，需列表对比扩增数据库与标注数据库在样本量、样本分布（注明扩增倍数）等差异，以

证实扩增数据库样本量的充分性以及样本分布的合理性。

（三）算法设计

1.算法选择

算法选择应当提供所用算法的名称、类型（如有监督学习、无监督学习，基于模型、基于数据，白盒、黑盒）、结构（如层数、参数规模）、输入输出数据类型、流程图、算法编程框架、运行环境等基本信息，并明确算法选用依据，包括选用的理由和基本原则。若组合使用集成学习、迁移学习、强化学习等，需提供算法基本信息以及算法选用依据。

2.算法训练

算法训练应当基于训练集、调优集进行训练和调优，考虑评估指标、训练方式、训练目标、调优方式、训练数据量—评估指标曲线等要求。

3.算法性能评估

算法性能评估应当基于测试集对算法设计结果进行评估，综合评估假阴性与假阳性、重复性与再现性、鲁棒性/健壮性、实时性等适用性，以证实算法性能满足算法设计目标，并作为软件验证、软件确认的基础。若使用第三方数据库开展算法性能评估，应当提供第三方数据库的基本信息（如名称、创建者、数据总量等）和使用情况（如测试数据样本量、评估指标、评估结果等）。

对于黑盒算法，应开展算法性能影响因素分析，并提供算法性能影响因素分析报告，明确影响算法性能的主要因素及其影响

程度，以及产品使用限制和必要警示提示信息。

（四）验证与确认

1.软件验证

软件验证应当基于软件需求予以开展，保证软件的安全有效性，并作为软件确认的基础。

2.软件确认

软件确认测试应当基于用户需求，由预期用户在真实或模拟使用场景下予以开展，涵盖现成软件、网络安全的测试要求，确定缺陷管理、风险管理、可追溯性分析、评审等活动要求，形成用户测试记录、测试报告以及评审记录并经批准，适时更新并经批准。可追溯性分析此时应当分析用户测试与用户需求、用户测试与风险管理的关系。同时，开展算法性能比较分析，若各类测试场景（含临床评价）算法性能变异度较大，详述原因并基于分析结果明确产品使用限制和必要警示提示信息。最后，结合算法训练、算法性能评估、临床评价等结果开展算法性能综合评价，针对训练样本量和测试样本量过少、测试结果明显低于算法设计目标、算法性能变异度过大等情况，对产品的适用范围、使用场景、核心功能进行必要限制。

基于测评数据库开展的，除满足数据库通用要求（如数据管理、网络安全与数据安全、可扩展性）外，还应满足权威性、科学性、规范性、多样性、封闭性和动态性要求。不应使用公开数据库作为测评数据库。

（五）部署运行

算法发布和更新时应在相关文件列明算法关键模块的功能、接口、版本、存储形式（如 pt、pth、bin、onnx、pb、keras、ckpt、pkl 等）；主要功能组件模块及相互依赖和接口关系；软件的前后端部署方式；对基础软件和硬件的依赖和兼容性要求等。

（六）更新控制

人工智能医疗器械若发生算法更新、软件更新，均应当按照质量管理体系的要求，开展与算法更新、软件更新的类型、内容和程度相适宜的验证与确认活动，将风险管理、可追溯分析贯穿于更新全程，形成记录。此外，算法更新、软件更新均需考虑引入回滚机制，以保证医疗业务的连续性，特别是对风险较高的软件。

软件版本控制应当基于合规性要求确定软件版本命名规则，涵盖自研软件、现成软件、算法（算法驱动型更新或数据驱动型更新）网络安全的全部软件更新类型，明确并区分重大软件更新和轻微软件更新，并符合软件版本命名规则的要求。

