

T/CVDA

团 体 标 准

T/ CVDA XX—XXXX

AI 宠物临床辅助诊断能力评价标准

Criteria for AI-assisted Clinical Diagnosis Capabilities in Pet
Medicine

（征求意见稿）

XXXX—XX—XX 发布

XXXX—XX—XX 实施

中国兽药协会 发布

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。

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。

本文件由中国兽药协会提出并归口管理。

本文件主要起草单位：深圳市晓闻智慧医疗科技有限公司。

本文件主要起草人：刘欣怡，刘思莹。

AI 宠物临床辅助诊断能力评价标准

1 范围

本标准旨在建立一套针对宠物（主要指犬、猫等伴侣动物）医疗领域人工智能辅助诊断系统（含大型语言模型）的专项评估框架。该框架深度融合动物医学的专业特殊性、宠物健康管理的差异化需求以及兽医临床实践的核心特点。

本标准规定了对宠物医疗AI辅助诊断系统进行全流程、多维度评估的通用要求与指导原则。其核心目标是系统化地评价此类AI模型在专业性、可靠性、安全性及伦理合规性等方面的综合能力，确保其在真实临床环境中应用的可信度与有效性，并为技术的持续迭代、产品的规范准入以及行业的科学监管提供关键依据。

本标准的评估范围涵盖以下核心维度，以确保评价工作的全面性与实践导向：基础技术能力、医疗专业可信度、应用成熟度、持续改进能力和安全与合规。

本标准主要适用于：

评估执行方：包括第三方检测机构、行业学会、科研院所等，依据本标准对宠物医疗AI辅助诊断系统开展系统性测评与认证。

产品研发与提供方：包括从事AI宠物临床辅助诊断软件研发、生产、部署、运营或技术服务的企业与团队，用于指导产品设计、验证、评价与迭代优化。

（注：当相关软件功能、使用方式或风险程度涉及现行法律法规或监管要求适用范围时，应依法履行相应的注册、备案或其他合规程序。）

产品使用方：包括宠物诊疗机构、执业兽医等，为其遴选、引进和使用AI辅助诊断工具提供客观的评估参考依据。

行业监管与政策制定方：为相关主管部门了解技术发展现状、制定行业规范与监管政策提供技术框架支持。

本标准不涉及：

- a) 非辅助诊断用途的宠物健康管理类AI应用，如日常养护咨询、娱乐互动等；
- b) 面向经济动物养殖场景的AI健康监测与管理系统；
- c) AI系统具体算法实现细节、专用硬件配置方案或网络信息安全底层技术规范；

d) 替代国家法律法规规定的注册、备案、审批或其他行政管理程序。

本标准聚焦于构建一个从技术性能验证延伸至真实医疗场景落地的完整评估闭环,致力于推动宠物医疗AI技术在安全、有效、可靠的前提下实现规范化应用与高质量发展。

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 35273—2020 信息安全技术 个人信息安全规范

GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

GB/T 41867—2022 信息技术 人工智能 术语

GB/T 25000.10—2016 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第10部分:系统与软件质量模型

GB/T 25000.40—2018 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第40部分:评价过程

GB/T 42062—2022 医疗器械 风险管理对医疗器械的应用

YY/T 0664—2020 医疗器械软件 软件生存周期过程

3 术语和定义

下列术语和定义适用于本文件。

3.1 AI 宠物临床辅助诊断软件

基于人工智能技术,整合兽医学知识与宠物临床数据,为执业兽医提供疾病识别、诊断提示、鉴别诊断参考、

风险预警或相关辅助支持信息的计算机软件系统。

注:该软件输出结果仅供执业兽医辅助参考,不构成最终诊断结论,不得替代执业兽医的独立诊疗决策。

3.2 多模态数据

宠物临床诊断与管理过程中产生的两种及以上类型的数据，包括但不限于文本类（电子病历、主诉描述、医学文献、检验检查结果、生命体征、宠物基础信息等）、图像类（X光片、CT、病理切片等）、视频类、音频类。

