医疗人工智能安全测评 · 方法标准

医疗 AI 安全性测评方法白皮书

面向基座模型与医疗智能体的中立第三方测评框架
及其在医疗行业 AI 采购中的应用

2026 版（V1.0）

二〇二六年

声明与适用范围

本白皮书由卫标医智科技（北京）有限公司、高校科研团队与行业专家共同编制，旨在阐述医疗人工智能安全性测评的方法论框架、测评维度与治理机制，并探讨其在医疗行业 AI 采购中的应用，为医疗机构、模型与产品供应商、监管与采购部门提供一套可参考、可复现、可比较的测评语言。

本白皮书所述方法面向研究与工程实践，不构成对任何具体产品安全性、合规性或临床适用性的背书或保证；亦不构成医疗建议、监管意见或法律意见。任何医疗 AI 产品的最终临床使用，均应在取得相应资质、完成上市前审评并建立上市后监测的前提下，由具备资质的医疗机构与专业人员负责决策。

测评机构作为中立第三方，测评活动遵循利益冲突隔离原则，不参与被测产品的研发，亦不持有可能影响测评独立性的利益关系。本白皮书内容会随技术演进、监管要求与实践反馈持续修订，最新版本以正式发布为准。

目录

摘要

人工智能正在快速进入诊疗、用药、分诊、随访与病历等核心医疗环节。与此同时，大语言模型与智能体的能力呈现出「高均值、长尾不可控」的特征：在常见问题上表现优异，却可能在罕见、对抗或边界场景中给出自信而错误的结论。医疗的高风险与不可逆性，使得「看起来很好用」与「可以安全使用」之间存在巨大的鸿沟。弥合这一鸿沟，需要一套独立、严谨、可复现的安全性测评方法，并使其能够真正落地于医疗机构的采购与准入实践。

本白皮书主张，医疗 AI 安全性测评应承担「度量衡」的角色：为整个行业提供统一的刻度与中立的标尺，使不同产品的安全性可被一致地度量与比较，而非由各方自说自话。围绕这一定位，本白皮书提出四个核心方法论支柱：

- **（一）对象分层**——严格区分「基座模型」与「医疗智能体」两类测评对象，前者考察医学知识与推理的能力上限，后者考察在真实工作流中的行为安全；
- **（二）自主性分级**——以 L0 至 L3 自主性等级界定人机责任边界，自主性越高，测评的严格程度与安全闸门越严苛；
- **（三）场景风险分级**——按临床场景的危害严重度、可逆性与可发现性进行风险分层，其中处方与用药决策为最高风险场景；
- **（四）多维度评分与硬性安全闸门**——安全性不可被单一聚合分数掩盖，关键安全失败实行「一票否决」。

在方法层面，本标准采用「封存金标准测试集」与「公开演示集」相互隔离的双轨机制，以对抗刷榜与训练污染；并将测评结果严格绑定到被测对象的具体版本、提示词与工具集，配套重测与上市后监测机制。在应用层面，本白皮书进一步阐述如何把安全测评结论转化为采购准入标识，并嵌入医疗 AI 采购的全生命周期，使「安全到什么程度」成为采购决策可核验、可归档、可追责的客观依据。

一、背景与挑战

1.1 医疗 AI 进入高风险临床环节

过去两年，医疗 AI 的应用形态发生了根本性变化。早期的医疗 AI 多为单点、封闭的判别式模型（如影像病灶检出），其输入输出边界清晰、可被传统医疗器械审评框架覆盖。而以大语言模型与智能体为代表的新一代医疗 AI，具备开放式输入、生成式输出与跨工具协作的能力，开始介入问诊、鉴别诊断、用药建议、检查开立、随访管理乃至部分流程的自动执行。这意味着 AI 的潜在影响半径，从「辅助某一判断」扩展到「参与甚至主导临床决策链条」。

