医学人工智能领域的隐私保护、安全性与临床验证：研究报告

I. 问题概要

医学人工智能（AI）的飞速发展为医疗保健带来了前所未有的机遇，但同时也伴随着对患者隐私、系统安全和临床有效性的严峻挑战。这三大支柱——隐私保护、安全性、临床验证——构成了医学 AI 伦理和实践的基石，其重要性日益凸显。

A. 医学 AI 中隐私、安全与临床验证的核心定义

1. **隐私保护 (Privacy)**：在医学 AI 领域，隐私保护的核心在于保障个体对其个人健康信息（Personal Health Information, PHI）的控制权。这包括在 AI 模型的训练、测试和部署全过程中，确保患者数据的机密性，防止未经授权的访问、使用或泄露 [1]。关键实践包括获取患者的知情同意、对数据进行匿名化或去标识化处理，以及严格遵守如《健康保险流通与责任法案》（HIPAA）和《通用数据保护条例》（GDPR）等法规要求 [1]。医学 AI 模型开发所需的海量数据集 [4]，若管理不当，极易引发隐私风险。值得注意的是，AI 时代对隐私的理解已超越传统数据访问控制，延伸至防范通过模型逆向工程等手段实现的再识别风险 [6]。
2. **安全性 (Security)**：安全性涵盖了保护医学 AI 系统及其相关数据免受各类威胁、破坏和未经授权的修改 [1]。这不仅包括传统的网络安全措施，如数据加密、访问控制，以防止数据泄露，也特别强调确保 AI 模型自身输出的完整性和可靠性，防范针对模型的恶意操纵，例如对抗性攻击 [8]。医学 AI 系统因其数据丰富性及对患者护理的关键作用，已成为网络攻击的高价值目标 [3]。AI 模型本身也可能成为新的攻击媒介，使得安全性问题更加复杂。
3. **临床验证 (Clinical Validation)**：临床验证是指对医学 AI 工具进行系统性、严格的评估过程，以确保其在真实医疗环境中的安全性、有效性、准确性，并能带来切实的临床效益 [11]。这通常涉及监管机构（如 FDA、EMA）的审批，遵循如 CONSORT-AI 等报告指南进行临床试验，以及持续的上市后监测 [14]。AI 工具直接影响临床决策和患者结局，因此其性能必须得到充分验证，以防范潜在伤害并确保获益 [5]。对于能够持

续学习和适应的 AI 系统以及处理多模态输入的 AI，临床验证的方法和标准亦在不断演进 [4]。

B. 三大支柱的内在关联与核心地位

隐私保护、安全性和临床验证并非孤立存在，而是相互交织、紧密联系的。例如，一次安全漏洞（如对抗性攻击）不仅可能泄露患者隐私，还可能导致错误的临床输出，从而使临床验证结果失效 [10]。即使 AI 系统在隐私和名义上的安全措施方面做得很好，如果缺乏充分的临床验证，也可能导致部署不安全或无效的 AI 工具 [12]。透明度，作为可解释性 AI（XAI）的重要组成部分，对这三者都至关重要：理解数据如何被使用（隐私）、识别模型潜在的脆弱点（安全）以及洞察模型的决策过程（验证）[21]。

全面应对这三大挑战，是建立患者、临床医生和监管机构信任的基础，这对于医学 AI 的广泛采纳和成功至关重要 [3]。解决这些问题不仅在于规避负面后果（如医疗事故、数据泄露），更是释放医学 AI 巨大潜力的前提。若缺乏信任和安全保障，AI 在诊断、预后和治疗方面的先进能力将难以得到充分利用，甚至会遭到抵制。因此，对这些核心问题的有效管理，从风险规避的视角转向价值创造的视角，是推动医学 AI 健康发展的关键。

II. 问题背景及应用价值

A. 医学 AI 的兴起及其变革潜力

医学 AI 经历了从早期基于规则的专家系统（如 20 世纪 70 年代的 MYCIN）到如今能够执行复杂任务（如诊断、影像分析、预测分析和药物研发）的精密深度学习模型的演进 [5]。其应用已广泛渗透到放射学、病理学、心脏病学、肿瘤学等多个医学领域，展现出在提升准确性、效率、早期疾病检测以及实现个性化医疗方面的巨大潜力 [5]。特别是多模态 AI，通过整合不同类型的数据（如影像、基因组学、临床记录），进一步放大了这种潜力，为疾病的诊断和治疗提供了更全面的视角 [30]。

B. 解决隐私、安全与临床验证问题的至关重要性

1. **患者安全与信任**：由于验证不足或安全漏洞导致的 AI 决策失误，可能直接对患者造成伤害（如误诊、错误治疗方案）[10]。隐私泄露则会侵蚀患者对医疗服务提供者和 AI 技术的信任，可能导致患者不愿分享其健康数据——而这些数据对于 AI 发展和个性化医疗至关重要[1]。调查显示，高达九成的患者认为隐私是一项权利，且 75%的民众对健康数据隐私保护表示担忧[3]。
2. **伦理的 AI 部署**：确保 AI 系统的公平性、无伤害性、受益性和自主性是核心伦理要求[1]。隐私保护是基本伦理原则之一；安全的系统能防止恶意使用；经过验证的系统则能确保其有益且不造成伤害。此外，必须警惕和避免算法偏见。如果 AI 系统未经不同人群的充分验证，或其训练数据本身就带有历史偏见，算法偏见就可能被放大，这与临床验证中使用的数据质量密切相关[1]。
3. **法规遵从性**：严格遵守数据保护法规（如 HIPAA、GDPR、CCPA）是强制性的，违规将面临严厉处罚[1]。AI 系统的设计从一开始就必须将合规性纳入考量。同时，医学 AI 工具常被归类为医疗器械，需要通过监管审批（如 FDA、EMA），这必然要求提供强有力的临床验证数据[11]。
4. **医学 AI 的可持续创新与采纳**：缺乏信任或无法证明其安全性和有效性，将严重阻碍临床医生和患者对 AI 技术的采纳，从而扼杀创新[21]。清晰的验证框架以及应对隐私安全风险的机制，能为开发者提供更明确的市场准入路径，从而鼓励投资和研发[12]。其应用价值在于使 AI 能够充分发挥其潜力：提高诊断准确性[28]、实现个性化治疗[5]、提升医疗效率[18]，并最终改善患者结局和公共卫生水平[5]。

解决这些核心问题的"应用价值"远不止于个体患者的护理，更延伸至系统性的医疗保健改进。例如，经过验证且安全的 AI 能够优化医疗资源分配[26]，降低医疗成本（例如，通过减少不必要的手术或缩短住院时间[29]），并加强公共卫生监测能力[5]。这种多层次的价值实现，依赖于对隐私、安全和临床验证问题的有效管理。

此外，一个潜在的积极反馈循环值得关注：当医学 AI 在隐私保护、安全性及临床有效性方面表现出稳健性时，公众和专业人士的信任度会随之提升。这种信任反过来会促进更大范围（在知情同意的前提下）和更多样化的数据共享，从而为 AI 模型的改进提供更优质

的“养料”[28]。更优的模型将带来更好的临床效果和更高的应用价值[5]，而这些积极的成果和健全的治理（涵盖隐私、安全、验证）又会进一步巩固信任。这是一个良性循环。反之，任何一个环节的重大失误，如大规模隐私泄露或因安全缺陷导致的临床 AI 应用失败，都可能严重打击信任，使整个领域的发展受挫。