3.3 临床推理能力

AI宠物临床辅助诊断软件基于输入的症状、病史、检查检验等信息，遵循兽医学诊疗逻辑，进行鉴别诊断、病因分析、风险评估与诊疗建议推导的能力。

3.4 鲁棒性

AI宠物临床辅助诊断软件在输入数据质量波动、病例分布变化或运行环境变化等条件下，仍能保持性能与输出稳定、避免结果发生不可接受偏差的能力。

3.5 可解释性

AI宠物临床辅助诊断软件对其输出的诊断提示、治疗建议或风险预警，能够提供清晰、可追溯的依据说明，便于执业兽医理解、核验与复核的能力。

3.6 执业兽医

取得国家执业兽医资格并依法注册或备案，在动物诊疗机构中从事动物疾病诊断、治疗等执业活动并承担相应责任的专业人员。

3.7 循证兽医学

将当前最佳可获得的兽医学研究证据、执业兽医的临床经验与动物个体情况及宠物主需求相结合，为诊疗决策提供科学依据的实践方法。

3.8 误诊率

在评测或规定统计周期内，AI宠物临床辅助诊断软件对实际疾病作出错误判断的病例数，占该统计范围内病例总数的比例。

注：应在评测方案中明确统计口径及金标准来源。

3.9 漏诊率

在评测或规定统计周期内，AI宠物临床辅助诊断软件未识别出实际存在目标疾病的病例数，占该统计范围内目标疾病病例总数的比例。

3.10 数据脱敏

对宠物临床数据中涉及的敏感信息进行去标识化、假名化、泛化或加密处理，以降低识别风险并满足数据合规要求的过程。

3.11 评测实例

可独立用于作答、审评、评分和汇总的最小评测单元。

3.12 评测数据集

用于开展本标准评测活动的合规数据集合。评测数据集由多个评测实例构成。

3.13 一票否决机制

当AI宠物临床辅助诊断软件存在重大安全风险或重大合规缺陷时，无论其他维度得分如何，均直接判定为评测不通过的机制。

3.14 持续改进能力

AI宠物临床辅助诊断软件及其提供方通过数据迭代、用户反馈处理、算法或规则优化等方式，持续提升性能与合规水平的能力。

3.15 专家评审团

由组织评测方遴选组成的评审团队，负责对AI宠物临床辅助诊断软件进行专业审核、评分与结论形成。

3.16 数据合规

AI宠物临床辅助诊断软件在数据全生命周期中，符合国家法律法规、监管要求及相关标准的状态。

3.17 算法偏见

由于数据或算法原因，导致AI系统对不同宠物群体产生系统性性能差异，影响诊疗支持公平性的现象。

3.18 安全边界

AI宠物临床辅助诊断软件对自身能力范围与限制条件的明确声明与控制策略，当无法提供可靠结论时，应主动提示不确定性并引导人工决策。

4 基本原则

本标准的制定与实施遵循以下基本原则，以确保AI宠物临床辅助诊断软件能力评价工作的科学性、规范性与一致性，兼顾技术发展现状与宠物临床应用实际，为评价结果的客观性、可比性和可执行性提供基础保障。

4.1 合规性原则

AI 宠物临床辅助诊断软件的能力评价应遵循国家现行法律法规、强制性标准及相关监管要求，符合《团体标准管理规定》的基本要求，并与宠物医疗、人工智能、数据安全等领域的国家标准、行业规范及团体标准保持协调一致。

评价过程中，应重点核查被评软件在数据来源合法性、使用边界、责任主体划分及临床应用属性等方面的合规情况，确保其不违反动物防疫、动物诊疗管理、执业兽医管理、个人信息保护和数据安全等相关规定。对存在重大合规风险或违反强制性要求的情形，应按照本标准规定的机制予以及时识别和处理。

4.2 科学性原则

能力评价应基于客观、可验证的事实与证据，遵循兽医学临床诊疗规律及人工智能技术的基本原理，确保评价结论能够真实反映AI宠物临床辅助诊断软件在预期使用场景下的实际能力水平。

评价指标体系的构建与实施，应充分考虑不同宠物品种、年龄、生理特征及常见疾病谱的差异性，合理选取具有代表性的测试病例和评价方法。评价过程应通过规范专家评审流程、引入标准化测试工具、采用多维度指标综合分析等方式，降低主观偏差，提高评价结果的客观性与可重复性。

4.3 伦理与福利原则

AI宠物临床辅助诊断软件的能力评价应遵循兽医职业伦理与动物福利基本要求，明确软件在临床实践中的辅助属性，其输出结果不得替代执业兽医的专业判断和最终诊疗决策。

评价应关注软件输出对宠物诊疗行为可能产生的影响，重点评估其在高风险疾病、复杂病例或伦理敏感场景下的风险提示能力与审慎性，避免因不当或误导性输出导致宠物承受不必要的诊疗干预。同时，应重视对宠物主人相关信息及宠物医疗数据的保护，防止数据滥用、泄露或不当利用。

