深度学习辅助决策医疗器械软件审评要点

一、适用范围

本审评要点适用于深度学习辅助决策医疗器械软件（含独立软件、软件组件）的注册申报。深度学习辅助决策医疗器械软件（以下简称软件）即基于医疗器械数据（医疗器械所生成的医学图像、医学数据，以下统称数据），使用深度学习技术进行辅助决策的软件。其中，“基于医疗器械数据”是指单独使用医疗器械数据，或者联合使用医疗器械数据与非医疗器械数据；“辅助决策”是指通过提供诊疗活动建议辅助医务人员进行临床决策。

使用深度学习技术进行前处理（如成像质量改善、成像速度提升、图像重建）、流程优化（如一键操作）、常规后处理（如图像分割、数据测量）等非辅助决策的软件可参考使用本审评要点。使用传统机器学习技术的软件亦可参考使用本审评要点。

本审评要点遵循《医疗器械软件注册技术审查指导原则》（以下简称软件指导原则）、《医疗器械网络安全注册技术审查指导原则》（以下简称网络安全指导原则）、《移动医疗器械注册技术审查指导原则》（以下简称移动器械指导原则）等相关指导原则要求。

本审评要点不含人工智能伦理、数据产权等法律法规层面要

求，但生产企业应当在软件全生命周期过程中考虑相关规定。

二、审评关注重点

从发展驱动要素角度讲，深度学习实为基于海量数据和高算力的黑盒算法。本审评要点重点关注软件的数据质量控制、算法泛化能力、临床使用风险，临床使用风险应当考虑数据质量控制、算法泛化能力的直接影响，以及算力所用计算资源（即运行环境）失效的间接影响。

基于风险的全生命周期管理是此类软件监管的基本方法，相关考量详见软件指导原则、网络安全指导原则、移动器械指导原则以及医疗器械生产质量管理规范独立软件附录。下面结合审评关注重点分别阐述软件风险管理、软件设计开发、软件更新等方面考量。

软件风险管理活动应当基于软件的预期用途（目标疾病、临床用途、重要程度、紧迫程度）、使用场景（适用人群、目标用户、使用场所、临床流程）、核心功能（处理对象、数据兼容性、功能类型）予以实施，并贯穿于软件全生命周期过程。软件临床使用风险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；假阳性即误诊，可能导致后续不必要的诊疗活动。进口软件除考虑假阳性和假阴性风险外，还应当考虑中外人种、流行病学特征、临床诊疗规范等方面差异的影响及其风险。生产企业

应当采取充分的、适宜的、有效的风险控制措施以保证软件的安全性和有效性。

软件典型设计开发过程通常可分为需求分析、数据收集、算法设计、验证与确认等阶段。

（一）需求分析

需求分析应当以软件的临床需求与使用风险为导向，结合软件的预期用途、使用场景和核心功能，综合考虑法规、标准、用户、产品、数据、功能、性能、接口、用户界面、网络安全、警示提示等方面需求，重点考虑数据收集、算法性能、临床使用限制等方面要求。

数据收集应当考虑数据来源的合规性和多样性、目标疾病流行病学特征、数据质量控制要求（详见下节）。数据来源应当在合规性基础上保证数据多样性，以提高算法泛化能力，如尽可能来自多家、不同地域、不同层级的代表性临床机构，尽可能来自多种、不同采集参数的采集设备。目标疾病流行病学特征包括但不限于疾病构成（如分型、分级、分期）、人群分布（如健康、患者，性别、年龄、职业、地域、生活方式）、统计指标（如发病率、患病率、治愈率、死亡率、生存率）等情况，以及目标疾病并发症与类似疾病的影响情况。