最后，问题的背景不仅是技术性的，也具有深刻的社会文化维度。公众对 AI 的认知、现存的医疗不平等状况以及不同群体的数字素养水平，都会影响这些挑战如何被感知和应对。例如，对 AI 可能加剧健康差距的担忧[5]，与临床验证中的数据偏倚问题以及从弱势群体收集数据时的隐私顾虑紧密相关。因此，解决隐私、安全，特别是涉及公平性评估的临床验证问题，需要一种社会技术综合方法，而非单纯的技术路径。这与伦理 AI 部署的理念一脉相承。

III. 研究现状及瓶颈

A. 医学 AI 中的隐私保护

1. 当前技术与研究现状：

   为应对医学 AI 带来的隐私挑战，研究界和产业界已探索并应用了多种隐私增强技术（Privacy-Enhancing Technologies, PETs）。

   - **联邦学习 (Federated Learning, FL)**：允许在不共享原始患者数据的情况下，跨多个去中心化数据集协作训练模型。各参与方仅共享模型更新（如梯度或参数），而非敏感数据本身[39]。这对于医院等机构间无法直接共享患者记录的场景尤为重要[40]。例如，Owkin 公司的 Owkin Connect 平台支持多机构间的联邦学习研究[41]。当前的研究焦点包括提升联邦学习对抗梯度反演等攻击的安全性、处理非独立同分布（non-IID）数据以及减少通信开销[40]。
   - **同态加密 (Homomorphic Encryption, HE)**：允许直接对加密数据进行计算，使得数据在处理过程中始终保持加密状态，从而保护数据隐私[2]。研究方向主要集中在降低其巨大的计算开销，提高实用性，例如 FAS 方法结合了选择性同态加密与其他 PETs 以平衡隐私与效率[40]。
   - **差分隐私 (Differential Privacy, DP)**：通过向数据或模型输出中添加统计噪声，

使得攻击者难以从聚合结果中区分或识别出单个个体的信息，从而提供形式化的隐私保障 [2]。差分隐私常被应用于联邦学习中，以保护模型更新的隐私 [40]。研究难点在于如何在保证隐私强度的同时，最大限度地保留模型效用（即隐私-效用权衡），以及如何有效地调整差分隐私的参数 [44]。

- **匿名化/去标识化 (Anonymization/De-identification)**：通过移除或遮蔽个人可识别信息（Personally Identifiable Information, PII）来保护隐私 [1]。当前研究致力于发展更强大的技术以抵抗高级的再识别攻击 [1]。
- **安全多方计算 (Secure Multi-Party Computation, SMPC)**：允许多个参与方在不泄露各自私有输入的前提下，共同计算一个函数的结果 [2]。
- **混合方法 (Hybrid Approaches)**：实践中常常结合多种 PETs 的优势，例如联邦学习与同态加密、差分隐私的结合（如 FAS 模型 [40] 和 SMHEA 算法 [39]）。

2. 瓶颈与挑战：

尽管 PETs 取得了显著进展，但在医学 AI 领域的应用仍面临诸多挑战：

- **再识别风险 (Re-identification Risks)**：即使是经过匿名化或聚合处理的数据，在面对复杂的模型反演攻击时，仍可能存在再识别风险，尤其对于基因数据或罕见病等具有高度独特性的数据集 [1]。
- **计算开销与可扩展性 (Computational Overhead and Scalability)**：同态加密和某些安全多方计算技术计算量巨大，阻碍了其在实时应用和大规模场景下的部署 [40]。
- **数据效用与隐私保护的权衡 (Data Utility vs. Privacy Trade-off)**：更强的隐私保护措施（如在差分隐私中加入更多噪声）往往会降低数据质量和模型性能（如准确率）[44]。如何在两者间取得最佳平衡是核心难题。
- **确保知情同意的复杂性 (Ensuring Informed Consent)**：在涉及多源数据融合和数据二次利用（如用于后续研究）的复杂数据生态系统中，获取并有效管理真正意义上的知情同意极具挑战性 [1]。
- **标准化与互操作性缺乏 (Lack of Standardization and Interoperability)**：PETs 的实现方式缺乏统一标准，可能阻碍不同系统或机构间的协作和数据安全共享 [41]。

- **实施的复杂性 (Complexity of Implementation)**：部署和管理这些高级密码学技术需要高度专业化的知识，并非所有医疗机构都具备相应能力 [50]。
- **监管模糊性 (Regulatory Ambiguity)**：尽管已有 GDPR 和 HIPAA 等法规，但其如何具体适用于新兴的 AI 技术和不断发展的 PETs 仍存在不明确之处，给开发者带来不确定性 [1]。

一个重要的观察是，隐私增强技术的发展与破解隐私的方法之间存在一种持续的"军备竞赛"动态。当 PETs 变得越来越复杂（例如 FAS 结合了 HE、DP 和数据置乱 [40]），旨在推断敏感数据的模型反演或基于梯度的攻击技术也在不断进步 [6]。这意味着没有一种 PETs 可以一劳永逸地解决问题；深度防御和自适应的隐私策略是必要的。在这种背景下，"目的限制"原则 [2] 和健全的知情同意机制 [1] 变得尤为关键，因为纯技术解决方案本身存在固有的局限性和权衡。

表 1：医学 AI 中的关键隐私增强技术（PETs）

技术	核心原理/机制	医学 AI 应用示例	隐私保护主要优势	主要挑战/局限性	相关文献
联邦学习 (Federated Learning, FL)	多方协作训练模型，仅交换模型参数而非原始数据	跨机构疾病预测模型训练、药物研发 [41]	数据不出本地，保护原始数据隐私，促进数据协作	梯度泄露风险、通信开销、非独立同分布数据处理、模型聚合安全性 [40]	39
同态加密 (Homomorphic Encryption, HE)	对加密数据直接进行计算，计算结果解密后与对明文数据计算结果一	安全的多方数据分析、外包计算 [2]	数据在计算过程中全程加密，提供强隐私保障	计算开销极大，性能瓶颈，目前主要适用于特定计算任务 [40]	2

	致				
差分隐私 (Differential Privacy, DP)	对查询结果或模型参数添加噪声，使个体数据贡献难以区分	保护训练数据集隐私、发布统计数据、保护联邦学习中的模型更新 [2]	提供可量化的严格隐私保证	隐私-效用权衡，噪声添加可能降低模型准确性，参数调整复杂 [44]	[2]
安全多方计算 (SMPC)	多方不泄露各自输入数据的前提下，共同完成某个函数的计算	联合数据分析、隐私保护的基因组数据分析 [2]	允许多方协作计算而无需信任中心方或暴露原始数据	通信复杂度和计算开销较高，可扩展性受限，对参与方在线状态有要求 [46]	[2]
匿名化/去标识化	移除或转换数据中的个人可识别信息（PII）	构建用于研究的去标识化医疗数据集 [2]	简单易行，是数据共享的基础步骤	再识别风险（尤其面对高级攻击和关联数据），可能损失部分数据维度和信息 [1]	[2]