对数据集进行用途（如训练、调优、测试、验证等）的变更，应按照数据集管理体系的要求进行确认形成记录。高控制等级的数据集停用后数据可流入低控制等级的数据集，不允许低控制等级的数据集向高控制等级流动（测试集数据可变更为训练和调优使用，不可将训练集、调优集的数据变更为测试使用）。数据集数据的变更，应按照建立数据集过程相同的质量体系进行管理，

并记录更新内容和版本变更。

（七）网络与数据安全

人工智能医疗器械全生命周期过程中应当考虑网络安全与数据安全问题，对网络与数据安全过程的控制要求形成文件，包括数据转移要求，数据整理、数据集构建、算法训练等内部活动开展过程中的数据污染防护措施，以及数据标注、软件确认等涉及外部活动开展过程中的数据污染防护措施及数据接口要求。

各数据库（集）均需进行数据备份，明确备份的方法、频次、数据恢复方法。

数据库和数据集访问应明确授权访问管理要求，形成文件及记录。

人工智能医疗器械软件应明确算法的软件安全性级别（轻微、中等、严重）并详述判定理由。应形成算法风险管理资料，明确过拟合与欠拟合、假阴性与假阳性、数据污染与数据偏倚（如数据扩增引入的偏倚）、中外差异等风险的控制措施。

（八）可追溯性分析

可追溯性分析应当建立控制程序，涵盖软件、现成软件、算法及数据、网络安全的控制要求，形成软件及算法的可追溯性分析报告。使用可追溯性分析工具保证软件开发、软件更新过程满足可追溯性要求，并贯穿于软件生存周期全过程。提供算法可追溯性分析报告等相关文件及记录，即追溯算法需求、算法设计、算法实现、算法验证与确认、风险管理、数据集的管理。若无单

独文档可提供软件可追溯性分析报告，并注明算法可追溯性分析所在位置。

五、采购

生产企业应确保采购物品符合法律法规的相关规定和国家强制性标准的相关要求，包括数据来源及以开源方式获得的软件等。

数据来源应当提供数据来源的合规性声明，列明数据来源机构名称、所在地域、数据收集量、伦理批件（或科研合作协议）编号等信息。

产品部署运行所需的软硬件，应当记录采购信息，其中以开源方式获得的软件组件、基础包、基础软件、集成环境等，应列明来源、下载地址、版本、开源协议等信息。

六、销售和售后服务

人工智能医疗器械软件在产品设计具有持续学习/自适应学习能力的，需要在软件部署阶段确认自学习功能已关闭，并提供记录。

七、不良事件监测、分析和改进

上市后需要持续开展算法泛化能力研究的，需要结合用户投诉、不良事件和召回等情况识别前期未预见的风险，并采取有效的风险控制措施将风险降至可接受水平。此外，根据产品更新需求，经评估后实施更新活动，开展与之相适宜的验证与确认活动，保证算法泛化能力持续满足用户需求。

附件：1.设计开发检查要点举例说明

2.人工智能医疗器械常用术语

3.参考文献

附件 1

设计开发检查要点举例说明

1.算法策划

算法开发策划阶段，应对算法需求、数据收集、算法设计、验证与确认、算法更新、风险管理、可追溯性分析等过程进行算法生命周期策划，输出《算法开发计划书》。

2.算法需求

算法需求分析以用户需求与风险为导向，结合产品的预期用途、使用场景、核心功能，综合分析法律、法规、规章、标准、用户、产品、功能、性能、接口、用户界面、网络安全、警示提示等需求，重点考虑数据收集、算法性能、使用限制等要求。输出《算法需求规范》。

3.数据收集

3.1 数据入选、排除标准

在《算法需求规范》中明确算法数据入选、排除标准。

3.2 数据来源及质控

数据收集应确保数据来源的合规性，数据质控的充分性、有效性、准确性。

3.3 数据采集

数据采集需考虑采集设备、采集过程、数据脱敏等质控要求，

并建立《数据采集操作规范》。

3.4 数据整理

脱敏数据汇总形成原始数据库，不同模态的数据在原始数据库中需加以区分。数据整理基于原始数据库考虑数据清洗、数据预处理的质控要求。输出《数据整理规范》，同时《数据整理规范》中需明确样本类型、样本量、样品分布等信息，数据经整理后形成基础数据库。