1.2 生成式 AI 带来的新型安全风险

生成式与智能体形态引入了传统测评方法难以覆盖的风险类型：

- **幻觉与自信错误：**模型可能编造不存在的指南、剂量或文献，且语气坚定，难以被非专业用户识别；

- **长尾失效：**在罕见病、复杂合并症、特殊人群（孕产妇、儿童、老年、肝肾功能不全）等场景中性能骤降；
- **脆弱性：**对输入扰动、方言口语、错别字、信息缺失或对抗性诱导高度敏感，输出可被轻易操纵；
- **行为不可控：**智能体在多轮交互与工具调用中可能偏离任务、越权操作或在错误状态下持续推进；
- **版本漂移：**底层模型静默更新后，既有的安全表现可能在无感知的情况下退化。

1.3 现有评测范式的不足

当前流行的公开榜单（如各类医学考试题库、问答基准）在衡量「能力上限」上具有一定价值，但在「安全性」维度存在结构性缺陷：题目高度公开，易被纳入训练数据形成污染；以单一准确率为核心，掩盖了长尾失效与危害的非对称性；以静态选择题为主，无法反映智能体在真实工作流中的动态行为；且评测主体往往与被评对象存在利益关联，缺乏中立性。安全性测评必须从「考能力」转向「验风险」，并由独立第三方以受控、保密、可复现的方式实施。

二、测评范式与中立第三方定位

2.1 「度量衡」理念

医疗 AI 安全性测评应被定位为行业基础设施意义上的「度量衡」。度量衡的价值不在于评判某一件商品的好坏，而在于提供统一、稳定、被各方共同信任的刻度，使交易、监管与协作得以在同一语言下进行。安全性测评同理：其首要目标不是宣布某产品「最好」，而是建立一套使「安全到什么程度」可被一致度量、可被横向比较、可被复核追溯的标尺。

2.2 中立性的三重隔离

中立第三方的公信力源于结构性的利益隔离，而非主观承诺。本标准要求测评机构在制度上确立三重隔离：

1. 研发隔离：测评机构不参与被测产品的设计、训练与调优，避免「既当运动员又当裁判」；
2. 利益隔离：测评结论不与被测方的商业利益挂钩，收费与结论解耦，杜绝「付费过评」；
3. 数据隔离：核心金标准测试集对被测方与公众保密、独立维护，防止针对性优化与污染。

2.3 服务双方而不偏向任一方

中立第三方测评同时服务于需求侧（医疗机构、采购与监管部门）与供给侧（模型与产品供应商）。对需求侧，测评提供采购与准入的客观依据，降低信息不对称；对供给侧，测评提供可对标、可改进的反馈，使安全投入获得可验证的回报。中立的本质，是对方法负责、对刻度负责，而非对任何一方负责。

三、测评对象分层：基座模型与医疗智能体

基座模型与医疗智能体是两类性质不同的测评对象，必须采用不同的测评逻辑。混淆二者，是当前医疗 AI 评测中最常见的方法论错误。

维度	基座模型	医疗智能体
测评指向	医学知识与推理的能力上限	在真实工作流中的行为安全
性质	静态、能力导向	动态、行为导向
输入	受控的题目或病例片段	多轮对话、真实或仿真环境状态
关注点	知识正确性、推理、校准、幻觉率	任务完成、工具调用、越权、纠错、人机协同
典型方法	封存题库、回顾性病例	仿真临床环境、红队、流程审计
责任边界	提供能力基础	能力 × 系统设计 × 部署上下文

3.1 基座模型测评：考察能力上限

基座模型测评回答的问题是：模型本身掌握了多少正确的医学知识，其推理与校准是否可靠。它衡量的是潜力与下限，而非最终产品的安全性。一个高分基座模型并不等于一个安全的医疗产品，但一个能力不足的基座模型几乎注定无法支撑安全的应用。基座模型测评聚焦于知识准确性、推理链条的合理性、不确定性校准与幻觉倾向。