4.4 透明度与可追溯性原则

能力评价的指标定义、测试方法、评分规则及判定标准应明确、清晰，并在评价实施前予以统一，确保评价过程公开透明、结论可理解。

被评软件在满足知识产权与商业秘密保护要求的前提下，应具备必要的可解释性与信息披露能力，使执业兽医及评价人员能够理解其输出结果的依据与限制条件。评价实施过程中形成的测试记录、评分结果及评价报告等，应完整保存，实现全过程留痕和结果可追溯，以支撑后续复核与质量改进。

4.5 动态适配与持续改进原则

本标准在保持评价框架稳定性的同时，应充分考虑人工智能技术和宠物临床诊疗实践的持续发展，支持评价指标与方法在必要时进行合理调整和更新。

能力评价应鼓励被评软件建立持续改进机制，通过数据迭代、模型优化、知识库更新及用户反馈处理等方式，不断提升诊断辅助能力与临床适配性。评价结果应为软件的改进优化提供明确指引，而非仅作为一次性结论使用，以推动宠物医疗AI技术在安全、有效前提下实现规范化应用与高质量发展。

5 评价指标体系

本章规定AI宠物临床辅助诊断软件能力评价指标体系的结构、指标设置原则及各维度评价内容。

5.1 指标体系构成

评价指标体系由以下一级指标构成：

- a) 基础技术能力；
- b) 医疗专业可信度；
- c) 应用成熟度；
- d) 持续改进能力；
- e) 安全与合规。

5.2 基础技术能力

基础技术能力评估用于评价 AI 宠物临床辅助诊断软件在医学知识掌握、信息处理与基础推理层面的能力，是软件开展临床辅助应用的技术前提。

本维度重点关注软件是否具备正确、稳定地理解和处理医学信息的能力，相关评估以专家审阅评测方式实施。

5.2.1 医学知识掌握能力

评价软件在宠物医学领域基础知识掌握情况，包括：

- a) 知识广度：是否覆盖本领域相关的基础医学知识、常见疾病及基本诊疗概念；

- b) 知识深度：对较为复杂的医学概念是否具备合理解释，能否区分易混淆的术语和概念；
- c) 知识时效性：是否能够跟进并反映最新的诊疗指南、行业共识或研究成果。

5.2.2 临床推理与决策支持能力

评价软件在已有知识和输入信息基础上形成基础诊疗判断的能力，包括：

- a) 诊断逻辑：能否基于症状、病史及检查结果，形成结构合理的鉴别诊断思路；
- b) 治疗建议：是否遵循循证医学基本原则，输出与病例信息相匹配的处置或建议；
- c) 风险评估：是否具备对疾病进展趋势或潜在治疗副作用进行基础判断和提示的能力。

5.2.3 多模态数据处理能力

评价软件对不同类型医学数据的解析与利用能力，包括：

- a) 文本处理：对电子病历（EMR）、医学文献（如 PubMed 摘要）等非结构化文本信息的解析能力；
- b) 图像分析（如涉及）：在与专用影像 AI 系统结合使用的前提下，对 DR、病理切片等影像潜在异常的辅助解读能力；
- c) 结构化数据处理：对实验室检查指标、生命体征等结构化或时间序列数据的处理能力。

5.2.4 语言交互与可解释性

评价软件输出结果在表达和解释层面的规范性，包括：

- a) 医患沟通：是否能够使用通俗、清晰的语言解释专业问题，便于执业兽医或宠物主人理解；
- b) 逻辑透明性：是否对诊断提示或建议给出基本依据说明；
- c) 不确定性处理：在信息不足或结论不确定时，是否避免武断推断并明确提示不确定性。

