《医学人工智能治理综合评价指标体系》 解读

一、编制背景

2017年7月8日，国务院印发了《新一代人工智能发展规划》，要求初步建立人工智能法律法规、伦理规范和政策体系，形成人工智能安全评估和管控能力。2018年10月，习近平总书记在十九届中央政治局第九次集体学习时明确指出：“要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控。”

2021年9月25日，《新一代人工智能伦理规范》发布，提出了增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等6项基本伦理要求，并提出了人工智能管理、研发、供应、使用等特定活动的18项具体伦理要求。2022年3月20日，国务院办公厅印发《关于加强科技伦理治理的意见》，要求“进一步完善科技伦理体系，提升科技伦理治理能力，有效防控科技伦理风险，不断推动科技向善、造福人类”，将人工智能伦理纳入科技伦理的综合治理中。2023年5月，国家网信办发布《生成式人工智能服务管理暂行办法》，旨在促进生成式人工智能的健康发展和规范应用，维护国家安全、社会公共利益及公民、法人和其他组织的合法权益。2025年3月《人工智能示范法

3.0》发布，要构建起“动态适配、前瞻治理”的监管框架，为人工智能产业高质量发展提供法治保障；鼓励支持中小企业创新与开源，支持人工智能产业健康发展；新增内部吹哨人制度，强化人工智能研发与应用的安全监督。同月29日中国医学科学院医学信息研究所召集起草的《医疗机构部署DeepSeek专家共识》在中关村论坛医学AI创新与发展论坛上发布，该共识为全国医疗机构提供了可操作的AI部署指南，旨在推动我国医疗机构部署人工智能规范化发展。

随着人工智能技术的应用普及，人工智能技术滥用等问题不断出现，同时也带来了信息安全问题。2023年4月，美国斯坦福大学以人为本人工智能研究所（Stanford HAI）发布的《2023年人工智能指数报告》（Artificial Intelligence Index Report 2023）分析了人工智能的影响和年度趋势。报告发现，关于人工智能滥用的事件数量正在迅速上升。通过追踪人工智能道德滥用相关事件的人工智能算法和自动化事件和争议（AIAAIC）数据库，2012年至2022年间人工智能相关争议事件增加近26倍。针对各类人工智能问题，世界主要国家和地区陆续从安全性审查、技术禁用等角度出台相关监管措施以实现人工智能的有效治理。

探索合理、有效的治理方案，实现人工智能健康有序发展已经成为全球重要课题。在第三届“一带一路”国际合作高峰论坛上，中国发布的《全球人工智能治理倡议》（下称《倡议》）赢得国际社会广泛赞誉。《倡议》站在构建人类命运共同体、维护全人类福祉的历史高度，就人工智能发展、

安全、治理三个方面公平公正地提出建设性解决措施，为全球人工智能治理提供中国方案。2023年11月首届人工智能安全峰会在英国布莱奇利园举行并发布了《布莱奇利宣言》，美国、中国、日本、德国、印度等20多个国家的政府代表以及联合国、经合组织、国际电信联盟等多个国际组织的代表参会，28个国家和欧盟共同签署了《布莱奇利宣言》，同意通过国际合作，建立人工智能（AI）监管方法。2024年7月，欧盟发布全球第一部关于人工智能的综合性法律《人工智能法》，确保AI系统的开发、部署和使用符合欧盟的价值观和法律规定，同时保护健康、安全和基本权利。2024年1月，WHO发布的《卫生领域人工智能的伦理与治理：多模态大模型指南》提出六大伦理原则：保护自主权、促进人类福祉、确保透明度、责任担当、包容公平、可持续性。

目前人工智能技术主要造成三类问题。一是信息安全问题。人工智能需要实时更新海量数据用于模型训练。当前，在个人隐私领域中，相关企业、平台尚未达成遵循知情同意原则的共识，存在越权使用现象。因此，用户如果缺乏相关意识，在使用人工智能技术时，很可能导致个人隐私数据泄露。二是虚假信息问题。尽管技术中立，但人为投放不实消息或片面事实可能会导致人工智能模型产出虚假信息，也就是“深度伪造”。三是数字鸿沟问题。人工智能具备高效的信息检索和整合能力，虽然为人们的工作和生活提供了极大方便，但其产出的真实性和准确性并不牢靠。因此，信息甄别、自主决策能力强的使用者将更善于使用这项工具，而盲