3.2 智能体测评：考察行为安全

医疗智能体的安全性不能由其底层模型的能力推断得出。本白皮书主张以如下分解理解智能体安全：

智能体安全 = 基座能力 × 系统设计（提示词、工具、护栏、回退机制） × 部署上下文（场景、用户、监督强度）

这意味着同一基座模型，配以不同的系统设计与部署约束，可能呈现截然不同的安全水平。因此智能体测评必须在尽可能接近真实部署的条件下进行：考察其在多轮交互中是否偏离任务、是否在信息不足时恰当地澄清或拒绝、工具调用是否越权、遇到错误状态能否识别并回退、以及人机交接是否清晰可靠。

四、自主性分级（L0 至 L3）

自主性等级界定了 AI 与人类之间的责任分配，是决定测评严格程度的首要变量。自主性越高，AI 的输出越可能直接转化为对患者的行动，人类纠错的机会越少，测评的安全闸门就越严苛。

等级	名称	定义与人机关系	监督强度
L0	纯信息提供	提供一般性医学信息与参考，不针对特定患者给出个体化建议	信息仅供参考
L1	辅助决策	针对个体给出建议，但人类对每一条输出进行	逐条人工审核

等级	名称	定义与人机关系	监督强度
		全程审核后才采纳	
L2	条件自主	在限定范围内自主产出，人类以例外或事后方式审核	例外或事后审核
L3	高度自主	在严格限定边界内闭环执行，常规情形无需人工介入	边界内闭环、超界升级

分级的关键不在于标榜自主性，而在于使产品声明的自主性与其实际安全能力相匹配。测评中应核验三件事：产品所声称的自主等级、其在该等级下的实际行为表现、以及当超出能力边界时是否能可靠地降级或升级至人工。一个声称 L1 却在实际中绕过人工审核的产品，或一个声称 L3 却无法识别自身边界的产品，都构成严重的安全缺陷。

五、临床场景与风险分级

不同临床场景的风险并不对等。本方法以风险为导向分配测评资源与严格度，避免在低风险场景上过度消耗、在高风险场景上覆盖不足。场景风险由四个因子共同决定：

场景风险 = 危害严重度 × 不可逆性 × 错误不可发现性 × 自主性等级

- **危害严重度：**一旦出错对患者造成伤害的严重程度；
- **不可逆性：**错误一旦发生能否被纠正或挽回；
- **错误不可发现性：**错误在造成后果前被人类察觉的难易程度；
- **自主性等级：**AI 输出转化为实际行动所需的人工介入程度。

5.1 场景风险分层参考

风险等级	典型场景	可逆性	测评严格度
极高	处方与用药决策、剂量计算、危急值处置	低	最高（多重硬性闸门）
高	诊断与鉴别诊断、急诊分诊、检查开立	中	高
中	慢病管理、随访建议、患者教育	中高	中
较低	病历整理与结构化、医学知识检索	高	中（侧重隐私与准确）
低	行政流程、排班、科普问答	高	基础

5.2 为何处方与用药决策为最高风险场景

处方与用药决策同时满足风险四因子的最不利组合：用药错误（错误药品、剂量、给药途径、相互作用或禁忌）可直接造成严重甚至致命伤害；错误一旦执行往往难以逆转；剂量或相互作用的细微错误在外观上与正确处方高度相似，极易逃过人工核验；而在追求效率的部署中，此类决策又常被赋予较高自主性。因此对涉及处方与用药的测评施加最严格的标准，并设置多重一票否决式的硬性安全闸门，包括禁忌识别、剂量边界、相互作用检测与特殊人群处理。

六、安全性测评维度体系

安全性是一个多维构念，不能被任何单一指标代表。本标准的测评维度体系覆盖八个核心维度，每个维度在不同场景与自主等级下被赋予不同权重，但安全相关维度始终具有否决权。

6.1 安全性

核心维度。考察产品避免造成伤害的能力，包括：禁忌与相互作用识别、剂量与给药安全、危急情况的升级与转诊提示、以及在超出能力或证据不足时是否恰当地拒答或寻求人工。安全性的关键不是「答对率」，而是「危害事件率」与「在边界处的恰当退避」。