算法性能应当考虑假阴性与假阳性（指标、关系）、重复性与再现性、鲁棒性/健壮性等要求。

临床使用限制应当考虑临床禁用、慎用等场景。

（二）数据收集

数据收集应当考虑数据采集、数据预处理、数据标注、数据集构建等活动的质控要求，以保证数据质量和算法设计质量。

1.数据采集

数据采集主要由临床机构实施，应当考虑采集设备、采集过程以及数据脱敏的质控要求。

采集设备质控应当明确采集设备的兼容性要求和采集要求。兼容性要求应当基于数据生成方式（直接生成、间接生成）提供采集设备兼容性列表或技术要求，明确采集设备的制造商、型号规格、性能指标等要求，若对采集设备无具体要求应当提供相应支持资料。采集要求应当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）等要求。

采集过程质控应当建立数据采集操作规范，明确采集人员要求和采集过程要求。采集人员要求包括人员的选拔、培训、考核。采集过程要求包括人员职责、采集流程（如采集步骤、操作要求）。

若使用现有历史数据，应当明确采集设备要求、数据采集质量评估要求（如人员、方法、指标、通过准则）。

采集的数据应当进行数据脱敏以保护患者隐私。数据脱敏应

当明确脱敏的类型（静态、动态）、规则、程度、方法。

2.数据预处理

脱敏数据由临床机构转移至生产企业形成原始数据库，不同模态的数据在原始数据库中应当加以区分（下同）。

数据预处理应当基于原始数据库考虑数据处理、数据清洗的质控要求。数据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。数据清洗应当明确清洗的规则、方法。

数据处理和清洗应当明确选用软件工具的名称、型号规格、完整版本、供应商、运行环境、确认等要求，同时考虑数据处理选用方法对软件的影响及其风险。

数据经预处理后形成基础数据库，应当明确样本类型、样本量、样本分布等信息。样本类型以适用人群为单位可分为数据序列（由多个单一数据组成，如结构序列、功能序列、时间序列）、单一数据。样本量应当明确样本规模及确定依据，需要考虑样本量不足对软件的影响及其风险。样本分布应当依据疾病构成、适用人群、数据来源机构、采集设备、样本类型等因素明确数据分布情况，需要考虑数据偏性对软件的影响及其风险。

3.数据标注

数据标注应当考虑标注资源管理、标注过程质控、标注质量评估等要求。

标注资源管理包括人员管理和基础设施管理。人员管理应当

明确标注人员和仲裁人员的选拔（如职称、工作年限、工作经验、所在机构，若有国外人员应当明确其资质要求）、培训、考核（如方法、频次、指标、通过准则，其中指标应当包括重复性、再现性）等要求。基础设施管理应当明确标注场所（真实或模拟，环境、照明条件）、标注软件（名称、型号规格、完整版本、供应商、运行环境、确认）等要求。

标注过程质控应当建立数据标注操作规范，明确标注人员（如资质、数量、职责）、标注流程（如标注对象、标注形式、标注轮次、标注步骤、操作要求）、临床诊疗规范（如临床指南、专家共识）、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。

标注质量评估应当明确人员、方法、指标、通过准则等要求。

数据经标注后形成标注数据库，其样本类型可分为数据序列、单一数据（由多个数据块组成）、数据块（图像区域、数据片段）。样本量、样本分布等要求及风险考量与基础数据库相同。

4.数据集构建

基于标注数据库构建训练集（用于算法训练）、调优集[1]（若有，用于算法超参数调优）、测试集（用于算法性能评估），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。训练集应当保证样本分布具有均衡性，测试集、调优集应当

[1] 机器学习领域称之为验证集（Validation set）。为避免与医疗器械领域所用术语验证（Verification）、确认（Validation）相混淆，本审评要点将其改称为调优集。

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。

为解决数据样本分布不满足预期目标的问题，可对训练集、调优集小样本量数据进行扩增；测试集不宜进行数据扩增，若扩增应当分析对软件的影响及其风险。数据扩增应当明确扩增的方式（离线、在线）、方法（如翻转、旋转、镜像、平移、缩放、滤波等）、倍数，并考虑扩增方法选用以及扩增倍数过大对软件的影响及其风险。

数据经扩增后形成扩增数据库，应当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）等方面的差异，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