B. 医学 AI 的安全性

1. 当前态势与研究现状：

   医学 AI 的安全性不仅涉及传统网络安全措施，更面临 AI 模型自身带来的新挑战。

   - **通用网络安全措施**：静态和传输中数据加密、严格的访问控制、全面的审计日志以及网络安全防护是保障医学 AI 系统安全的基础 [1]。
   - **对抗性攻击 (Adversarial Attacks)**：这是当前医学 AI 安全领域的研究热点，因为机器学习模型，特别是深度学习模型，对精心构造的微小扰动异常敏感。
     - **攻击类型**：
       - *逃逸攻击 (Evasion Attacks)*：在模型推理阶段，通过对输入数据（如医

学影像）添加人眼难以察觉的扰动，诱导模型产生错误分类[8]。常见算法包括快速梯度符号法（FGSM）、投影梯度下降法（PGD）、DeepFool、C&W 攻击等[10]。

- *投毒攻击 (Poisoning Attacks)*：在模型训练阶段，通过向训练数据中注入恶意样本，破坏模型的学习过程，植入后门或降低模型在特定输入上的性能[8]。
- *模型反演/窃取 (Model Inversion/Stealing)*：试图从模型中提取敏感的训练数据信息或复制模型本身[6]。
- *成员推断攻击 (Membership Inference)*：判断某个特定个体的数据是否被用于训练模型[51]。
- *提示注入 (Prompt Injection) (针对大语言模型 LLMs)*：通过构造特定的输入提示，操纵大语言模型的行为，使其产生有害或非预期的输出[51]。

- **潜在影响**：对抗性攻击可导致严重的临床后果，如误诊（例如，将良性病变误判为恶性肿瘤[53]，或干扰癌症检测[20]）、危及患者安全、并侵蚀对 AI 系统的信任[10]。研究表明，这类攻击的成功率可能非常高（可达 70-90%）[8]。
- **防御机制**：
  - *对抗性训练 (Adversarial Training)*：将对抗样本加入训练集，以增强模型的鲁棒性[8]。
  - *输入预处理/净化 (Input Preprocessing/Sanitization)*：采用随机裁剪、色彩抖动、添加噪声、特征压缩等技术，以消除或减弱对抗性扰动的影响[8]。
  - *防御性蒸馏 (Defensive Distillation)*：训练一个更鲁棒的小模型来近似一个复杂大模型的行为，从而平滑决策边界，降低对抗性攻击的成功率[8]。
  - *异常检测 (Anomaly Detection)*：识别输入数据中可能指示对抗性攻击的异常模式[8]。
  - *随机平滑 (Randomized Smoothing)*：通过对输入进行随机变换并聚合预测结果，来降低对抗性扰动的影响[8]。

2. **瓶颈与挑战**：

- **持续演变的威胁环境 (Evolving Threat Landscape)**：攻击者不断开发新的、更复杂的对抗性攻击技术。
- **通用防御的困难性 (Difficulty in Creating Universal Defenses)**：针对某种特定攻击有效的防御机制，可能对其他类型的攻击无效，领域内常说的“没有免费午餐”定理在此适用。
- **鲁棒性与性能的权衡 (Trade-off between Robustness and Performance)**：一些防御机制虽然能提升模型鲁棒性，但也可能导致模型在干净、非对抗性数据上的准确率下降，或增加计算成本 [10]。
- **模型的“黑箱”特性 (Black-Box Nature of Models)**：许多深度学习模型的不透明性，使得全面理解其脆弱点并设计针对性防御措施变得困难 [21]。
- **攻击的可转移性 (Transferability of Attacks)**：为一个模型精心构造的对抗样本，往往也能成功欺骗其他具有不同架构的模型，这构成了更广泛的安全威胁 [10]。
- **真实世界攻击的可行性 (Real-world Feasibility)**：将数字领域的对抗性攻击转化为对物理世界（如实际的医学影像设备或流程）的攻击，带来了新的挑战，是一个值得关注的新兴问题 [52]。
- **缺乏标准化的 AI 安全测试协议 (Lack of Standardized Security Testing Protocols for AI)**：与传统软件相比，AI 模型的安全测试尚不成熟。
- **人类监督的局限性 (Human Oversight Limitations)**：过度依赖 AI 安全解决方案而缺乏充分的人工监督，可能产生虚假的安全感，并使一些漏洞未被发现 [7]。

对抗性攻击之所以有效，很大程度上是因为它们利用了机器学习模型（尤其是深度学习模型）学习方式的基本特点，即基于梯度的优化和模式识别。这意味着当前深度学习范式中几乎内在地存在脆弱性。医学影像 AI 系统之所以成为对抗性攻击的高发区 [10]，是因为这类模型在识别那些人眼难以察觉的细微视觉特征方面表现出色，但也正是这些细微特征，如果模型未能稳健地学习，就容易受到对抗性扰动的操纵。因此，防御措施不能仅仅停留在表面，可能需要从根本上改变模型的学习或信息表示方式，甚至可能与提升模型的因果理解能力相关联——如果模型能学习到真正的因果特征，或许能减少对由对抗性噪声引起的虚假相关性的依赖。

表 2：医学 AI 中的对抗性攻击类型与防御策略总结

攻击类别	具体攻击类型 (示例)	攻击机制描述	对医学 AI 的潜在影响	常用防御策略	防御机制	防御局限性	相关文献
逃逸攻击 (Evasion)	FGSM, PGD, DeepFool, C&W	在模型推理时，对输入数据（如图像）添加微小、人眼难察的扰动，导致模型错误分类。	误诊（如良性判为恶性）、治疗方案错误、患者安全风险 [10]	对抗性训练、输入预处理（随机化、特征压缩）、防御性蒸馏、梯度隐藏/混淆	增强模型对扰动的鲁棒性、消除或减弱扰动、平滑决策边界 [8]	可能降低在干净数据上的性能、计算开销大、对未知攻击类型效果有限 [10]	[8]
投毒攻击 (Poisoning)	数据投毒、后门攻击	在模型训练阶段，向训练数据中注入恶意构造的样本，以破坏模型学习过程或植入特定行为。	模型性能下降、产生特定偏见、后门被激活导致特定错误输出 [8]	数据清洗与过滤、异常检测、鲁棒训练方法、模型正则化	识别并移除恶意训练数据、限制恶意数据对模型的影响 [8]	难以检测隐蔽的投毒数据、大规模数据集清洗成本高	[8]
模型窃取/反演	模型提取、模型反演、成	试图通过查询模型输出来复	知识产权泄露、患者隐私泄	模型加水印、差分隐私训	增加窃取难度、保护训练数	可能影响模型可用性、对某	[6]

	员推断	制模型功能、重建部分训练数据或判断特定数据是否在训练集中。	露 [6]	练、限制查询接口、输出混淆	据隐私 [51]	些攻击效果有限	
提示注入 (LLMs)	间接提示注入、越狱提示	针对大语言模型，通过精心设计的输入提示，绕过安全限制，诱导模型产生不当或有害输出。	生成错误医疗信息、泄露敏感数据、执行非授权操作 [51]	输入过滤与净化、指令微调、输出监测、强化学习（基于人类反馈）	识别和阻止恶意提示、限制模型输出范围 [51]	难以穷举所有恶意提示模式、可能误伤正常用户输入	[51]