6.2 准确性与有效性

以金标准为参照衡量输出的正确性。需区分不同错误类型的代价：漏诊与误诊、假阴性与假阳性在医疗中代价高度不对称，应分别报告而非合并为单一准确率。

6.3 鲁棒性

考察在非理想输入下的稳定性：信息缺失、口语方言、错别字、单位混淆、多语言、以及对抗性诱导（如诱导越权、绕过安全约束）。医疗现实中的输入极少是规整的，鲁棒性直接决定真实环境下的安全下限。

6.4 校准与不确定性

考察模型的置信度是否与其实际正确率相符，以及它是否「知道自己不知道」。一个在错误时仍表达高置信的系统，比一个会恰当表达不确定性的系统危险得多。校准是连接能力与安全的关键桥梁。

6.5 公平性

考察性能在不同性别、年龄、地域、人群中的一致性，识别系统性偏差，避免 AI 放大既有的医疗资源与诊疗不平等。

6.6 可解释性与可追溯

考察输出是否提供可供专业人员审查的依据与推理路径，以及关键决策是否可被记录、复盘与追责。可追溯性是责任落地的前提。

6.7 隐私与合规

考察对患者个人信息与健康数据的保护、最小化使用、留存与传输安全，以及对适用法律法规与行业规范的符合性。

6.8 一致性与稳定性

考察相同输入是否产出一致结论（输出稳定性），以及在底层模型更新前后安全表现是否保持稳定（时间稳定性）。版本漂移是医疗 AI 长期安全的隐性风险。

七、测评方法论

7.1 双轨数据集：封存金标准与公开演示集

测评结果的可信度，首先取决于测试数据是否未被污染。本方法采用严格隔离的双轨机制：

- **封存金标准测试集：**由临床专家构建并多方仲裁、独立保密维护、绝不公开、不向被测方披露，用于产出正式测评结论，从根本上对抗刷榜与训练污染；
- **公开演示集：**内容、难度分布与金标准集对齐但完全独立，可对外公开，用于方法演示、能力沟通与自检，但不用于正式定级。

两集严格隔离、定期轮换并监测潜在泄露。一旦金标准集出现污染迹象，相关条目即行退役并由新条目替换，以维持刻度的长期有效性。

7.2 数据来源与专家仲裁

测评数据综合采用真实世界回顾性病例（脱敏）与专家构造的高风险边界案例。每一条目的标准答案由多名具备相应专科资质的临床专家独立标注，分歧通过仲裁机制解决，并记录评者间一致性作为条目质量的内部指标。低一致性条目需复核或退役。

7.3 对抗性与红队测试

针对生成式系统的脆弱性，测评设置专门的红队流程，主动构造诱导越权、绕过安全约束、误导性前提、隐藏关键信息等对抗输入，考察系统在被攻击条件下是否仍能守住安全边界，而非仅在友善输入下表现良好。

7.4 仿真临床环境（面向智能体）

对医疗智能体，静态题库无法反映其动态行为。测评在受控的仿真临床环境中进行：通过患者模拟器与电子病历沙盒重建多轮交互与工具调用场景，观测智能体在完整任务链条中的行为安全、越权检测、错误状态识别与回退、以及人机交接的可靠性。仿真环境使高风险行为得以在不伤害真实患者的前提下被充分暴露。

八、评分体系与安全闸门

8.1 反对单一聚合分数

本标准明确反对以单一聚合分数概括医疗 AI 的安全性。聚合分数会用高频场景的优异表现掩盖高风险场景的致命缺陷，掩盖错误代价的非对称性，制造虚假的安全感。测评结论以分场景、分维度的结构化报告呈现，而非一个排名数字。

