《医学人工智能治理综合评价指南 第2部分：安全评价》（征求意见稿）编制说明

《医学人工智能治理综合评价指南　第2部分：安全评价》

标准编制组

二〇二五年十一月

目 录

一、 编制的目的和意义

（一）研究背景

如今，人工智能技术飞速发展，与医学领域的深度融合与应用，在提升诊疗效率、促进精准医疗的同时，也引发了复杂的数据安全、隐私泄露和医疗风险等治理挑战。我国高度重视人工智能的安全与可控发展。2017 年国务院印发的《新一代人工智能发展规划》明确提出要建立人工智能安全评估与管控能力。后续出台的《生成式人工智能服务管理暂行办法》（2023 年）与《全球人工智能治理倡议》（2023 年）进一步强调了发展必须与安全治理并重，要求采取有效措施防范风险。至 2025 年，《人工智能安全治理框架》2.0 版的发布，标志着我国人工智能治理从原则共识迈向实操深化的新阶段；同期《国务院关于深入实施“人工智能+”行动的意见》则明确将“安全可控”作为推动人工智能与各行业深度融合的基本要求。这一系列顶层设计为开展医学人工智能安全评价提供了根本遵循与政策依据。

在国际层面，人工智能安全风险已演变为全球性挑战。斯坦福大学《2025 人工智能指数报告》显示，相比 2023 年 2024 年全球公开报道的人工智能安全事件数量激增 56.4%，凸显了安全治理的紧迫性。世界卫生组织(WHO)于 2024 年发布《Ethics and governance of artificial intelligence for health:

Guidance on large multi-modal models》，明确指出其在医疗应用中的独特风险（如数据偏见、隐私泄露、诊断错误等），并系统提出了确保安全、可信与问责的核心原则与治理措施。这为我国构建本土化的医学人工智能安全评价框架提供了重要的国际视野与专业参考。

我国已初步构建了覆盖网络安全、数据安全、个人信息保护的人工智能标准体系。国家标准如《网络安全技术 生成式人工智能服务安全基本要求》（GB/T 45654-2025）、《信息安全技术 个人信息安全规范》（GB/T 35273-2020）《数据安全技术 数据安全和个人信息保护社会责任指南》（GB/T 46071—2025）等，地方标准如《医学人工智能治理综合评价指标体系》(DB4403/T 634—2025)和《信息安全 人工智能数据安全通用要求》（DB11/T 2251—2024）及行业标准《电信网和互联网数据安全风险评估实施方法》（YD/T 3801—2020）、《金融数据安全 数据安全分级指南》（JR/T 0197—2020）等共同为医学人工智能安全评估奠定了基础。

然而，现有标准在医学这一高风险特殊领域的适配性、精细度和系统性方面仍显不足。面对医学人工智能在数据全生命周期管理、敏感隐私信息保护以及临床应用可靠性等方面的严峻挑战，亟需一部专门针对医学场景的安全评价标准，在此背景下，制定《医学人工智能治理综合评价指南 第2部分：安全评价》势在必行。本部分作为该系列标准中承上启下的关键一

环，建立一套聚焦数据安全、隐私安全与医疗安全三大核心维度的评价框架，明确具体评价要素、方法与准则，为医学人工智能系统的研发、测试、部署与监管提供统一、可操作的安全评价工具，确保其在全生命周期内安全、可靠、可控，最终护航“人工智能+医疗”的高质量发展。

（二）编制目的

随着医学人工智能技术在疾病诊断、治疗决策、健康管理等核心医疗环节的深度融合与应用，其引发的数据安全、隐私泄露及临床误诊等风险日益凸显。医疗数据的高敏感性、算法决策的不可控性以及系统失效可能带来的直接生命健康威胁，使得建立一套科学、统一、可操作的安全评价体系成为行业发展的紧迫需求。目前，人工智能领域虽已存在部分通用安全标准，但缺乏针对医疗场景高风险特性的专项安全评价规范。这导致医疗机构、技术研发方及监管机构在实践层面面临评价维度不一、方法各异、结论互认困难等问题，严重制约了医学人工智能的安全落地与可信发展。