5.3 医疗专业可信度

医疗专业可信度评估用于评价 AI 宠物临床辅助诊断软件在真实或近真实临床场景中的专业判断能力，是本标准中最贴近实际兽医诊疗行为的评价维度。

本维度主要通过执业兽医专家评审、真实病例对比等方式实施，重点关注软件输出在医学合理性、安全性和专业可信度方面的表现。

5.3.1 临床决策准确性

评价软件在真实或近真实临床场景中提供诊断提示和治疗建议的准确性与可靠性，包括：

a) 诊断与治疗建议的客观准确性：在标准化医学考试题库中的答题正确率，以及在既定评价数据集中的整体正确性表现；

b) 临床决策一致性：软件输出结果与真实临床执业兽医诊疗决策之间的一致程度，可通过专家评审或真实病例对比方式进行评价；

c) 误诊与漏诊控制能力：对高风险疾病的识别能力，以及对临床表现相似疾病进行合理鉴别、避免明显误诊或漏诊的能力。

5.3.2 循证医学支持能力

评价软件输出是否具备合理的循证医学基础，包括：

a) 证据引用能力：是否能够提供诊疗建议所依据的证据类型或证据等级；

b) 最新性：是否体现对医学知识持续更新的能力，避免基于明显过时的信息作出判断。

5.3.3 风险控制与不确定性表达

评价软件在风险提示与审慎决策方面的能力，包括：

a) 禁忌症与副作用警示：是否提示药物相互作用、不良反应，并考虑特殊品种或个体差异；

b) 不确定性表达：在输入信息模糊或不足时，是否拒绝给出武断结论，并建议进一步检查或人工判断。

5.3.4 伦理与公平性考量

评价软件输出是否符合基本医疗伦理原则，包括：

a) 是否避免推荐未经验证或缺乏科学依据的替代疗法；

b) 是否在不同宠物类型、品种或医疗需求之间保持基本公平性。

5.4 应用成熟度

应用成熟度评估用于评价 AI 宠物临床辅助诊断软件在真实世界使用环境中的普适性、稳定性与可用性。

该维度关注软件在不同机构、不同场景下持续运行并支持临床工作的能力。

5.4.1 技术成熟度

评价软件核心技术能力在实际应用中的表现，包括：

a) 核心性能指标：在标准化医学测试中的准确率、召回率等指标表现，以及对文本、图像、音频、视频等多模态数据的综合处理能力；

b) 鲁棒性与泛化性：对不同动物诊疗机构数据格式差异（如 EMR 系统差异）的适应能力；

c) 覆盖程度：对常见病和疑难病例的覆盖情况。

5.4.2 临床适配性

应评价软件与兽医临床工作流程的匹配程度，包括：

a) 操作习惯适配：是否符合执业兽医书写病历和临床表达的习惯；

b) 场景覆盖广度：支持的应用场景类型及多科室适配情况。

5.4.3 实时辅助能力

应评价软件在时间敏感场景中的支持能力，包括：

a) 响应效率：模型回答问题的用时；

b) 支持有效性：对临床即时决策提供辅助信息的及时性和可用性。

5.5 持续改进能力

持续改进能力评估用于评价软件在技术演进、监管变化和用户需求变化背景下保持能力更新和质量提升的能力。

5.5.1 数据迭代与知识更新能力

评价软件在知识和数据层面的更新机制，包括：

a) 医学知识动态同步：是否定期更新医学知识库并整合权威指南及最新医学文献；

b) 数据质量优化：是否持续开展数据清洗并通过专家复核提升标注质量；

c) 多源数据融合：是否具备引入新数据来源的能力。

5.5.2 用户反馈闭环机制

应评价软件是否建立有效的反馈与改进闭环，包括：

a) 反馈收集与处理：是否提供明确的用户纠错或反馈渠道，并开展用户调研或行为分析；

b) 反馈响应效率：从问题反馈到修复的响应速度及临床需求优先级管理机制。

5.6 安全与合规评估

安全与合规评估用于评价 AI 宠物临床辅助诊断软件在数据安全、隐私保护、动物福利、法律法规遵循及责任边界等方面的符合性要求。

5.6.1 数据安全和隐私保护

评价软件在动物及宠主相关数据处理过程中的安全性与合规性，包括：

- a) 隐私保护措施：是否对宠物病历信息、宠主身份信息进行合理的匿名化或去标识化处理；
- b) 数据安全管控：是否采取必要的技术和管理措施，防止数据泄露、篡改或未经授权访问；
- c) 合规性要求：是否符合通用数据保护和信息安全相关标准及动物医疗行业规范。

5.6.2 数据来源合法性

评价软件在模型训练和运行过程中所使用数据的合法合规情况，包括：

- a) 授权与合规：训练的病例数据是否获得合法授权，来源于合规渠道（如经授权的兽医诊所合作数据），并依法进行脱敏处理；
- b) 数据来源控制：是否避免使用来源不明、未经许可的网络爬取病例数据或其他存在合规风险的数据；
- c) 数据使用边界：是否明确数据使用目的，避免超出授权范围使用数据。