8.2 硬性安全闸门（一票否决）

在极高风险场景中，若产品在关键安全项上出现失败，无论其在其他维度表现多么优异，均不得通过相应等级的测评。硬性闸门示例包括（视场景而定）：

- 在用药场景中出现明确的禁忌或严重相互作用漏检；
- 给出可造成严重伤害的剂量或给药途径错误；
- 对危急值或急危重情形未触发升级或转诊提示；
- 在超出能力边界时仍以高置信给出个体化处置建议而不退避；
- 声称的自主性等级与实际行为不符，且无可靠的人工兜底。

8.3 风险加权与分层报告

在通过硬性闸门的前提下，各维度得分按场景风险与自主等级进行风险加权，并以「场景 × 维度 × 自主等级」的矩阵形式分层报告。报告同时给出适用边界声明，明确产品在何种场景、何种自主等级、何种监督强度下的测评结论成立，避免结论被超范围引用。

九、测评治理与结果有效性

9.1 版本绑定

测评结论严格绑定于被测对象的具体构成：模型版本、系统提示词、工具集与关键配置。任一要素变更，原结论即不再自动适用。测评报告中完整记录被测对象的版本指纹，使结论可被精确复核与追溯。版本绑定是后续采购合同得以约束的技术前提。

9.2 重测与上市后监测

医疗 AI 的安全性不是一次性结论。底层模型更新、提示词调整或工具变更，都可能引起安全表现的漂移。本标准主张建立周期性重测机制，并配合上市后监测持续追踪真实部署中的安全事件，使测评从「一次发证」转向「持续守护」。

9.3 利益冲突与测试集泄露防护

测评机构应在制度上执行利益冲突申报与隔离，测评人员与被测方之间不存在影响独立性的利益关系。针对测试集，建立访问审计、内容轮换与泄露监测机制；一旦发现条目外泄或被定向优化，立即退役并替换，保护刻度的长期公信力。

十、安全测评在医疗 AI 采购中的应用

安全测评的价值最终要在采购与准入环节兑现。对医疗机构与采购管理部门而言，本标准提供的不只是一份技术报告，而是一套可嵌入采购全流程、可写入合同、可审计追责的安全决策依据。本章阐述如何把安全测评应用于医疗 AI 的采购实践。

10.1 采购环节的安全信息不对称

医疗 AI 采购区别于传统医疗器械或信息系统采购，其核心困难在于安全信息的严重不对称：产品能力以生成式形式呈现，难以通过短期演示判断其长尾与边界表现；底层模型可在交付后静默更新，使采

购时的判断迅速失效；供应商对自身产品的安全边界往往缺乏充分披露。由此衍生出五类典型的采购风险：

- **榜单刷分误导：**以公开基准的高分作为安全证明，而公开题库易被污染，分数无法反映真实安全；
- **演示特调：**演示场景经过精心挑选与特别优化，不代表真实部署中的表现；
- **版本掉包：**交付或上线运行的版本与演示、测评版本不一致，或上线后未经告知即更新；
- **自主性夸大：**宣称的自主能力超出其实际安全边界，缺乏可靠的人工兜底；
- **责任真空：**安全事件发生后，供应商、平台方与使用机构之间责任不清。

独立第三方安全测评的作用，正是为采购方提供一道可信的风险防线，把上述不对称转化为可核验、可归档、可追责的客观依据。在以机构背书与合规证明为权重的采购评审中，一份载明版本指纹的独立测评报告，恰好提供了采购决策所需的留痕与依据。

10.2 安全分级与采购准入标识

为使测评结论可直接服务于采购决策，本标准将测评结果转化为与适用边界绑定的采购准入标识。标识在给定的「场景 × 自主性等级」组合下给出明确结论：

准入结论	含义	采购建议
准予采用	在该场景与自主等级下通过全部硬性安全闸门，各维度达标	可纳入采购，按合同绑定版本
限定条件采用	总体达标，但存在需控制的残余风险	在附加监督或限定使用范围条件下采用
不予采用	存在未闭合的硬性安全闸门失败	不应在该场景或自主等级下采购使用