3.5 数据标注

3.5.1 数据标注前应建立《数据标注操作规范》，明确标注资源管理、标注过程质控、标注质量评估等要求。数据经标注后形成标注数据库。

3.5.2 数据标注可使用自动标注软件，但自动标注结果不得直接使用，应由标注人员审核后方可使用；同时，自动标注软件亦需明确名称、型号规格、完整版本、制造商、运行环境等信息，并进行软件确认。

3.6 数据集构建

3.6.1 基于标注数据库构建训练集、调优集、测试集，明确训练集、调优集、测试集的划分方法、划分依据、数据分配比例，输出《数据集构建标准》。

3.6.2 训练集应当保证样本分布具有均衡性，测试集、调优集应当保证样本分布符合临床实际情况，训练集、调优集、测试集的样本应当两两无交集并通过查重予以验证。

3.6.3 为解决样本分布不满足预期的问题，可对训练集、调优集小样本量数据进行扩增，原则上不得对测试集进行数据扩增，对抗测试除外。数据扩增需明确扩增的对象、方式（离线、在线）、方法（如翻转、旋转、镜像、平移、缩放、滤波、生成对抗网络等）、倍数，扩增倍数过大应考虑数据偏倚的影响及风险。若采用生成对抗网络进行数据扩增，需明确算法基本信息以及算法选用依据。

3.6.4 数据经扩增后形成扩增数据库，需列表对比扩增数据库与标注数据库在样本量、样本分布（注明扩增倍数）等差异，以证实扩增数据库样本量的充分性以及样本分布的合理性。

3.7 数据库管理

3.7.1 数据库管理应遵循真实性、完整性、可用性、合规性、可追溯性、临床代表性、时效性、安全性、准确性 9 大原则。

3.7.2 应定期对逻辑数据库的数据以及文件数据进行备份，备份文件保存在不同机架的机器磁盘上以提高备份的安全性。

3.7.3 在整个数据转移过程中，应当明确数据转移方法、数据防污染措施以及数据销毁方式。

4.算法设计

人工智能算法设计主要考虑算法选择、算法目标设定分析、算法训练、算法性能评估等要求，形成《算法设计说明书》。对于黑盒算法，算法设计应开展算法性能影响因素分析，同时与现有医学知识建立关联，以提升算法可解释性。

5.验证与确认

算法训练过程中，需要明确算法训练环境，应当对算法进行性能评估，以确保选择的算法准确、有效。算法验证阶段，明确算法性能评估环境，应完成算法性能指标评估、压力测试、对抗测试，黑盒需要算法性能影响因素分析，输出《算法性能评估报告》。

同时，开展算法性能比较分析，详述各类测试场景（含临床评价）算法性能变异度较大的原因，基于分析结果明确产品使用限制和必要警示提示信息，输出《算法性能比较分析报告》。

最后，结合算法训练、算法性能评估、临床评价等结果开展算法性能综合评价，针对训练样本量和测试样本量过少、测试结果明显低于算法设计目标、算法性能变异度过大等情况，对产品的适用范围、使用场景、核心功能进行必要限制。输出《算法性能综合评价报告》。

6.算法风险管理

人工智能医疗器械的软件安全性级别可基于产品的预期用途、使用场景、核心功能进行综合判定，并开展风险管理活动，采取有效的风险控制措施将风险降至可接受水平，并贯穿于人工智能医疗器械全生命周期过程。

人工智能医疗器械的主要风险从算法角度包括过拟合和欠拟合。从用途角度，辅助决策主要包括假阴性和假阳性，其中假阴性即漏诊，可能导致后续诊疗活动延误，特别是要考虑快速进