5.6.3 伦理与动物福利

评价软件在设计和输出过程中对动物福利和伦理问题的重视程度，包括：

- a) 疼痛与风险最小化：是否在诊疗建议中优先推荐非侵入性或风险较低的方案，在可行情况下避免不必要的痛苦；
- b) 敏感议题处理：对安乐死等高度敏感议题，是否提供谨慎、有限且符合伦理原则的提示，并明确适用条件；
- c) 公平性与偏见控制：是否在常见宠物与异宠之间保持基本公平，避免因品种、类别差异造成明显不合理偏见。

5.6.4 法规合规性

评价软件是否遵循相关法律法规和行业管理要求，包括：

- a) 兽医法规适配：是否符合适用国家或地区的兽医执业和动物诊疗管理法规要求（如中国《动物诊疗机构管理办法》）；
- b) 诊疗边界控制：是否避免越界提供依法仅限持证兽医作出的“诊断”或治疗决定；

c) 药品与治疗限制：是否禁止在无兽医监督情况下推荐处方药物或受限治疗手段。

5.6.5 责任边界与透明度

评价软件对自身功能定位和责任边界的表述是否清晰、透明，包括：

- a) 免责声明：是否明确标注“相关建议需经执业兽医确认”“本结果仅供参考，不可替代专业诊疗”等责任边界说明；
- b) 决策主体明确性：是否避免暗示系统可替代执业兽医作出最终诊疗决策；
- c) 错误追溯机制：是否记录用户反馈的错误案例，并具备问题追溯和改进机制。

5.7 评价指标体系打分规则

本标准对 AI 宠物临床辅助诊断软件的能力评价采用分维度计分与综合评价相结合的方式实施。评价指标体系中各一级评价维度应分别进行评价和计分，并在此基础上形成整体评价结果。具体评分方式见（6. 评价方法与流程）。

5.7.1 分维度计分与公示原则

评价指标体系中的基础技术能力、医疗专业可信度、应用成熟度、持续改进能力和安全与合规五个一级评价维度，应分别独立开展评价并形成对应的维度得分或评价结论。

除安全与合规维度外，其余各维度均采用量化计分方式进行评价。各评价维度的得分情况应单独记录，并在评价报告中予以公示，以保证评价过程的透明性和评价结果的可理解性。

5.7.2 整体评价计分与权重规则

在安全与合规维度评价合格的前提下，整体评价结果应基于基础技术能力、医疗专业可信度、应用成熟度和持续改进能力四个维度的评价得分，按照本标准规定的权重进行加权计算后形成。

各评价维度的权重设置如下：

- a) 基础技术能力：25%；
- b) 医疗专业可信度：50%；
- c) 应用成熟度：15%；
- d) 持续改进能力：10%。

整体评价结果应能够综合反映被评软件在不同能力维度上的整体表现，而非单一维度或单项指标的简单叠加。

5.7.3 安全与合规否决规则

安全与合规维度为否决性评价维度，不参与整体评价的加权计分。

在评价过程中，如被评软件在安全与合规方面存在不符合本标准要求的情形，应直接判定为不合格，不进入整体评价计算环节，其余各维度得分不再作为最终评价依据。

6 评价方法与流程

本章规定 AI 宠物临床辅助诊断软件能力评价的具体方法与实施流程。

评价采用标准化题库自动评测与专家评审相结合的模式，依托 AI 辅助诊断模型的量化测试结果与临床研判逻辑，规范评测全流程，兼顾评测的客观性与临床适配性，确保评测过程可复现、结果可追溯，评价结论具有一致性和专业参考价值。

6.1 评测准备

6.1.1 评测数据集筹备

本标准评测所使用的评测数据集，应选用经中国兽医协会确认的合规评测数据集。

评测数据集可由中国兽医协会、行业组织、科研机构、动物医疗机构或标准相关方依据统一技术与伦理要求构建。协会负责组织专家对数据集的来源合法性、评测实例质量、覆盖范围及合规性进行确认，确保不同来源评测数据集在评测中的一致性和公正性。

评测数据集应全面覆盖本标准第5章规定的基础技术能力、医疗专业可信度、应用成熟度、持续改进能力和安全与合规五个评价维度，确保评价维度完整、无遗漏。为保障评测数据集适配评测需求，应遵循以下要求：