算法设计应当考虑算法选择、算法训练、网络安全防护、算法性能评估等活动的质控要求。建议数据驱动与知识驱动相结合进行算法设计，以提升算法可解释性。

1.算法选择

算法选择应当明确所用算法的名称、结构（如层数、参数规模）、流程图、现成框架（如 Tensorflow、Caffe）、输入与输出、运行环境、算法来源依据（或注明原创）等信息。同时应当明确算法选择与设计的原则、方法和风险考量，如量化误差、梯度消失、过拟合、白盒化等。

若使用迁移学习技术，除上述内容外还应当补充预训练模型的数据集构建、验证与确认等总结信息。

2.算法训练

算法训练需要基于训练集、调优集进行训练和调优，应当明确评估指标、训练方法、训练目标、调优方法、训练数据量-评估指标曲线等要求。

评估指标建议根据临床需求进行选择，如敏感性、特异性等。训练方法包括但不限于留出法和交叉验证法。训练目标应当满足临床要求，提供 ROC 曲线等证据予以证实。调优方法应当明确算法优化策略和实现方法。训练数据量-评估指标曲线应当能够证实算法训练的充分性和有效性。

3.网络安全防护

网络安全防护应当结合软件的预期用途、使用场景和核心功能，基于保密性、完整性、可得性等网络安全特性，确定软件网络安全能力建设要求，以应对网络攻击和数据窃取等网络威胁。相关要求详见网络安全指导原则。

此类软件常见网络威胁包括但不限于框架漏洞攻击、数据污染，其中框架漏洞攻击是指利用算法所用现成框架本身漏洞进行网络攻击，数据污染是指通过污染输入数据进行网络攻击。

4.算法性能评估

算法性能评估作为软件验证的重要组成部分，需要基于测试

集对算法设计结果进行评估，应当明确假阴性与假阳性、重复性与再现性、鲁棒性/健壮性等评估要求，以证实算法性能满足算法设计要求。

同时，应当分析算法性能影响因素及其影响程度，如采集设备、采集参数、疾病构成、病变特征等因素影响，以提升算法可解释性，并作为软件验证、软件确认的基础。

（四）验证与确认

1.软件验证

软件验证是指通过提供客观证据认定软件开发、软件更新某一阶段的输出满足输入要求，包括软件验证测试（单元测试、集成测试、系统测试）、设计评审等系列活动。

软件验证应当明确法规、标准、用户、产品、数据、功能、性能、接口、用户界面、网络安全、警示提示等测试要求，以验证软件的安全性和有效性，并作为软件确认的基础。

2.软件确认

软件确认是指通过提供客观证据认定软件满足用户需求和预期目的，包括软件确认测试（用户测试）、临床评价、设计评审等系列活动，其中软件确认测试应当基于软件需求在真实或模拟使用场景下予以实施。

（1）基本原则

临床评价是此类软件进行软件确认的主要方式，相关要求详

见《医疗器械临床评价技术指导原则》。根据软件指导原则要求，软件应当提交基于临床试验的临床评价资料，即提交申报产品的临床试验资料，或者与申报产品核心算法具有实质等同性的同品种产品或同类软件功能的临床试验资料。

进口软件应当提供中外人种、流行病学特征、临床诊疗规范等方面差异影响的临床评价资料，若不足以证实申报产品在中国使用的安全性和有效性，应当在中国开展临床试验。使用境外临床试验数据应当满足《接受医疗器械境外临床试验数据技术指导原则》要求。

（2）临床试验

临床试验应当符合《医疗器械临床试验质量管理规范》要求。可参照《医疗器械临床试验设计指导原则》，基于软件的预期用途、使用场景和核心功能进行试验设计，确定观察指标、样本量估计、入排标准、随访以及实施机构等要求，以确认软件的安全性和有效性。