展疾病的诊疗活动延误风险，而假阳性即误诊，可能导致后续不必要的诊疗活动；非辅助决策从算法设计目标能否得以实现角度，亦可参考辅助决策分为假阴性和假阳性。应输出《风险管理报告》，明确过拟合与欠拟合、假阴性与假阳性、数据扩增与数据偏倚等风险的控制措施。

7.算法可追溯性分析

医疗器械全生命周期管理中，应实现算法的可追溯性，并形成算法可追溯性分析报告等相关文件及记录，即追溯算法需求、算法设计、算法实现、算法验证与确认、风险管理、数据集的管理。

在数据收集过程中，数据采集、数据整理、数据标注、数据集构建过程中形成《数据脱敏交接记录》《数据清洗记录》《数据标注记录》《数据审核记录》《数据仲裁记录》《数据集构建记录》，以上记录均由操作人员签字确认。

数据集管理过程中，每一例数据都可追溯到唯一识别号、脱敏人员、清洗人员、标注人员、审核人员、仲裁人员、入库人员，保证了数据收集各环节的数据和操作人员的可追溯。

8.算法更新控制

人工智能医疗器械若发生算法更新、软件更新，均应当按照质量管理体系的要求，开展与算法更新、软件更新的类型、内容和程度相适宜的验证与确认活动，将风险管理、可追溯分析贯穿于更新全程，形成记录以供体系核查。

对于算法更新，无论算法驱动型更新还是数据驱动型更新，均应开展算法性能评估、临床评价等验证与确认活动，以保证算法更新的安全有效性。对于软件更新，具体要求详见医疗器械软件指导原则、医疗器械独立软件生产质量现场检查指导原则。

人工智能医疗器械所含的每个人工智能算法，均应独立开展需求分析、数据收集、算法设计、验证与确认、更新控制等活动，同时考虑人工智能算法组合的整体评价要求，以保证产品的安全有效性。

附件 2

人工智能医疗器械常用术语

人工智能 artificial intelligence（AI）: 表现出与人类智能（如推理和学习）相关的各种功能的功能单元的能力。

人工智能医疗器械 artificial intelligence medical device（AIMD）: 采用 AI 技术实现其预期用途的医疗器械。

注 1: 如采用机器学习、模式识别、规则推理等技术实现其医疗用途的独立软件。

注 2: 如采用内嵌 AI 算法、AI 芯片实现其医疗用途的医疗器械。

机器学习 machine learning: 功能单元通过获取新知识或技能，或通过整理已有的知识或技能来改进其性能的过程。

注：也可称为自动学习。

深度学习 deep learning： 通过训练具有多个隐层的神经网络来获得输入输出间映射关系的机器学习方法。

训练 training: 基于机器学习算法，利用训练数据，建立或改进机器学习模型参数的过程。

监督学习 supervised learning: 一种学习策略，获得的知识的正确性通过来自外部知识源的反馈加以测试的学习策略。

注：也可称为监督式学习。

无监督学习 unsupervised learning: 一种学习策略，它在于观察并分析不同的实体以及确定某些子集能分组到一定的类别里，而无需在获得的知识上通过来自外部知识源的反馈，以实现任何正确性测试。

注 1: 一旦形成概念，就对它给出名称，该名称就可以用于其他概念的后续学习了;

注 2: 也可称为无师（式）学习。

强化学习 reinforcement learning: 一种学习策略，它强调从环境状态到动作映射的过程，目标是使动作从环境中获得的累积奖赏值最大。

集成学习 ensemble learning: 通过结合多个学习器来解决问题的一种机器学习范式。

注: 其常见形式是利用一个基学习算法从训练集产生多个基学习器，然后通过投票等机制将基学习器进行结合。

迁移学习 transfer learning

利用一个学习领域 A 上有关学习问题 T(A)的知识，改进学习领域 B 上相关学习问题 T(B)的学习算法的性能。

过拟合 overfitting: 学习器对训练样本过度学习，导致训练样本中不具有普遍性的模式被学习器当作一般规律，降低了泛化性能；典型表现是训练集上的性能越高，测试集上的性能越低。