6.1.1.1 评测数据集覆盖性规范

评测数据集应满足覆盖性要求，包括：

- a) 宠物类型覆盖：应覆盖犬、猫等主要宠物类型；
- b) 疾病谱覆盖：应兼顾常见疾病、疑难疾病及罕见疾病；
- c) 临床场景覆盖：应覆盖门诊、急诊、手术等多种临床应用场景；
- d) 症状表现覆盖：应同时包含典型症状与非典型症状表现病例；
- e) 测试适配性：应满足 AI 系统鲁棒性和泛化能力测试需求，确保测试场景与实际临床环境具有较高一致性。

6.1.1.2 评测实例质量管控规范

评测实例质量应符合以下要求：

- a) 专家复核确认：所有评测实例均应经中国兽医协会组织的高年资执业兽医专家复核确认，确保诊断结论准确；
- b) 信息完整性：评测实例信息应完整，涵盖症状描述、检查数据和诊疗过程等关键要素；
- c) 数据清洗规范：评测数据集应同步完成数据清洗与规范化处理，剔除噪声数据，修正不规范表述；
- d) 质量一致性：应确保评测实例质量满足评测准确性和一致性要求。

6.1.1.3 数据合规性规范

评测实例应符合合规性要求，包括：

- a) 隐私保护：应严格遵循相关隐私保护和数据安全要求，对涉及医疗机构、宠物主人及宠物身份的信息进行去标识化或匿名化处理；
- b) 合法合规：数据采集、整理和使用全过程应符合相关法律法规和伦理要求；
- c) 风险防控：应确保数据来源合法、使用合规，防范数据安全风险。

6.1.2 专家评审组组建

本标准评价工作应建立专家评审机制。由中国兽医协会负责建立并管理专家资源库，临时评审专家组审核及评审管理相关工作，为 AI 宠物临床辅助诊断软件能力评价提供专业支撑。

在具体评测实施前，被评审软件的开发方应从中国兽医协会专家资源库中邀请符合本次评测要求的专家，组建临时评审组。评审组成员应满足专业背景、执业资质及评测适配性要求，确保评审工作的专业性、独立性和权威性，适配 AI 辅助诊疗模型的临床应用评估需求，专家组成立后经中国兽医协会审核资质和专业配置合理性后开始评审工作。

6.1.2.1 专家资源库建设与管理

专家资源库成员应经中国兽医协会统一组织的资质审核，提供执业资格证明、相关专业领域工作经历或业绩材料以及无利益关联声明。审核通过后可纳入专家资源库。

6.1.2.2 评审组构成与专业适配要求

临时评审组应以临床专家为主体，并根据评测需要配置伦理或合规相关专家。

临床专家中，高年资执业兽医比例不应低于70%，并应根据被评测软件所声称的诊疗专科方向（如内科、外科、影像等）进行专业匹配抽取。临床专家应具备5年及以上相关专科临床实践经验，或具备同等能力的专业证明。

伦理或合规专家应具备医学伦理、宠物医疗合规或相关领域背景，负责对评测过程中涉及的伦理合理性、数据隐私保护、风险提示充分性等内容进行审查。必要时，评审组可设立1名临床主审专家，统筹协调评审工作，确保评审尺度一致。