C. 医学 AI 的临床验证

1. 当前方法与研究现状：

   临床验证是确保医学 AI 工具安全有效进入临床应用的关键环节。

   - **监管框架 (Regulatory Frameworks)**：
     - **美国 FDA**：通过上市前批准（PMA）、重新分类（De Novo）或 510(k)途径批准 AI/ML 赋能的医疗器械 [55]。已批准超过 1000 款此类设备，其中放射学领域占比最大（约 76%）[56]。为应对 AI/ML 医疗软件（SaMD）的持续学习特性，FDA 推出了“预定变更控制计划”（Predetermined Change Control Plan, PCCP）指南 [12]。
     - **欧洲 EMA 及欧盟 AI 法案**：EMA 负责欧洲药品和部分医疗器 CDE 的评估。欧盟《AI 法案》对 AI 应用采取基于风险的分类管理方法，医学 AI 设备通常

被划为高风险类别，需满足严格的合规性评估、风险管理和透明度要求[11]。

- **报告指南 (Reporting Guidelines)**：为提升 AI 临床试验报告的透明度和完整性，学术界和国际组织制定了一系列指南：
  - **CONSORT-AI**：用于报告 AI 干预措施的临床试验结果[14]。
  - **SPIRIT-AI**：用于规范 AI 干预措施临床试验方案的撰写[14]。
  - **STARD-AI**：针对 AI 诊断准确性研究报告的标准（STARD 的 AI 扩展版，仍在制定中）[15]。
  - **CLAIM**：医学影像 AI 应用报告清单[15]。
  - **TRIPOD+AI**：用于报告多变量预测模型的研究[60]。
- **性能评估指标 (Performance Metrics)**：传统上关注准确率、灵敏度、特异度、AUC 等统计指标[4]。日益认识到需要超越这些技术指标，评估 AI 的临床实用性、工作流程整合度以及对患者结局的实际影响[18]。针对大语言模型，也开始评估其临床推理能力，如 CRAFT-MD 框架[61]。
- **数据质量与管理 (Data Quality and Management)**：强调使用高质量、有代表性的数据集进行训练和验证，以避免偏倚[11]。数据预处理、去标识化和数据管理至关重要[28]。
- **上市后监测 (Post-Market Surveillance, PMS)**：AI 设备在真实世界临床应用中，需要持续监测其性能，以识别潜在风险，确保长期的安全性和有效性[16]。FDA 的 MAUDE 数据库用于追踪不良事件报告[17]。
- **数据漂移与模型退化管理 (Managing Data and Model Drift)**：认识到随着时间推移和真实世界数据分布的变化（概念漂移、数据漂移），AI 模型性能可能下降。应对策略包括持续监控、定期重训练和适应性学习[60]。

2. **瓶颈与挑战**：
   - **泛化性与鲁棒性 (Generalizability and Robustness)**：在一个医疗机构或特定人群中验证有效的模型，在新的应用场景（如不同的人群、设备、临床实践）中性能可能显著下降[13]。确保模型在不同人群亚组间的公平性也是一大挑战[1]。
   - **持续学习/自适应系统的验证 (Validation of Continuously Learning/Adaptive**

Systems)：传统的静态验证方法难以适用于能够随新数据不断学习和演进的 AI 系统。PCCP 是应对这一挑战的尝试，但更完善的方法论仍在发展中 [1]。

- **临床实用性的定义与衡量 (Defining and Measuring Clinical Utility)**：如何从技术准确性转向证明 AI 在真实临床中能带来实际益处（如改善患者结局、降低成本、提升工作流程效率）是一个复杂问题，需要精心设计的前瞻性研究 [13]。
- **数据可及性与质量 (Data Availability and Quality)**：获取大规模、多样化、高质量且标注良好的数据集用于 AI 训练和验证，仍然是一个重大障碍，尤其对于罕见病或代表性不足的人群 [54]。
- **缺乏标准化的评估指标 (Lack of Standardized Evaluation Metrics)**：除了常用的统计指标外，目前尚缺乏公认的用于衡量临床实用性、可解释性和安全性的标准化指标 [54]。
- **临床工作流程整合 (Integration into Clinical Workflows)**：将 AI 工具无缝整合到现有医院信息系统和临床工作流程中存在实际障碍，这可能影响其有效验证和最终采纳 [67]。
- **严格验证的成本与时间 (Cost and Time of Rigorous Validation)**：为 AI 进行大规模、前瞻性、随机对照临床试验，成本高昂且耗时漫长 [71]。
- **上市后监测不足 (Insufficient Post-Market Surveillance)**：现有的上市后监测体系可能不足以应对 AI/ML 设备的独特性，特别是在检测模型性能的细微退化或新出现的偏倚方面 [17]。
- **透明度与可重复性 (Transparency and Reproducibility)**：AI 研究报告的不完整性使得评估其有效性和重复研究结果变得困难 [14]。

医学 AI 的研发与临床成功应用之间存在一个显著的“死亡之谷”。许多模型在研究阶段显示出巨大潜力，但在真实世界验证中却因数据漂移、泛化能力不足或难以融入临床工作流程等问题而失败 [13]。当前的监管和验证框架仍在努力追赶 AI 技术（尤其是自适应 AI）的发展步伐。这意味着临床验证本身需要成为一个持续的、适应性的过程，整合真实世界证据、上市后监测以及管理数据漂移的方法，而不仅仅是一次性的上市前审批。“负责任的 AI”理念，涵盖整个生命周期的公平性、安全性和透明度，是克服这一挑战的关键 [13]。

表 3：医学 AI 临床验证指南与框架概览

指南/框架	发布机构/主要倡导者 (如适用)	主要目的/范围	AI 相关关键组成/建议	目标受众	相关文献
FDA PCCP (预定变更控制计划)	美国食品药品监督管理局 (FDA)	规范 AI/ML 医疗软件（SaMD）在上市后发生预期变更的管理，减少重复审批。	变更描述、变更方案（包括验证方法）、影响评估；允许在授权范围内修改模型而无需提交新的上市申请 [12]。	医疗器械制造商、AI 开发者	[12]
CONSORT-AI (AI 临床试验报告统一标准)	SPIRIT-AI 与 CONSORT-AI 工作组 (国际多方协作)	提高 AI 干预措施临床试验报告的透明度和完整性。	针对 AI 特性的扩展条目，如 AI 系统描述、输入数据、模型训练、人类 AI 交互、错误分析等 [14]。	研究人员、期刊编辑、审稿人、临床医生	[14]
SPIRIT-AI (AI 临床试验方案标准条目)	SPIRIT-AI 与 CONSORT-AI 工作组 (国际多方协作)	提高 AI 干预措施临床试验方案的透明度和完整性。	针对 AI 特性的扩展条目，涵盖试验设计、数据收集、AI 干预细节、评估计划等 [14]。	研究人员、伦理审查委员会、资助机构	[14]