欠拟合 underfitting: 学习器对训练样本学习不充分，导致训练样本中包含的重要模式没有被学习器获取，降低了泛化性能;

典型表现是训练集上的性能可以继续提高，测试集上的性能同时得以提高。

人工智能医疗器械生存周期模型 AIMD lifecycle model: 人工智能医疗器械从起始到退役的整个演进过程的框架。

注 1: 包括: 需求分析，设计与开发，验证与确认，部署，运维与监控，再评价直至停运。

注 2: 在人工智能医疗器械生存周期中，某些活动可出现在不同的过程中，个别过程可重复出现。例如为了修复系统的隐错和更新系统，需要反复实施开发过程和部署过程。

数据 data: 信息的可再解释的形式化表示，以适用于通信、解释或处理。

注: 可以通过人工或自动手段处理数据。

个人敏感数据 personal sensitive data: 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。

注: 个人敏感信息包括身份证件号码、个人生物识别信息、银行账号、通信记录和内容、财产信息、征信信息、行踪轨迹、住宿信息、健康生理信息、交易信息、14 岁以下（含）儿童的个人信息等。

健康数据 health data: 与身体或心理健康相关的个人敏感数据。

注: 由于目前全球规定了不同的隐私合规性法律和法规。例

如，在欧洲，可能需要采取的要求和参考变更为“个人数据”和“敏感数据”，在美国，健康数据可能会变更为“受保护的健康信息（PHI）”，这需要不同国家或地区的制造商进一步考虑中国当地的法律或法规。

数据集 data set: 具有一定主题，可以标识并可以被计算机化处理的数据集合。

训练集 training set: 用于训练人工智能算法的数据集，其外部知识源可用于算法参数的计算。

调优集 tuning set: 用于优化人工智能算法的数据集，其外部知识源可用于算法超参数的选择。

注：为避免与医疗器械领域所用术语“确认”进行区分，这里不使用通用人工智能领域的 validation set，二者含义一致。

测试集 testing set:

用于测试人工智能算法性能的数据集，其外部知识源可用于对算法的评估。

参考标准 reference standard: 筛查、诊断和治疗过程或基于标注过程建立的基准。

注：参考标准可包含疾病、生理状态或生理异常以及位置和程度等信息标签。

金标准 gold standard: 筛查、诊断和治疗可依据的最佳参考标准。

数据清洗 data cleaning: 检测和修正数据集合中错误数据

项的预处理过程。

数据采集 data acquisition: 数据由生成装置按照数据采集规范生成，以数字化格式存储并传输到目标系统的过程。

数据脱敏 data masking: 通过去标识化或匿名化，实现对个人敏感信息的可靠保护。

数据标注 data annotation: 对数据进行分析，添加外部知识的过程。

仲裁 arbitration: 多名标注人员对同一原始数据的标注结果不一致时用于决定最终结果的过程。

软件质量 software quality: 在规定条件下使用时，软件产品满足明确或隐含要求的能力。

软件质量保证 software quality assurance:

a)为使某项目或产品遵循已建立的技术需求提供足够的置信度，而必须采取的有计划的和有系统的全部动作的模式。

b)设计以估算产品开发或制造过程的一组活动。

可靠性 reliability: 在规定时间间隔内和规定条件下，系统或部件执行所要求功能的能力。

完整性 integrity: 保护数据准确性和完备性的性质。

一致性 consistency: 在数据集的各阶段、部分之间，一致、标准化、无矛盾的程度。

重复性 repeatability: 由同一操作员按相同的方法、使用相同的测试或测量设施、在短时间间隔内对同一测试/测量对象

进行测试/测量，所获得的独立测试/测量结果间的一致程度。

再现性 reproducibility: 由不同的操作员按相同的方法，使用不同的测试或测量设施，对同一测试/测量对象进行观测以获得独立测试/测量结果，所获得的独立测试/测量结果间的一致程度。