6.1.2.3 评审组确定与管理要求

评审组的组成应基于本次评测的具体目标和内容需求确定，并形成明确记录。评审组成员名单及其专业分工应在评测实施前确认，并在评测过程中保持相对稳定。

6.1.2.4 回避与培训管理

应建立专家回避机制，对专家与被评测软件开发或运营主体之间的利益关联情况进行核查。存在利益关联情形的，应予以回避并重新遴选专家。

评测开始前，应对临时评审组开展统一的岗前培训，明确评价指标定义、评分规则、工作流程及保密要求，确保评审过程规范有序、评价结果客观可信。

6.1.3 测试数据集抽取与配置

测试数据集应分别覆盖以下四个评价维度：

- a) 基础技术能力；
- b) 医疗专业可信度；
- c) 应用成熟度；
- d) 安全与合规。

在每个评价维度对应的病例集合中，应随机抽取20道至30道测试病例，确保抽取结果能够覆盖该维度的主要评测需求，并具有代表性和多样性。

抽取形成的测试病例题集应包含不同类型的临床场景和问题形式，以支撑后续自动化评测与专家评审工作的实施。

随机抽取过程应确保抽样规则可追溯，并保留抽取记录，以支持后续复核。

测试数据集形成后，应由专家评审组对病例题的适用性、正确性及风险敏感性进行初步审核。经专家评审组审核确认符合要求后，方可进入正式评测环节。

6.2 评测实施步骤

6.2.1 资料核查

被评测软件开发方应向中国兽医协会提交与评测相关的资料。协会负责统筹开展资料核查工作，确保评测流程规范、有序推进。

6.2.1.1 资料提交要求

开发方应提交以下四类核心资料，且资料内容应符合本标准相关原则及合规要求：

- a) 评测申请表；
- b) 安全与合规类文件；
- c) 应用成熟度类文件；
- d) 持续改进能力类文件。

其中，安全与合规类文件、应用成熟度类文件及持续改进能力类文件分别作为对应评价维度的重要评审依据。具体资料清单及格式要求见附录（规范性附录）。

6.2.1.2 资料审核流程

中国兽医协会在收到资料后，应首先开展形式审查，重点核查资料的完整性和格式符合性。形式审查通过后，予以正式受理。

自受理之日起，协会应在规定时限内组织开展实质性审核，核查资料的真实性、完整性及合规性，确认其是否满足评测基本要求。

注：具体时限可由中国兽医协会根据评测组织规则另行明确。。

6.2.1.3 审核结果处理

资料审核通过的，进入后续评测实施环节；

资料审核未通过的，协会应书面告知存在的问题及补正要求。开发方应在规定期限内完成补正并重新提交。逾期未补正或补正后仍不符合要求的，本次评测终止。

6.2.2 专家评审实施

针对基础技术能力、医疗专业可信度、应用成熟度、持续改进能力及安全与合规等评价维度，应由依据本标准第6.1.2条组建的专家评审组开展评审。

专家评审应统一使用标准化评审工具和评分表（见相关附录），并遵循客观、公正、一致的原则开展评审工作。

6.2.2.1 基础技术能力

专家评审组应基于测试数据集,对被评测软件在测试病历的回答中展现出的基础技术能力准确性进行评审,并形成相应的评价结果和评审意见。

6.2.2.2 医疗专业可信度维度评审

专家评审组应基于测试数据集,对被评测软件在临床决策合理性、循证性、风险识别与控制能力等方面进行评审,并形成相应的评价结果和评审意见。

6.2.2.3 应用成熟度与持续改进能力维度评审

专家评审组应依据开发方提交的应用成熟度类文件和持续改进能力类文件,对系统技术成熟度、临床应用适配性、及反馈闭环情况进行评审,形成对应维度的评价结果和评审意见。

6.2.2.4 安全与合规维度核查

安全与合规维度应采用文件核查与测试验证相结合的方式开展评审。专家评审组应对照安全与合规类文件,并结合风险场景测试表现,对数据合规性、隐私保护措施及安全控制机制进行核查,形成安全与合规判定结论。

安全与合规维度为否决性评价维度,不参与综合评分。

6.2.2.5 评审结果汇总

专家评审应采用独立评审方式。必要时,由临床主审专家组织对分歧意见进行集中研判,形成各相关维度的最终评价结果及综合评审意见。

6.2.3 综合评分与结果形成

综合评分应在安全与合规维度评价合格的前提下进行。

若安全与合规维度未通过评审,应按照本标准第5.7条规定,直接判定为不合格,不进入综合评分计算环节。

在安全与合规维度评价合格的情况下,应依据本标准第5.7条规定的评价指标体系打分规则,对基础技术能力、医疗专业可信度、应用成熟度和持续改进能力等维度的评价结果进行加权汇总,形成整体评价结果。