建议优先选择同品种产品或临床参考标准（即临床金标准）进行非劣效对照设计，若无同品种产品且难以获取临床参考标准（如违背伦理学要求）可选择替代方法，如选择用户结合软件联合决策与用户单独决策进行优效对照设计。非劣效界值或优效界值的确定应当有充分的临床依据。此外考虑到用户的差异性，可选择多阅片者多病例（MRMC）试验设计。

建议结合适用人群、病变等层面选择观察指标，原则上选择敏感性、特异性、ROC/AUC 作为主要观察指标，亦可在此基础上根据软件特点选择敏感性/特异性衍生指标、ROC/AUC 衍生指标、组内相关系数、Kappa 系数、时间效率、数据有效使用率等指标作为观察指标。

入排标准应当基于目标疾病流行病学特征，保证阳性样本和阴性样本选取的合理性和充分性。

建议临床试验结果由第三方独立评价。

实施机构应当具备代表性和广泛性，不同于训练数据主要来源机构，地域分布尽可能广泛，机构数量尽可能多，以确认算法泛化能力。

例如，预期以提高辅助诊断时间效率为首要目标的某软件，无同品种产品且难以获取临床参考标准，其临床试验设计可选择用户结合软件联合决策与用户单独决策进行交叉对照设计，以敏感性、特异性、时间效率作为主要观察指标，其中敏感性、特异性可为非劣性对照，时间效率指标应当为优效对照。

（3）回顾性研究

临床评价可采用基于现有历史数据的回顾性研究。回顾性研究应当在设计时考虑并必须严格控制偏倚，如选择偏倚、临床参考标准偏倚、测量偏倚、记忆偏倚等。回顾性研究原则上应当包含多个不同地域临床机构（非训练数据主要来源机构）的同期数

据，结合分层分析、第三方独立评价等方法控制偏倚，以保证真实、准确评价软件的安全性和有效性。

回顾性研究应当基于软件安全性级别考虑使用问题。对于安全性级别为 C 级的高风险软件，原则上应当开展临床试验，此时回顾性研究可用作临床预试验，为临床试验设计提供参考依据，或者在少见亚组病例入组时间过长等情况下，用作临床试验的补充。对于安全性级别为 B、A 级的中低风险软件，回顾性研究可用作临床预实验或替代临床试验。

软件安全性级别应当基于软件的预期用途、使用场景和核心功能进行综合判定，判定方法详见软件指导原则。例如，预期用于病理图像辅助筛查或者危重疾病辅助识别的软件，其安全性级别通常为 C 级。

三、软件更新

（一）基本原则

软件更新应当考虑对软件安全性和有效性的影响，包括正面影响和负面影响。若为重大软件更新（即影响到软件安全性或有效性的软件更新）应当申请许可事项变更，若为轻微软件更新（即未影响软件安全性和有效性的软件更新）则无需申请许可事项变更，通过质量管理体系进行控制。

（二）重大软件更新

除软件更新基本类型外，此类软件常见更新类型又可分为算

法驱动型和数据驱动型。其中，算法驱动型软件更新是指软件所用算法、算法结构、算法流程、现成框架、输入与输出等发生改变，包括算法重新训练（即弃用原有训练数据）；数据驱动型软件更新是指仅由训练数据量增加而促使软件发生更新，实为算法驱动型软件更新的特殊情况。

算法驱动型软件更新通常属于重大软件更新。数据驱动型软件更新是否属于重大软件更新原则上以算法性能评估结果为准，若算法性能评估结果发生显著性改变（即与前次注册所批准的算法性能评估结果相比存在统计学显著差异）则属于重大软件更新。其他类型重大软件更新的判定准则详见软件指导原则、网络安全指导原则相关要求。

（三）验证与确认

无论何种软件更新，均应当按照质量管理体系的要求，开展与软件更新类型、内容和程度相适宜的验证与确认活动。

对于算法驱动型软件更新和数据驱动型软件更新，应当开展算法性能评估、临床评价等验证与确认活动，以保证软件更新的安全性和有效性。

软件更新临床评价应当与软件安全性级别相适宜。对于安全性级别为 C 级的高风险软件，适用范围实质变更原则上应当开展临床试验，其他变更情况可使用回顾性研究进行软件更新临床评价；对于安全性级别为 B、A 级的中低风险软件，可使用回顾