可达性 accessibility: 组成软件的各部分便于选择使用或维护的程度。

可得性 availability:

a）软件（系统或部件）在投入使用时可操作或可访问的程度或能实现其制定系统功能的概率;

b）系统正常工作时间和总的运行时间之比;

c）在运行时，某一配置项实现指定功能的能力。

保密性 confidentiality: 数据对未授权的个人、实体或过程不可用或不泄露的特性。

网络安全 cybersecurity: 通过采取必要措施，防范对数据、模型等攻击、侵入、干扰、破坏和非法使用以及意外事故，使设备处于稳定可靠运行的状态，以及保障数据、模型等的完整性、保密性、可得性的能力。

安全性 safety: 免除于不可接受的风险。

鲁棒性/稳健性: 在存在无效输入或急迫的环境条件下，系统或部件其功能正确的程度。

泛化能力 generalizability: 机器学习算法对陌生样本的适应

能力。

可追溯性 traceability: 系统对其决策过程及输出进行记录的特性。

公平性 fairness: 系统做出不涉及喜好和偏袒决策的性质。

可解释性 explainability: 以人能理解的方式，对系统决策因素进行说明的能力。

黑盒测试 black-box testing: 忽略系统或部件的内部机制只集中于响应所选择的输入和执行条件产生的输出的一种测试。

白盒测试 glass-box testing: 侧重于系统或部件内部机制的测试。类型包括分支测试、路径测试、语句测试等。

对抗[措施] countermeasure: 为减小脆弱性而采用的行动、装置、过程、技术或其他措施。

对抗样本 adversarial sample: 基于原始数据上添加扰动达到混淆系统判别目的新样本。

对抗测试 adversarial test: 使用对抗性样本开展的测试，或采用不同目标样本分布的特选数据作为压力数据集进行的测试。

阳性样本 positive sample: 由参考标准确定为带有某一种或几种特定特征的样本。

阴性样本 negative sample: 除阳性样本以外的样本。

真阳性 true positive(TP): 被算法判为阳性的阳性样本。

假阳性 false positive(FP): 被算法判为阳性的阴性样本。

真阴性 true negative(TN): 被算法判为阴性的阴性样本。

假阴性 false negative(FN): 被算法判为阴性的阳性样本。

目标区域 target region: 在影像评价中，根据参考标准从原始数据中划分出的若干个包含特定类别目标的最小数据子集(子集元素为像素，体素等)。

分割区域 segmentation region: 在影像评价中，从原始数据中划分出的若干个包含特定类别目标的最小数据子集(子集元素为像素，体素等)。

病变定位 lesion localization: 算法检出病变位置正确标识出参考标准确定的病变位置。

非病变定位 non-lesion localization: 算法检出病变位置未能正确标识出参考标准确定的病变所在位置。

病变定位率 lesion localization rate: 病变定位数量占由参考标准确定的全体病变数量的比例。

非病变定位率 non-lesion localization rate: 非病变定位数量占全体病例数量的比例，非病变定位率可以大于 1。

假阳性率 false positive rate: 假阳性病例数量（阴性病例中包含非病变定位）占全部阴性病例数量的比例。

灵敏度 sensitivity

召回率（查全率）recall:真阳性样本占全体阳性样本的比例。

特异度 specificity:真阴性样本占全体阴性样本的比例。

漏检率 miss rate:1 减去灵敏度。

精确度（查准率） precision

阳性预测值 positive prediction value: 真阳性样本占被算法判为阳性样本的比例。

阴性预测值 negative prediction value: 真阴性样本占被算法判为阴性样本的比例。

准确率 accuracy: 算法判断正确的样本占全体样本的比例。

F_1度量 F_1-measure: 召回率和精确度的调和平均数。

约登指数 Youden index: 灵敏度与特异度之和减去 1。

受试者操作特征曲线 receiver operating characteristics curve（ROC curve）: 以假阳性率为横坐标、真阳性率为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