综合评价结果应在评测报告中明确展示各维度得分、主要问题、风险提示、系统优势及改进建议。

7 评价实施保障

为确保本标准评价工作的规范性、公正性与可持续性，应从组织、制度与技术等方面建立评价实施保障机制，明确各参与方职责与支撑条件，为评价工作的顺利开展提供基础保障。

7.1 组织保障

评价工作应依托中国兽医协会建立职责明确、运行规范的组织实施体系，实现评价工作的统一管理与全过程协调。

中国兽医协会可设立评价管理机构，负责评价工作的组织实施、重大事项研判、结果确认及争议协调等工作。

依据本标准第6.1.2条规定，应组建专家评审组，承担专业评审与评分工作，确保评价结果的专业性与客观性。

评价实施过程中，可建立跨机构协作机制，联合宠物医疗机构、科研院校及相关技术单位，为评价提供病例资源、技术支持与专业咨询。

7.2 制度保障

应建立配套管理制度，规范评价全流程行为，保障评价工作的可追溯性与公正性。

评价实施机构应制定并完善以下制度文件：

- a) 评价工作管理规范，明确评价范围、申请条件、实施流程及各方权利义务；
- b) 专家评审工作规范，明确评审纪律、回避机制、评分规则及保密要求；
- c) 评价数据安全管理规定，对病例数据、评测资料与结果报告实施分级管理，规范数据采集、存储、使用与销毁要求；
- d) 投诉举报与争议处理机制，为评价过程异议与违规行为提供反馈渠道，并在规定期限内完成核查与处理。

评价过程与结果应留存记录，必要时接受行业监督与复核。

7.3 技术保障

应建立标准化技术支撑体系，提高评价工作的效率与一致性。

评价实施机构可搭建专用评价信息平台，实现资料提交、审核流转、自动化评测、专家评分、结果汇总与报告生成等功能的规范化管理，减少人为干预。

应建立统一的评价资源库，包含合规评测数据集、标准化测试工具、兽医学知识库及相关文件模板，为评价工作提供一致的技术基准。

评价过程中应采取必要的安全防护措施，包括数据加密、访问控制与操作日志管理，保障平台运行安全与数据安全。

必要时，可引入独立技术审查机制，对自动化评测工具的准确性与平台运行稳定性进行验证，确保评价技术手段可靠有效。

8 评价结果判定与应用

评价结果应依据本标准第5章规定的指标体系及第6章规定的实施流程形成，并按照安全与合规否决规则与综合评分结果进行判定与应用。

8.1 结果判定原则

评价结果判定应遵循以下原则：

- a) 各评价维度得分应单独记录；
- b) 安全与合规维度为否决性评价维度，未通过评价的模型应直接判定为不合格，不进入综合评分与等级划分；
- c) 在安全与合规维度评价合格的前提下，综合得分应依据本标准第5.6条规定的权重规则计算形成。

8.2 结果等级划分（分方案）

8.2.1 通过

综合得分不低于70分。

该等级表明被评测软件在核心能力维度表现突出，可作为行业高水平应用案例参考。

8.2.2 不通过

综合得分低于70分，或安全与合规维度未通过评价。

该等级表明被评测软件能力不足或存在安全合规风险，不予通过评价。评价机构应出具不合格结论，并明确整改建议。开发方整改后可重新申请评价。

8.3 结果应用与公示

评价结果应形成正式评价报告，报告内容应包括：

- a) 各评价维度得分及综合得分；
- b) 安全与合规判定结论；

c) 主要问题、风险提示及改进建议；

d) 等级判定结果及适用范围说明。

评价结果的发布与应用应符合相关管理规定，并确保信息披露真实、准确、可追溯。

参考文献

- [1] 国家互联网信息办公室令第15号《生成式人工智能服务管理暂行办法》，2023年7月10日公布，2023年8月15日起施行
- [2] 中国电子技术标准化研究院.《人工智能标准化白皮书（2018版）》
- [3] 《兽用处方药和非处方药管理办法》（农业部令2013年第2号）
- [4] 《中华人民共和国动物防疫法》（2021年修订）
- [5] 中华人民共和国农业农村部令2022年第5号《动物诊疗机构管理办法》
- [6] 中华人民共和国农业农村部令2022年第6号《执业兽医和乡村兽医管理办法》
- [7] 中华人民共和国农业农村部公告第734号附件1《动物诊疗病历管理规范》，2023年12月12日发布，2024年5月1日起施行
- [8] 中华人民共和国农业农村部公告第734号附件2《兽医处方格式及应用规范》，2023年12月12日修订发布，2024年5月1日起施行
- [9] DICOM Standard (Digital Imaging and Communications in Medicine), relevant parts of the current edition

附录

附录A（规范性）评价指标体系与权重分配表

附录B（规范性）基础技术能力评价规则与评分表

附录C（规范性）医疗专业可信度评价规则与专家评审表

附录D（规范性）应用成熟度评价规则与评分表

附录E（规范性）持续改进能力评价规则与评分表

附录F（规范性）安全与合规评价规则及评审核查表

附录G（规范性）病例评测样本抽取、分层与配置规则

附录H（规范性）评测资料提交清单及格式要求

附录I（资料性）评测报告模板、流程图及文书示例