为此，依据《医学人工智能治理综合评价指南　第 1 部分：总则》的总体框架，特制定本标准《医学人工智能治理综合评价指南　第 2 部分：安全评价》。本标准将构建一个覆盖数据安全、隐私安全与医疗安全三大核心维度的专项评价体系，明确其在全生命周期中的具体评价要素、方法与准则。将为各相关方开展医学人工智能安全评价提供统一的技术依据和操作指

引，有效识别、预警和防范关键安全风险，确保医学人工智能系统在整个生命周期内安全、可靠、可控。

二、 任务来源及编制原则和依据

（一）任务来源

本标准编制任务来源于浙江省数理医学学会于 2025 年 5 月 2 日下达的浙数医 [2025]11 号关于批准《医学人工智能治理综合评价指南　第 1 部分：总则》等两项团体标准立项的通知，归口单位为浙江省数理医学学会，标准名称为《医学人工智能治理综合评价指南　第 2 部分：安全评价》，项目编号:ZSMM-2025-005。

（二）编制原则

本标准的制定工作遵循“统一性、协调性、适用性、一致性、规范性”原则，本着先进性、科学性、合理性和可操作性的原则，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》给出的规则编写。

（三）编制依据

本文件的编制主要参考与依据以下文件：

1.《标准起草规则　第 8 部分：评价标准》（GB/T 20001.8—2023　）

2.《信息安全技术　健康医疗数据安全指南》（GB/T 39725—2020　）

3.《医学人工智能治理综合评价指标体系 》（DB4403/T 643—2025　）

4.《标准化工作导则 第 1 部分:标准化文件的结构和起草规则》（GB/T 1.1—2020）

5.《网络安全技术　生成式人工智能服务安全基本要求》（GB/T 45654—2025 ）

6.《信息技术服务　从业人员能力评价要求》（GB/T 37696—2019）

7.《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》（GB/T 45652—2025）

8.《网络安全技术　人工智能生成合成内容标识方法》（GB/T 45438—2025）

9.《数据安全技术　数据安全和个人信息保护社会责任指南》（GB/T 46071—2025）

10.糖尿病视网膜病变人工智能筛查应用规范（DB52/T 1726—2023）

11.《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）

三、 编制过程

1、实地调研阶段,2025 年 1 月至 2 月，编制组通过问卷调查、实地走访、关键人物访谈、小组访谈等方式对科研院校、医疗卫生机构、人工智能相关企业及相关政府主管部门的医学人工智能软件在研发、使用、推广的过程中的数据安全、隐私安全以及医疗安全的评价要素构成展开实地调研和讨论。

2、规划准备阶段,2025 年 2 月 21 日编制组正式成立，由医学人工智能领域、卫生行政管理研究领域、卫生法学领域、医学伦理领域、卫生经济学评价领域、数据安全领域的学者专家、行政管理者、卫生技术人员、工程师等组成。编制组制定了详细的编制计划方案，形成了明确的分工机制。编制组开展前期文献研究，收集和整理国内外相关法律法规、政策文本、标准规范和研究论文，分析适宜医学人工智能安全治理评价的规范性要素、技术要点和框架结构。

3、标准起草阶段,2025 年 3 月至 4 月，在充分调研和理论研究的基础上，编制组开始标准文本的起草工作，形成了《医学人工智能治理综合评价指南　第 2 部分：安全评价》工作组讨论稿。

4、申请立项阶段，2025 年 4 月 5 日，标准编制组向浙江省数理医学学会递交团体标准立项申请表，于 2025 年 4 月 9 日收到受理通知书。

5、立项论证阶段，2025 年 4 月 24 日，浙江省数理医学学会标准化工作委员会组织召开立项论证会，《医学人工智能治理综合评价指南　第 2 部分：安全评价》通过立项论证评审，经公示，于 2025 年 5 月 2 日成功获批立项。

6、标准研制阶段，2025 年 5 月至 11 月，标准编制组根据专家意见，经过多轮研讨，对《医学人工智能治理综合评价指南　第 2 部分：安全评价》工作组讨论稿进行修改，于 2025 年 11 月 28 日完成《医学人工智能治理综合评价指南　第 2 部分：安全评价》征求意见稿与编制说明。

四、 主要内容的确定

本文件的重要技术内容系基于对国家相关政策法规的系统研究、对国内外相关标准的参考借鉴，并经由起草组专家多轮专题研讨后最终确定。《医学人工智能治理综合评价指南　第 2 部分：安全评价》有五个章节及附录和参考文献。其中主要内容包括：（一）范围；（二）规范性引用文件；（三）术语和定义；（四）安全评价指标体系内容；（五）安全评价指标内涵。