关键在于：任何准入结论都严格绑定其适用边界——离开既定场景、自主等级与监督强度，结论即不成立。采购方不得将某一场景的结论外推至更高风险场景，亦不得将低自主等级下的结论用于支撑更高自主性的部署。

10.3 将安全测评嵌入采购全流程

安全测评不是采购链条末端的一次性盖章，而应贯穿采购的全生命周期。各阶段的作用与关键动作如下：

采购阶段	安全测评的作用	关键动作
需求与立项	明确场景风险与自主等级，规格化安全要求	在需求文件写明目标场景、自主等级与最低安全等级
招标与选型	以测评等级作为资格门槛与评分依据	设安全准入为否决项，安全表现纳入技术评分
合同与协议	锁定被测版本，约定变更与重测义务	将版本指纹、重测触发、监测义务写入合同

采购阶段	安全测评的作用	关键动作
验收	以测评结论与版本一致性为验收依据	核验交付版本与测评版本一致
运维与监测	持续追踪安全表现，触发重测	建立安全事件报告与版本变更重测机制

10.4 招标文件中的安全要求设置

建议采购方在招标与评审中遵循以下原则：

- **安全准入作为否决项：**高风险场景须达到规定的最低安全等级，且无未闭合的硬性安全闸门失败，未达标者不进入后续评审；
- **安全表现纳入技术评分：**在准入基础上，将各安全维度的测评表现作为技术分的重要组成，避免「重功能、轻安全」；
- **要求独立第三方报告：**供应商须提交由中立第三方出具、载明被测版本指纹的安全测评报告，而非自评或公开榜单分数；
- **明确自主性与监督方案：**供应商须声明产品自主性等级，并提供与之匹配的人工监督与兜底方案。

10.5 合同条款与持续合规管理

医疗 AI 的快速迭代特性，要求采购从「一次性验收」转向「持续合规管理」。建议在合同与服务协议中至少约定：

- **版本绑定与一致性：**交付及运行版本须与测评版本一致；版本指纹作为验收与履约依据，不一致视为重大违约；
- **变更告知与重测义务：**底层模型、系统提示词或工具集发生变更的，供应商须事先告知并触发重新测评，重测未通过前不得在高风险场景继续使用；
- **上市后监测义务：**建立安全事件的监测、记录与报告机制，约定报告时限与响应措施；
- **安全事件责任与处置：**明确安全事件的分级、上报、处置流程与相应责任。

10.6 差异化采购：匹配场景风险与机构能力

采购的安全要求应同时匹配两个变量：场景风险与使用机构的人机协同能力。

- **按场景风险设定门槛：**高风险场景（如处方、诊断）应要求更高的最低安全等级、更严格的硬性闸门与更强的人工监督，并对高自主性声明施加更审慎的核查；
- **按机构能力限定自主性：**使用机构的监督能力决定可承接的自主性等级。具备充分专业审核能力的机构方可承接较高自主性的产品；监督能力有限的机构（如部分基层医疗机构）应优先采购低自主性、强人工兜底的配置。

换言之，同一产品在不同机构的可采购自主等级可能不同——采购决策必须把「产品能力」与「机构监督能力」一并纳入考量，避免将高自主性产品配置到缺乏相应监督能力的环境中。

10.7 责任划分

清晰的责任边界是采购关系可持续的前提。本标准建议在采购中明确四方责任：

主体	责任范围
供应商	产品安全设计、版本一致性、变更告知与重测配合、安全事件响应
第三方测评机构	测评方法的科学性、结论在适用边界内的可靠性、测评过程的独立与可追溯
采购方	规格化安全要求、将测评结论正确纳入采购、监督合同履约
使用机构	在适用边界与监督方案内使用、落实人工监督、上报安全事件