从的使用者则可能为其所害。随着两者使用技术的差距不断沉积，数字鸿沟也就此加深。

人工智能技术在医学领域使用同样面临以上的治理问题外，还会面临人工智能技术带来卫生系统资源配置的深刻影响。然而，关于医学人工智能治理评价指标体系，业内一直高度关注，但尚未形成广泛一致的意见。截至 2025 年 4 月，深圳市缺乏本地化的医学人工智能治理评价指标体系，各医疗卫生机构开展医学人工智能治理试验评价遵循的范式参差不齐，导致各级医疗机构间或区域内医学人工智能治理评价结果管理困难的实际境况，通过管控和治理发现医学人工智能可能带来的潜在风险尚未形成有效的监测与预警机制。

为了满足当前医学人工智能治理评价的标准需求，科学引导深圳市医学人工智能的研发与应用的同时，促进各类医疗机构及其他利益相关方组织共同参与医学人工智能治理活动的有序有效地开展，亟需形成医学人工智能治理评价指标体系，从安全、风险、效用、效率、效益的多元视角加强医学人工智能的治理影响加以监测与管控。

二、必要性和意义

深圳市响应国家《倡议》号召，为了更好推动医学人工智能应用的发展，在加强医疗人工智能治理能力的同时，亟需构建一套分类管理、综合评价、贯穿医学人工智能技术开发全生命周期的人工智能治理评价指标体系，指导医学人工智能治理活动的开展，动态评估不同类别医学人工智能在安

全、风险、效用、效率和效益层面上的表现，规范建设深圳市医学人工智能治理，以达到对医学人工智能发展潜在风险研判和防范，确保其安全、可靠、可控的发展。推动深圳市智慧城市医疗人工智能治理工作先行示范，助力深圳市粤港澳大湾区、中国特色社会主义先行示范区的“双区建设”。

为了促进医学人工智能健康发展和规范应用，维护国家安全和公共利益，保护公民、法人和其他组织的合法权益，根据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《中华人民共和国科学技术进步法》《生成式人工智能服务管理暂行办法》等政策法规，结合人工智能治理的相关行业标准，制定本评价指标体系。通过医疗人工智能治理综合评价，从安全、风险、效用、效率以及效益层面的动态评估结果为规范建设人工智能治理体系提供决策性参考。

《医学人工智能治理综合评价指标体系》的制定、发布和实施，明确了深圳市医学人工智能治理评价的行业指标体系架构，确定了深圳市医学人工智能治理综合评价的建设方向，为后续深圳市医学人工智能治理评价相关标准的编制提供了依据和规范性指引，从而促进技术健康发展，保护合法权益，增强公众信任，助力深圳在智慧城市和粤港澳大湾区建设中发挥示范作用。

三、主要内容

《指标体系》现有7个章节以及参考文献。包括范围、规范性引用文件、术语与定义、指标选取原则、指标体系内

容、指标体系架构图、指标内涵（一级指标、二级指标、三级指标）。以下对标准中的主要条款进行简要说明。

（一）范围

本文件提供了开展医学人工智能治理评价的指导，给出了指标选取原则、指标体系的构成、内容、架构图和指标内涵等方面的建议，并给出了相关信息。

本文件适用于评价深圳市行政区域内医学人工智能治理产生的现实或潜在影响的活动。

（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

本章节主要包括医学人工智能、医学人工智能治理、医学人工智能治理评价、训练数据、医学人工智能技术内在风险、医学人工智能技术应用风险的术语与定义。

（四）指标选取原则

本章节介绍了医学人工智能治理评价指标体系中评价指标选取的原则，包括可行性原则、全面性原则、代表性原则、时效性原则。

（五）指标体系内容

本章节介绍了指标体系的构成和各级指标包含的内容。医学人工智能治理评价指标体系的结构分3个层级，其中一级指标3个（安全与效用评价、风险评价、卫生经济评价），二级指标6个（安全评价、效用评价、技术内在风险评价、技术应用风险评价、效率评价、效益评价），三级指标18个

（数据安全、隐私安全、医疗安全、场景渗透、适用效能、受众体验、算法风险、训练数据风险、生成内容风险、社会安全风险、伦理风险、社会经济风险、成本效率、规模效率、配置效率、经济效益、社会效益、健康效益）。

（六）指标体系架构图

本章节给出了医学人工智能治理评价指标体系的架构图。

（七）指标内涵

本章节给出了医学人工智能治理评价指标体系中一级指标、二级指标、三级指标的内涵。

四、附则

本文件由深圳市卫生健康委员会提出并归口，起草单位包括深圳市卫生健康委员会、南方医科大学、深圳市卫生健康发展研究和数据管理中心、中国科学院自动化研究所、南方医科大学深圳医院。