曲线下面积 area under curve(AUC):曲线下与坐标轴围成的积分面积。

自由响应受试者操作特征曲线 free-response receiver operating characteristics curve(fROC): 以非病变定位率为横坐标、病变定位率为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

候选自由受试者操作特征曲线 alternative free receiver operating characteristics curve(AFROC curve): 以假阳性率为横坐标、病变定位率为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

精确度-召回率曲线 precision-recall curve(P-R): 以召回率

为横坐标、精确度为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

平均精确度 average precision(AP): 精确度-召回率曲线下与坐标轴围成的积分面积。

平均精确度均值 mean average precision(MAP): 在多目标检测问题上，算法对于各类目标的平均精确度的平均值。

交并比 intersection over union(IoU): 分割区域与目标区域的交集占分割区域与目标区域并集的比例

注：也可称为 Jaccard 系数。

Dice 系数 Dice coefficient: 分割区域与目标区域的交集占分割区域与目标区域平均值的比例。

中心点距离 central distance: 分割区域中心与目标区域中心的距离，该指标反映两个集合的接近程度。

混淆矩阵 confusion matrix: 一种矩阵，它按一组规则记录试探性实例的正确分类和不正确分类的个数。

注 1：通常矩阵的列代表人工智能的分类结果，而矩阵的行代表参考标准的分类结果;

注 2：也可称为含混矩阵。

Kappa 系数 Kappa coefficient: 一种用于评价结果一致性的指标。

信噪比 signal-to-noise ratio(SNR): 信号平均功率水平与噪声平均功率水平的比值。

峰值信噪比 peak signal-to-noise ratio: :信号最大可能功率与噪声平均功率水平的比值。

结构相似性 structural similarity: 是一种衡量两幅图像相似度的指标。

余弦相似度 cosine similarity: 通过测量两个向量的夹角的余弦值来度量它们之间的相似性。

困惑度 perplexity: 度量概率分布或概率模型的预测结果与样本的契合程度，困惑度越低则契合越准确。

字错率 word error rate: 将识别出来的字需要进行修改的字数与总字数的比值。

交叉熵 cross-entropy: 一种度量两个概率分布之间差异的指标。

互信息 mutual information: 对两个随机变量间相互依赖性的量度。

服务可用性 service availability: 服务客户发起服务请求后，服务可访问的时间占总服务时间的比例。

注：服务可用性的计算是在一系列预定义的时间段中，服务可用时间之和占预定义时间段之和的比例，可排除允许的服务不可用时间。

附件3

参考文献

[1]医疗器械生产质量管理规范附录独立软件

[2]医疗器械生产质量管理规范独立软件现场检查指导原则

[3]人工智能医疗器械注册审查指导原则

[4]医疗器械软件注册审查指导原则（2022 年修订版）

[5]医疗器械网络安全注册审查指导原则（2022 年修订版）

[6]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

[7]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

[8]YY/T 1833.3-2022 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

[9]YY/T 1833.4-2023 人工智能医疗器械 质量要求和评价 第 4 部分：可追溯性

[10]YY/T XXXX.X-XXXX《人工智能医疗器械 质量要求和评价 第 5 部分：预训练模型》征求意见稿

[11]GB/T 42061—2022 医疗器械 质量管理体系 用于法规的要求[S]

[12]GB/T 42062—2022 医疗器械 风险管理对医疗器械的

应用[S]

[13]YY/T 0664-2020 医疗器械软件 软件生存周期过程[S]

[14]IEEE Std 2801-2022 Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence 医学人工智能数据集质量管理推荐标准