STARD-AI (AI 诊断准确性研究报告标准)	(STARD 的 AI 扩展版，制定中)	规范 AI 诊断准确性研究的报告，提高可重复性和可比性。	预计将涵盖 AI 模型细节、训练与测试数据、性能评估方法、偏倚风险评估等，类似于 CLAIM [15]。	研究人员、诊断测试开发者、临床医生	[15]
CLAIM (医学影像 AI 清单)	Mongan J, et al. (Radiology: Artificial Intelligence)	为医学影像 AI 应用研究报告提供基本信息清单，确保研究结果的可解释性和可推广性。	研究注册、数据和代码可得性、偏倚来源报告、模型开发细节、性能评估指标等 [15]。	医学影像 AI 研究人员、期刊编辑	[15]
TRIPOD+AI (多变量预测模型报告指南 AI 扩展)	(TRIPOD 的 AI 扩展)	提高基于 AI 的多变量预测模型（用于个体预后或诊断）研究报告的质量和透明度。	针对模型开发、验证、性能评估和临床影响的 AI 特定报告要求 [60]。	预测模型研究人员、临床医生、政策制定者	[60]
欧盟 AI 法案 (EU AI Act) 及医疗器械法规 (MDR/IVDR)	欧盟委员会、欧洲议会	对 AI 系统（包括医疗器械中的 AI）进行基于风险的监管，确保安全、合法和符合伦理。	高风险 AI（多数医疗 AI 属此类）需进行合格评定、风险管理、数据质量控制、透明	AI 开发者、制造商、进口商、医疗机构、监管机构	[11]

			度、人类监督等[11]。		

D. 可解释 AI (XAI) 与因果推断的作用

1. **对解决核心挑战的贡献**：
   - **隐私保护**：XAI 可以通过揭示模型如何使用数据来增强透明度，从而辅助合规和建立信任，但解释本身也需妥善管理以防泄露敏感信息[21]。
   - **安全性**：理解模型的决策机制有助于识别其对对抗性攻击的脆弱点或可能被利用的偏倚[21]。
   - **临床验证**：XAI 帮助临床医生理解和信任 AI 的建议，审视其错误，并确保 AI 的推理过程与医学知识一致[13]。因果推断则致力于从"相关性"走向"因果性"，理解模型为何做出某种预测，这对于制定有效的干预措施和真正的临床决策支持至关重要[23]。
2. **当前主要方法**：
   - **XAI**：局部可解释模型无关解释（LIME）、SHAP 值、积分梯度、注意力机制等是常见的后设（post-hoc）解释方法，用于解释"黑箱"模型[78]。此外还有事前（ante-hoc）可解释模型，即模型本身设计为可解释的，如决策树、线性回归等[21]。
   - **因果推断**：结构因果模型（SCMs）、有向无环图（DAGs）、do-calculus（由 Judea Pearl 提出）是核心理论工具[71]。在实践中，随机对照试验（RCTs）被视为金标准，而模拟 RCTs、倾向性得分匹配、工具变量等方法则常用于处理观察性数据[71]。
3. **在医学背景下的局限性**：
   - **可解释性与保真度的权衡 (XAI)**：更简单的解释可能无法准确反映复杂模型的真实决策过程[49]。
   - **计算复杂度 (XAI)**：某些 XAI 方法（如 SHAP）对于大型模型或数据集而言计算成本高昂[54]。
   - **局部解释与全局解释 (XAI)**：LIME 等方法提供的是局部解释，其结论未必能推广到模型的全局行为[71]。

- **解释并非因果 (XAI)**：XAI 通常揭示的是特征重要性（相关性），而非真正的因果联系 [23]。
- **“可致因性”鸿沟 (Causability Gap)**：技术层面的可解释性（系统属性）与人类专家对因果关系的理解（个人属性）之间存在差距 [23]。目前的 XAI 方法可能无法完全实现临床医生所期望的“可致因性”。
- **因果推断的假设前提 (Causal Inference)**：基于观察性数据的因果推断依赖于一些强假设，而这些假设往往难以检验 [73]。混杂偏倚是主要挑战 [71]。
- **因果 AI 的数据需求 (Causal AI)**：通常不仅需要观察性数据，干预性数据或丰富的领域知识对于构建可靠的因果模型更为有利 [71]。
- **缺乏标准化的 XAI 评估方法 (XAI)**：难以客观衡量解释的“好坏”或其临床实用价值 [49]。

对 XAI 和因果推断的追求，并非仅仅源于学术好奇，而是由实际需求驱动的：临床医生需要信任并理解 AI 的建议 [21]，监管机构需要 AI 系统安全性和决策合理性的保证 [25]，而识别真正的因果因素对于制定有效治疗方案至关重要 [71]。然而，“可解释性”本身是一个多层面且时常带有主观性的概念 [21]。不同的利益相关者（如开发者、临床医生、患者）对解释的需求类型各不相同。当前的 XAI 方法（如 LIME、SHAP）更多是提供相关性信息（哪些特征影响了决策）[71]，而非深层次的因果理解 [75]。因果推断虽然致力于实现这种深层理解 [71]，但其自身也面临诸多挑战和局限性 [73]。这意味着单一的 XAI 解决方案不太可能满足所有需求。未来可能需要一种分层的、针对具体模型、利益相关者和临床情境定制的可解释性和因果推断方法。“可致因性”（causability）——即人类专家在多大程度上能够通过解释达成因果理解——是一个关键目标，而当前的技术性 XAI 方法可能尚未完全实现 [23]。

IV. 过往研究的里程碑

医学 AI 在隐私保护、安全性与临床验证方面的研究与实践，在过去数十年间取得了若干关键性进展。

A. 数据隐私法规的演进及其在 AI 领域的适用

- **早期数据保护理念**：在 AI 时代之前，医学数据保密的伦理原则和法律框架已初步建立。

- **HIPAA (1996, 美国)**：确立了保护敏感患者健康信息的国家标准，深刻影响了用于 AI 的数据收集和使用方式 [1]。
- **GDPR (2018, 欧盟)**：对数据处理、知情同意和数据主体权利施加了严格规定，因其域外效力而对全球 AI 研发和部署产生重大影响，特别关注自动化决策和用户画像等问题 [1]。欧洲数据保护委员会（EDPB）近期关于 AI 与 GDPR 的意见书进一步明确了 GDPR 的适用范围可能扩展至 AI 应用的训练和部署阶段，并对匿名化标准和合法利益作为处理依据提出了更严格的审视 [6]。这是对现有法律如何应用于 AI 的最新且重要的解读里程碑。
- **其他区域性法规 (如 CCPA)**：加州消费者隐私法案等进一步强化了区域性的数据隐私权 [25]。
- **AI 领域的适用性探索**：如何将这些普适性的数据保护法规具体应用于 AI 的独特性（如训练数据、模型输出、持续学习等）仍在持续的解释和指南制定过程中 [1]。

B. 医疗领域隐私增强技术（PETs）的关键发展

- **联邦学习（FL）的理念与早期实现**：谷歌于 2016 年提出联邦学习概念 [40]，随后被应用于医疗领域，以支持在不共享原始数据的情况下进行多机构协作研究 [41]。MELLODDY 联盟于 2020 年成功完成了首次大规模联邦学习运行，是其在药物研发领域应用的一个里程碑 [41]。
- **同态加密（HE）的进展**：尽管计算成本仍是主要障碍，但 HE 技术正朝着更实用的方向发展；选择性同态加密等方法（如 FAS）旨在平衡隐私保护强度与计算效率 [40]。
- **差分隐私（DP）的成熟**：从理论走向大规模系统中的实际应用（如苹果、谷歌等公司的实践 [44]），并被整合到联邦学习等框架中以增强隐私保护 [40]。
- **混合 PETs 解决方案的出现**：认识到单一 PETs 可能不足以应对复杂的隐私威胁，催生了结合多种技术优势的混合方法（如 SMHEA 算法 [39] 和 FAS 模型 [40]）。