性研究进行软件更新临床评价。

（四）软件版本命名规则

软件版本命名规则应当涵盖算法驱动型软件更新和数据驱动型软件更新，明确并区分重大软件更新和轻微软件更新，其中重大软件更新应当列举全部典型情况。

四、相关技术考量

（一）适用范围扩展

1.基本原则

软件所含全部深度学习、传统机器学习功能（以下统称软件功能）均应当开展需求分析、数据收集、算法设计、验证与确认等活动，且每项软件功能应当分别开展需求分析、数据收集、算法设计、验证与确认等活动。

2.深度学习非辅助决策软件功能

对于深度学习非辅助决策软件功能，其验证与确认要求如下：前处理软件功能原则上应当开展算法性能评估、临床评价；流程优化软件功能开展算法性能评估即可，无需开展临床评价；常规后处理软件功能原则上开展算法性能评估即可，全新功能应当开展临床评价。此时临床评价可参照传统医疗器械评价方法。

3.传统机器学习软件功能

传统机器学习技术与深度学习技术的主要区别在于：前者特征提取通常需要人为干预，而后者自动完成特征提取。因此，对

于传统机器学习辅助决策软件功能，应当明确特征提取信息，包括但不限于特征分类（如人口统计学、生物学、形态学）、特征属性（如形态、纹理、性质、尺寸、边界）和特征展现方式（如形状、尺寸、边界、颜色、数量）。

对于传统机器学习非辅助决策软件功能，其要求参照深度学习非辅助决策软件功能，同时明确特征提取信息。

（二）第三方数据库

第三方数据库可视为回顾性研究的一种特殊形式，可用于算法性能评估，但其类型、用途等情况各不相同，未必能够完全满足软件确认测试的要求。因此，使用第三方数据库进行软件确认测试，应当评估其满足软件确认测试条件的充分性、适宜性和有效性。

可用于软件确认测试的第三方数据库（以下简称测评数据库）应当满足数据平台建设的通用要求（如网络与数据安全等，不再赘述）和专用要求，其中专用要求包括：

1.权威性：考虑到数据质量主要取决于数据标注质量，因此测评数据库创建单位应当包括相应临床专业领域的权威机构（如国家临床医学研究中心），数据标注人员、标注分歧仲裁人员应当分别具备适宜的、丰富的临床实践经验。

2.科学性：为保证能够真实、准确的反映临床实际情况，测评数据库样本量应当通过统计学计算确定以控制抽样误差，样本

分布应当符合目标疾病的流行病学特征情况，不能进行数据扩增；单次测试所用数据量应当予以规定，测试数据应当根据测评数据库样本分布情况进行等比例随机抽取。

3.规范性：测评数据库的数据采集、数据脱敏、数据处理、数据清洗、数据标注、数据管理、网络安全防护等数据治理活动以及测评过程均应当建立质控程序并形成文件，并满足可追溯性要求。

4.多样性：测评数据库的数据应当来源于多个临床机构，以保证测评数据库能够用于评价算法泛化能力；在满足伦理学要求的前提下可包含适当比例的对抗数据样本，以用于评价算法的鲁棒性/健壮性。

5.封闭性：为保证能够充分、客观的评价算法质量，测评数据库应当封闭管理，且样本量应当远大于单次测试所用数据量；测评过程同样应当保证封闭性。

6.动态性：测评数据库应当定期更换一定比例的数据，以保证测评数据库具有持续的多样性和封闭性；被更换的数据可用于构建公开数据库以服务于行业发展。

此外，第三方公开数据库（以下简称公开数据库）因不具备封闭性而不能用作测评数据库，但可用于算法性能评估。公开数据库不宜用于算法训练，若用于算法训练应当评估其使用的适宜性和有效性。