十一、方法的局限与展望

本白皮书坦承所述方法论存在固有局限，明确这些边界本身即是安全实践的一部分：

- **测评通过不等于部署安全：**测评衡量的是受控条件下的表现，真实部署中的人机交互、组织流程与使用习惯仍可能引入新风险；
- **静态测评不能替代持续监测：**任何时点的测评都是快照，必须辅以上市后监测；
- **金标准本身存在边界：**医学知识在演进，专家共识在更新，金标准需要持续维护与修订；
- **覆盖不可能穷尽：**测评以风险为导向尽力覆盖关键场景，但无法穷举所有长尾情形。

展望未来，编制组将持续投入三个方向：扩充并精炼覆盖更多专科与人群的金标准体系；深化面向智能体的仿真测评与动态行为度量；推动测评方法、采购规范与监管框架的衔接，使「安全到什么程度」成为医疗 AI 行业可共享、可信任的共同语言。编制组愿与医疗机构、供应商、监管与学术界一道，把医疗 AI 的安全建立在可度量、可复现、可追溯的基础之上。

附录 A　术语表

术语	说明
基座模型	提供医学知识与推理能力的底层大模型，测评其能力上限
医疗智能体	在工作流中调用工具、多轮交互、可具备一定自主性的应用系统
自主性等级	界定 AI 与人类责任分配的分级（L0 至 L3），越高越需严格测评
金标准测试集	封存保密、用于产出正式结论的高质量测评数据集
公开演示集	与金标准对齐但独立、可公开、不用于正式定级的数据集
硬性安全闸门	关键安全失败时一票否决的强制性测评条件
校准	模型置信度与实际正确率相符的程度
红队测试	主动构造对抗输入以暴露安全脆弱性的测评方法
版本指纹	标识被测对象具体构成的记录，含模型版本、提示词与工具集
采购准入标识	由测评结论转化、与适用边界绑定的采购结论（准予/限定/不予采用）
上市后监测	产品投入使用后对真实安全事件的持续追踪

附录 B　测评维度速览

维度	核心问题
安全性	是否会造成伤害？边界处是否恰当退避？
准确性与有效性	结论是否正确？不同错误类型的代价是否分别衡量？
鲁棒性	在非理想与对抗输入下是否仍稳定安全？
校准与不确定性	置信度是否可靠？是否知道自己不知道？
公平性	性能在不同人群间是否一致？
可解释性与可追溯	依据是否可审查？决策是否可追责？
隐私与合规	数据是否得到保护？是否符合法规？
一致性与稳定性	相同输入是否一致？版本更新后是否稳定？

附录 C　医疗 AI 采购安全要求清单

本清单供采购方在编制招标文件与合同条款时直接引用或裁剪，按六类要求组织。

一、资格与准入

- 供应商须提交由中立第三方出具的安全测评报告，载明被测对象的版本指纹（模型版本、系统提示词、工具集）。

- 目标为高风险场景的，产品须在该场景与声明自主等级下达到规定的最低安全等级，且无未闭合的硬性安全闸门失败。
- 不接受以公开榜单分数或供应商自评替代独立第三方测评报告。

二、版本与一致性

- 交付及实际运行版本须与测评版本一致；不一致视为重大违约。
- 上线运行期间未经告知的版本变更，构成违约并触发安全复核。

三、变更与重测

- 底层模型、系统提示词或工具集变更的，须于约定时限内事先告知采购方并触发重新测评。
- 高风险场景下，重测未通过前不得继续使用变更后版本。

四、自主性与监督

- 供应商须明确声明产品自主性等级及其适用边界。
- 须提供与自主等级相匹配的人工监督方案与超界升级（人工兜底）机制。

五、监测与责任

- 须建立上市后安全事件的监测、记录与报告机制，并约定报告时限。
- 须明确安全事件的分级、处置流程与各方责任划分。

六、隐私与合规

- 须提供患者数据保护与合规符合性证明，落实数据最小化与传输、存储安全。