C. FDA 对 AI 医疗器械的批准及监管思路的演变

- **首个 AI/ML 医疗器械获批 (1995 年)**：PAPNET 检测系统，一种半自动化的细胞学筛查系统，标志着 AI 技术首次获得 FDA 的医疗应用认可 [55]。

- **首个自主 AI 诊断平台获批 (IDx-DR, 2018 年)**：用于糖尿病视网膜病变早期检测的 IDx-DR 软件获得 De Novo 分类，是 AI 在诊断领域的一个里程碑事件 [55]。
- **AI 医疗器械批准数量激增**：近年来，FDA 批准的 AI/ML 赋能医疗器械数量显著增加，尤其是在放射学领域（2023 年批准 221 个，2024 年上半年批准 107 个）[56]。放射学相关的 AI 设备约占所有 AI 医疗器械批准总数的 76% [57]。
- **SaMD 监管框架草案 (2019 年)**：FDA 首次针对 AI/ML 医疗软件的独特性提出的监管框架草案，旨在解决其持续学习和迭代更新带来的挑战 [55]。
- **AI/ML SaMD 行动计划 (2021 年)**：在 2019 年草案的反馈基础上，FDA 发布了更为具体的行动计划，阐述了其监管 AI/ML 医疗软件的五大支柱 [55]。
- **PCCP 指南的发布 (草案 2023 年，最终版约 2024/2025 年)**："预定变更控制计划"指南允许制造商在初始审批时就预先规划并获得批准其 AI/ML 模型的后续修改，而无需为每次变更重新提交申请，这对于适应 AI 的动态特性具有重要意义 [12]。
- **FDA 突破性设备认定**：Paige Prostate Detect、Paige Breast Lymph Node 和 Paige PanCancer Detect 等创新 AI 工具获得了 FDA 的突破性设备认定，表明监管机构对其潜力的认可 [82]。

D. AI 临床试验报告指南的确立

- **需求认知**：早期 AI 研究报告质量参差不齐，信息不完整，阻碍了对其质量、偏倚和泛化能力的评估 [14]。
- **SPIRIT-AI 与 CONSORT-AI 倡议 (2019 年宣布，2020 年发布)**：由国际多方利益相关者共同制定的、基于共识的指南，分别用于提高 AI 干预临床试验方案（SPIRIT-AI）和试验报告（CONSORT-AI）的透明度和完整性。这些指南得到了 EQUATOR 网络的支持，并被主要医学期刊推荐使用 [14]。
- **STARD-AI、CLAIM 等指南的制定**：针对 AI 诊断准确性研究（STARD-AI）和医学影像 AI 研究（CLAIM）等特定领域，也相继出台或正在制定相应的报告标准和清单 [15]。

E. 对抗性攻防研究的重大突破

- **对抗性脆弱性的发现**：早期研究（如 Goodfellow 等人关于图像分类的研究 [8]）揭示了深度神经网络对微小、人眼难以察觉的扰动非常敏感。
- **攻击算法的开发**：研究人员设计出多种白盒攻击（如 FGSM、PGD）和黑盒攻击方法，系统地探索了模型的脆弱点 [8]。
- **医学领域的验证**：后续研究证实，医学影像 AI 系统同样易受此类攻击影响，可能导致误诊等严重后果 [10]。
- **防御策略的兴起**：针对对抗性攻击，研究界提出了对抗性训练、防御性蒸馏、输入净化等多种防御思想和技术 [8]。
- **真实世界攻击演示**：例如，通过物理手段欺骗特斯拉自动驾驶系统的案例 [51]，展示了对抗性攻击从数字领域向物理世界延伸的可能性。

F. XAI/因果 AI 在构建可信医学 AI 中的兴起与认可

- **早期基于规则系统的 XAI**：像 MYCIN 这样的早期专家系统，其决策过程可以通过追溯规则链条得到解释，具有内在的可解释性 [23]。
- **深度学习的"黑箱"问题**：随着高精度但高度不透明的深度学习模型的兴起，对可解释 AI（XAI）的需求日益迫切 [21]。
- **后设 XAI 方法的开发**：LIME、SHAP 等旨在解释"黑箱"模型行为的方法在 2010 年代中后期被相继提出 [22]。
- **Judea Pearl 的因果推断革命**：Judea Pearl 在结构因果模型、do-calculus 和因果关系之梯方面的工作（主要在 20 世纪 90 年代至 21 世纪初）为 AI 领域的因果推断奠定了理论基础 [79]。其著作《为什么：因果关系的新科学》（2018 年）进一步推广了这些思想 [79]。
- **因果推断在医学中的应用**：学术界和产业界日益认识到，在医学领域，理解因果关系对于有效的干预至关重要，这超越了单纯的预测 [23]。
- **XAI 在临床决策支持系统（CDSS）中的整合**：努力使 CDSS 的建议更加透明和值得信赖，提升其临床应用价值 [22]。
- **具有可解释特征的 AI 产品获批**：一些商业化的 AI 医疗产品开始融入可解释性元

素，以帮助临床医生理解其输出（尽管全面的 XAI 整合仍在发展中）。例如，Paige Prostate Detect 能够高亮显示可疑区域 [82]。

- **医疗 XAI 研究里程碑 (2021 年)**：FDA 批准用于诊断的 AI 工具、实时预测工具的出现以及 AI 与电子健康记录（EHR）系统的集成，标志着 XAI 在医疗领域的应用进入新阶段 [26]。

许多“里程碑”事件实际上是对已有问题或新兴挑战的响应，而非完全主动的规划。例如，GDPR 等隐私法规的制定具有普适性，其在 AI 领域的具体适用是一个持续解释和调整的过程 [6]。同样，对抗性攻击的研究往往是在发现新的模型脆弱性之后才得以深入。这表明在快速发展的 AI 领域，监管和伦理框架的建立常呈现出一种“追赶”技术发展的态势。

另一个值得注意的趋势是先前相对独立的学科领域正在加速融合。例如，隐私增强技术（源于密码学、统计学）正成为 AI（机器学习）不可或缺的一部分。医疗器械的监管科学正在与软件监管和数据治理原则相结合。可解释 AI 则借鉴了认知科学、哲学和计算机科学的成果。这种跨学科的融合是推动医学 AI 领域进步的关键动力，许多里程碑事件（如隐私保护机器学习的发展 [45] 或 AI 临床试验指南的制定 [14]）都发生在这些学科的交叉点。

此外，回顾这些里程碑的时间线可以发现，近年来的发展速度显著加快，特别是在 2015 年之后，这与深度学习技术的突破和计算能力的指数级增长密切相关。这种快速的迭代步伐本身也对建立稳定、持久的标准和法规构成了挑战。即使是几年前制定的框架和指南，也可能需要迅速更新以保持其相关性，FDA 对其 SaMD 指南的持续修订即为例证 [12]。这为开发者和监管者共同营造了一个充满活力但也存在不确定性的环境。