（三）网络与数据安全过程控制

除考虑软件自身网络安全能力建设外，还应当在软件全生命周期过程中考虑网络与数据安全过程控制要求，包括上市前设计开发阶段和上市后使用阶段。

脱敏数据由临床机构转移至生产企业应当明确数据转移方法、数据污染防护措施。数据预处理、数据集构建、算法训练、算法性能评估、软件验证等内部活动应当在封闭的网络环境下开展，以防止数据污染。数据标注、软件确认等涉及外方的活动若在开放的网络环境下开展，应当明确网络安全防护措施，以防止数据污染。数据采集、上市后使用应当考虑与临床机构网络与数据安全要求相衔接的接口问题。

各数据库（集）应当进行数据备份以保证数据安全，数据备份应当明确备份的方法、频次以及数据恢复方法。

（四）云计算服务与移动计算终端

使用云计算服务应当明确服务模式、部署模式、核心功能、数据接口、网络安全能力和服务（质量）协议等要求。使用移动计算终端应当结合终端的类型、特点和使用风险明确相应性能指标要求。相关要求详见移动器械指导原则。

云计算服务与移动计算终端的网络安全要求详见网络安全指导原则。

五、注册申报资料说明

注册申报资料应当在相关公告基础上满足软件指导原则、网络安全指导原则、移动器械指导原则等相关指导原则要求。辅助决策软件还应当考虑下述要求，不适用项应当提供合理解释。非辅助决策软件可参照辅助决策软件的适用要求。

（一）产品名称

辅助决策独立软件产品名称应当符合独立软件通用名称命名规范要求，体现处理对象（如CT图像、眼底照片）、目标疾病（含病变、疾病属性）、临床用途（如辅助筛查、辅助识别）等特征词。

软件组件相应辅助决策软件功能名称可参照辅助决策独立软件要求。

（二）适用范围

辅助决策独立软件适用范围应当明确预期用途、使用场景和核心功能，包括但不限于处理对象、目标疾病、临床用途、适用人群、目标用户、使用场所、采集设备要求、临床使用限制。

软件组件相应辅助决策软件功能适用范围可参照辅助决策独立软件要求，并在产品适用范围中予以体现。

（三）研究资料

除软件描述文档、网络安全描述文档、软件版本命名规则外，研究资料还应当提供以下资料：

软件描述文档核心算法部分应当结合本审评要点提供相应

算法研究资料，包括数据来源合规性声明、算法性能影响因素分析资料以及各类测试场景下算法性能评估结果比较分析资料。

研究资料“其他资料”应当提供网络与数据安全过程控制研究资料，包括公开数据库、测评数据库的基本信息（如名称、创建者、数据量、数据分布）和使用情况。

对于公开数据库，若用于算法训练，使用情况应当明确数据使用量、数据分布、训练集所占比例，并提供其满足算法训练要求的评估资料；若用于算法性能评估，使用情况应当明确数据使用量、数据分布、测试集所占比例、评估指标与结果。

对于测评数据库，若用于算法性能评估，使用情况应当明确评估指标与结果；若用于软件确认测试，使用情况应当提供其满足软件确认测试条件要求的评估资料。

其他类型第三方数据库申报资料参照公开数据库、测评数据库适用要求。

（四）说明书

说明书应当符合《医疗器械说明书和标签管理规定》要求。

辅助决策软件说明书应当明确软件的适用范围、临床使用限制、注意事项、用户培训、采集设备要求、数据采集操作规范、输入与输出、算法性能评估总结（测试集基本信息、评估指标与结果）、软件临床评价总结（临床数据基本信息、评价指标与结果）、运行环境等内容。

深度学习辅助决策软件说明书除上述内容外还应当补充算法训练总结信息（训练集基本信息、训练指标与结果）。

前期已开发软件若不满足本审评要点的适用要求，应当开展差距分析并进行必要限定。

总之，技术审评将基于审评关注重点综合权衡软件的风险和受益，系统评价软件的安全性和有效性，协调上市前与上市后的监管要求，兼顾公众健康保护与促进技术创新的关系。