(一)　范围

本文件提供了开展医学人工智能治理综合评价中关于安全维度评价的指导，给出了安全维度涉及的权利、需要考虑的评价要素，以及评价执行的相关信息。

本文件适用于开展医学人工智能对安全治理产生的现实或潜在影响的综合评价活动。

（二） 规范性引用文件

本章节主要包括了标准文本中规范性引用的文件。

（三） 术语和定义

本章节的术语和定义参考《医学人工智能治理综合评价指南 第一部分：总则》。

《医学人工智能治理综合评价指南 第一部分：总则》界定的术语和定义适用于本文件。

（四） 安全评价指标体系内容

本章介绍了医学人工智能治理综合评价指南中的安全评价指标体系内容，包含二级评价指标3个，三级评价指标19个。其中二级指标包括数据安全、隐私安全和医疗安全。指标体系内容编制依据包括《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）、《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）、《生成式人工智能服务管理暂行办法》（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 国家广播电视总局令第15号）和《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）等政策文件，经由起草组专家研讨审议后形成了本章指标体系内容。

（五） 安全评价指标内涵

本文件第五章 安全评价指标内涵，编制依据情况如下表。

二级指标	三级指标	依据
5.1.1 数据安全	5.2.1 数据使用授权 5.2.2 数据使用范畴 5.2.3 数据用益权 5.2.4 数据采集 5.2.5 数据加密 5.2.6 数据传输 5.2.7 数据存储 5.2.8 数据共享 5.2.9 数据销毁	《数据安全技术 数据安全风险评估方法》（GB/T 45577—2025） 《信息安全 人工智能数据安全通用要求》(DB11/T 2251—2024） 《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》（GB/T 45652—2025） 《网络安全技术 生成式人工智能数据标注安全规范》（GB/T 45674—2025） 《物联网数据质量评价方法》（GB/T 44811—2024） 《医学人工智能治理综合评价指标体系》(DB4403/T 634—2025) 《信息安全技术 个人信息安全规范》（GB/T35273—2020） 《智能交通数据安全服务》（GB/T37373—2019） 《信息安全技术健康医疗数据安全指南》（GB/T39725—2020） 《金融数据安全数据生命周期安全规范》（JR/T0223—2021） 《电信网和互联网数据安全评估规范》（YD/T3956—2021） 《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》（YY/T1833.3—2022）
5.1.2 隐私安全	5.2.10 身份信息 5.2.11 疾病信息 5.2.12 生物识别信息 5.2.13 地理空间信息	《数据安全技术 敏感个人信息处理安全要求》（GB/T 45574—2025） 《数据安全技术 数据安全和个人信息保护社会责任指南》（GB/T 46071—2025） 《医学人工智能治理综合评价指标体系》(DB4403/T 634—2025) 《信息安全技术个人信息安全规范》(GB/T35273—2020） 《信息安全技术健康医疗数据安全指南》（GB/T39725—2020） 《人工智能算法金融应用信息披露指南》（JR/T 0287—2023） 《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》（YY/T1833.3—2022）
5.1.3 医疗安全	5.2.14 医师信赖率 5.2.15 质控改善率 5.2.16 假阴性率 5.2.17 假阳性率 5.2.18 预警错误发生率 5.2.19 推荐错误发生率	依据《医学人工智能治理综合评价指标体系》（DB4403/T 634—2025）、《信息安全技术 健康医疗数据安全指南》（GB/T39725—2020）、《糖尿病视网膜病变人工智能筛查应用规范》（DB52/T 1726— 2023）以及相关行业专家意见共识编制。

五、 采标情况

无

六、 重大分歧意见的处理

本标准制定过程中无重大分歧。

七、 与国家法律法规和强制性标准的关系

本标准为指南类团体标准，与有关的现行法律、法规和强制性国家/行业标准无抵触。

八、 标准实施的建议

标准发布后视各方反映情况，可以举办培训班来指导标准的实施。

九、 其他应予说明的事项

无

《医学人工智能治理综合评价指南　第 2 部分：安全评价》

团体标准编制组

2025 年 11 月 28 日