V. 产业化案例

医学 AI 在隐私保护、安全性与临床验证方面的努力，已在众多商业化产品和平台中得到体现。这些案例不仅展示了技术的进步，也反映了行业为应对核心挑战所做的努力。

A. 产业界的隐私保护平台与技术应用

- **Owkin**：
  - **平台与技术**：Owkin Connect（前身为 Owkin Studio）是一款联邦学习软件，支持医院、研究中心和制药公司在不共享原始数据的情况下进行机器学习合作[41]。该平台确保数据所有权和合规性（如 GDPR/HIPAA）保留在各合作方，并通过安全聚合和差分隐私模型训练等技术防止数据泄露[41]。
  - **应用案例**：与法国医院合作，利用多模态数据（CT 影像、报告、临床数据）开发了 COVID-19 重症预测模型[41]。此外，通过 MELLODDY 联盟，与 10 家制药公司利用联邦学习共同开发基于庞大化合物库的药物发现模型[41]。
  - **解决的问题**：有效解决了多机构合作中敏感数据共享的难题，促进了大规模协作研究。
- **Inka Health (PROmAI 联盟)**：
  - **倡议与平台**：发起全球 AI 肿瘤学联盟（PROmAI），旨在整合真实世界数据和临床试验数据（分子、影像、临床等多模态数据）进行预测性癌症研究，以打破数据孤岛[88]。其 SynoGraph 平台利用 AI 驱动的因果推断技术，整合多模态数据，识别可能对特定治疗产生应答的患者群体[88]。
  - **解决的问题**：聚焦于提升肿瘤学 AI 的预测准确性、可解释性、可移植性和监管相关性，致力于构建透明、可信的 AI 应用[88]。
- **谷歌 (MedGemma, AMIE)**：
  - **产品与技术**：推出了用于医疗保健 AI 的开放模型 MedGemma[89]，以及能够通过解读视觉医疗信息来辅助诊断对话的 AMIE 智能体[89]。这些举措表明大型科技公司正积极布局医疗 AI 基础模型和工具，其设计可能内含隐私保护考量，尽管具体采用的 PETs 细节在当前资料中尚不明确。

B. 医学 AI 领域的安全解决方案与实践

尽管针对医学 AI 的特定商业化"AI 安全产品"在所提供的资料中没有广泛详述，但 AI 开发公司在实践中已普遍关注并应用相关安全原则。

- **通用安全实践**：像 Paige 这样的公司强调稳健的数据治理、安全措施和合规性（如

通过 ISO 27001 认证，符合 HIPAA、GDPR 要求）[90]。安全的 API 协议、速率限制、隐私保护日志记录等是系统架构层面的安全考量 [2]。

- **对抗性鲁棒性**：对于开发关键性 AI 诊断工具的公司而言，模型的鲁棒性（即抵抗扰动的能力）是其研发和质量控制的重要方面，这与对抗性攻击防御的研究紧密相关 [10]。工业界也逐渐认识到，需要超越传统的准确率指标来评估模型，将安全性纳入考量 [8]。

C. 产业界的临床验证成功案例与挑战

众多公司已成功将其 AI 产品推向市场，并通过了严格的临床验证和监管审批。

- **Paige**：
  - **产品系列**：提供包括 Paige Prostate Suite（用于前列腺癌检测、分级和量化）、Paige Breast Suite、Paige GI Suite 以及 Paige PanCancer Suite 在内的一系列 AI 辅助诊断应用 [82]。其 FullFocus®全视野数字切片阅片器也获得了 FDA 批准 [82]。
  - **监管里程碑**：Paige Prostate Detect 是首款获得 FDA 批准的病理学 AI 应用 [82]。多款产品获得 FDA 突破性设备认定（前列腺癌、乳腺癌淋巴结、泛癌种检测）[82]，并拥有 13 项 CE-IVDD 和 UKCA 认证 [82]。
  - **临床验证数据**：Paige Prostate Detect 的研究数据显示，其能显著提升工作效率（阅片时间减少高达 21.9%），缩短诊断周转时间（减少 65.5%），降低诊断错误率（减少 70%），并可能减少免疫组化（IHC）的使用及其相关成本 [38]。耶鲁大学的一项独立研究表明，在作为预筛查工具时，其特异性达 99.3%，灵敏度 97.7%，阴性预测值（NPV）99.2%，可使 68.6%的活检样本免于人工复核 [92]。Paige Breast Lymph Node 在检测转移灶方面灵敏度高达 98%，阅片时间缩短多达 55% [82]。Paige PanCancer Detect 在内部验证中对多种组织类型的样本级 AUC 达到 0.95 [82]。
  - **核心技术与方法**：Paige 的核心竞争力在于其海量高质量病理数据集（超过 2500 万张数字切片）、先进的基础模型（如 Virchow、PRISM）以及多模态能力（结合大视觉模型 LVMs 和大语言模型 LLMs）[90]。公司强调临床级 AI 的开

发、验证和卓越的监管实践[87]。

- **Viz.ai**：
  - **平台与产品**：Viz.ai One™是一个 AI 驱动的智能医疗协调平台，拥有超过 50 种经 FDA 批准的 AI 算法，用于分析医学影像（CT、心电图、超声心动图等）[93]。其应用覆盖卒中（Viz Neuro Suite，包括大血管闭塞 LVO、CT 灌注 CTP、出血等）、肥厚型心肌病（Viz HCM）、动脉瘤、肺栓塞（PE）等多个领域[93]。
  - **监管里程碑**：Viz HCM 是首个获得 FDA 批准用于肥厚型心肌病筛查的 AI 算法（2023 年 8 月通过 De Novo 途径获批）[94]。其众多模块均获得了 FDA 的许可。
  - **临床验证数据**：Viz HCM 在通过心电图检测肥厚型心肌病方面展现出高准确性，并有潜力使患者提前数年得到诊断[94]。一项研究显示，其针对心脏 MRI 确诊病例的灵敏度为 56%，特异性 100%，阳性预测值（PPV）100%[94]。Viz ICH Plus（用于颅内出血容积和中线移位等的自动测量）比传统的 mABC/2 方法更准确，且速度比手动半自动分割（SAS）快近 3 倍[95]。一项关于 Viz LVO 的 Meta 分析表明，其应用与卒中救治流程中关键时间节点的缩短相关（如 CT 至血管内治疗时间、门到腹股沟穿刺时间等），尽管在改善临床结局方面的统计学显著性尚待进一步证实[98]。
  - **核心技术与方法**：Viz.ai 的核心价值在于提供实时洞察、自动化评估，从而简化工作流程，加强医疗团队协作[93]。
- **Aidoc**：
  - **平台与产品**：aiOS™是一个统一的医疗 AI 平台[48]，其 CARE1™基础模型是其技术核心之一[70]。产品覆盖放射科、心脏科、神经血管科、血管外科等，据称可覆盖医疗系统中 75%的患者[48]。具体应用包括肋骨骨折分诊、肺栓塞检测、颅内出血检测等。
  - **监管里程碑**：其基于 CARE1™基础模型的肋骨骨折分诊 AI 解决方案获得了 FDA 的 CADt（计算机辅助分诊和通知）许可，是同类首个基于基础模型的 QFM（合格的医疗器械功能）SaMD 设备[70]。Aidoc 拥有市场领先数量的 FDA 许可[48]。
  - **临床验证数据**：在一项针对意外肺栓塞（IPE）的大样本研究中，Aidoc 的 PE 检

测算法显示出高灵敏度（96.4%）、高特异性（高达 99.8%）、高 PPV（87.1%）和高 NPV（高达 99.9%），总体准确率高达 99.7%，并成功标记了 11 例初次报告遗漏的 IPE 病例[100]。然而，另一项研究指出，当前 Aidoc 模型在区分急性颅内出血与脑血管造影后 CT 上的残留对比剂方面尚有不足（灵敏度 51.7%，特异性 50.0%），提示针对此特定场景的模型仍需优化[69]。公司宣称其 AI 工具可使 PE 患者的通知时间加快 31%，卒中患者的门到穿刺时间缩短 34%[48]。

  - **核心技术与方法**：利用基础模型实现 AI 解决方案的可扩展性和快速开发，强调与临床实践的无缝集成和企业级规模化部署[48]。
- **Tempus**：
  - **平台与数据**：拥有庞大的多模态数据库（超过 4000 万份研究记录，包含基因组、临床、影像等数据）[101]。Fuses 项目利用其新型基础模型进行诊断、预后和预测建模[102]。Loop 平台则专注于肿瘤学靶点发现和验证，整合了真实世界数据（RWD）、人源生物模型和 CRISPR 筛选技术[101]。
  - **应用领域**：AI 赋能的诊断（如免疫治疗相关的免疫特征评分 IPS）、治疗研究、药物发现、临床试验优化等[101]。
  - **解决的问题**：旨在通过从海量患者数据中学习普适性规律，使诊断更智能、更个性化。利用 RWD 和先进建模技术加速研究进程[101]。
- **SOPHiA GENETICS**：
  - **平台与技术**：SOPHiA DDM™多模态平台致力于整合和标准化肿瘤学领域的基因组、影像、临床和生物学数据[103]。SOPHiA CarePath®用于数据可视化、队列构建和分析，SOPHiA DDM™多模态工厂则用于机器学习模型的开发[103]。
  - **应用案例**：癌症研究，识别疾病复发、治疗反应和毒副反应的趋势。已开发多个预测模型，如用于非小细胞肺癌（NSCLC）的 DEEP-Lung-IV 研究和用于肾癌复发预测的 UroPredict 计算器[103]。
  - **解决的问题**：打破数据孤岛，加速数据驱动的医学发展，提高准确性，缩短周转时间，支持精准医疗[103]。其 SOPHiA UNITY 网络促进了多中心协作研究[103]。
- **Causaly**：

- **平台与技术**：一个 AI 驱动的生命科学研发平台，核心是一个能够区分因果关系与相关性的大规模、高精度知识图谱。结合了科学文献检索增强生成（Scientific RAG™）、生成式 AI 操作系统和企业数据结构 [104]。
- **应用领域**：主要用于药物靶点识别与优先级排序、生物标志物发现以及疾病病理生理学研究 [104]。
- **解决的问题**：通过高效查找、解读和共享生物医学信息，特别是强调因果联系的挖掘，旨在降低研发风险，提高研发效率 [104]。

- **PathAI**：
  - **平台与技术**：AISight®数字病理图像管理系统，以及一系列 AI 驱动的病理学算法 [105]。
  - **应用领域**：肿瘤检测、胶原蛋白自动量化、生物标志物发现、空间生物学、基于组织的临床研究。已应用于溃疡性结肠炎、克罗恩病、三阴性乳腺癌（TNBC）复发预测等研究中 [105]。
  - **产业合作**：与 Precision for Medicine 合作，部署 AISight®和 AI 算法，以增强生物样本分析、临床试验服务和多模态数据集的价值。目标是将复杂的生物标志物组合简化为可操作的指标，并解读复杂的组织生物学信息 [106]。
  - **解决的问题**：提高生物样本分析的一致性和数据可靠性；为已测序样本增加细胞层面的洞察，超越传统病理评估；优化病理工作流程和准确性 [105]。

从这些产业化案例中可以看出，成功的医学 AI 应用往往聚焦于特定的、具有高临床价值的需求点，在这些点上 AI 能够展现出超越现有方法的明确优势（例如 Viz.ai 在卒中分诊上的快速反应，Paige 在前列腺癌检测上的精度提升）。目前，通用的“全能型”医学 AI 尚未成为现实。

多模态数据整合是行业内一个强劲的趋势（如 Tempus、SOPHiA GENETICS、PathAI、Inka Health、Owkin 等公司的实践），这反映了业界共识，即融合不同类型的数据（基因组学、影像学、临床记录、真实世界数据等）能够产生更强大的洞察和更优越的模型。然而，这也无疑加剧了在隐私保护、系统安全和临床验证方面的挑战。

合作与联盟正成为重要的行业策略，尤其是在获取多样化数据（如 PROmAI 联盟 [88]）和进行前沿共性技术研究（如联邦学习领域的 MELLODDY 联盟 [41]）方面。这表明，没有任何单一机构能够拥有所有必要的数据或专业知识。

“基础模型”的概念也开始进入医学 AI 领域（例如 Aidoc 的 CARE1 [70]，Paige 的 Virchow/PRISM [90]，Tempus 的 Fuses 模型 [102]），它们有望加速新应用的开发并拓宽适用范围。然而，如何对这些大型、通用的基础模型进行有效的临床验证，并确保其不带有潜在的偏见，是亟待解决的新挑战。

VI. 总结与展望

A. 核心要素的相互作用回顾

医学人工智能的发展与应用，其核心在于如何在创新与责任之间取得平衡。隐私保护、系统安全性以及临床验证，这三大支柱并非相互独立，而是紧密交织、缺一不可。任何一个环节的疏漏都可能削弱其他环节的努力，并最终损害医学 AI 在医疗保健领域发挥其巨大潜力的基础。一个无法保障患者隐私的 AI 系统，即使技术上再先进，也难以获得公众信任；一个容易受到攻击而不安全的 AI 系统，其临床决策的可靠性将大打折扣；而一个未经充分临床验证的 AI 工具，无论其隐私和安全措施多么完善，都可能给患者带来无法预估的风险。因此，只有协同推进这三方面的工作，才能构建出真正值得信赖且行之有效的医学 AI。

B. 未来研究与发展方向

1. **隐私保护**：
    - 开发计算效率更高、鲁棒性更强、且能在隐私保护与数据效用之间取得更优平衡的隐私增强技术（PETs）。
    - 建立标准化的再识别风险评估方法学，尤其针对多模态和高维度医疗数据。
    - 研究和推广动态的、细粒度的用户知情同意管理系统，以适应复杂的数据共享和使用场景。

- 探索将“隐私融入设计（Privacy by Design）”原则更深度地整合到 AI 模型开发的全生命周期。

2. **安全性**：
   - 研发能够主动防御、甚至预测对抗性攻击的新型防御机制，并将其整合到模型设计和训练过程中。
   - 开发针对 AI 系统（特别是医学 AI）的自动化安全审计工具和标准化测试协议。
   - 加强对 AI 供应链安全的关注，确保第三方组件和数据的安全性。
   - 研究针对新兴 AI 范式（如生成式 AI、基础模型）的特有安全威胁和防护策略。
3. **临床验证**：
   - 为持续学习和自适应 AI 系统建立有效的、动态的临床验证方法论和监管路径。
   - 开发和推广能够全面评估 AI 临床实用性、人机交互效率以及对患者长期结局影响的新型评估指标和框架。
   - 构建高效的真实世界证据（RWE）生成和利用体系，以及针对 AI 的持续性上市后监测（PMS）机制。
   - 深入研究并制定有效策略，以检测、量化和减轻 AI 模型中的偏倚，确保其在不同人群和医疗环境中的公平性和普适性。
4. **可解释 AI（XAI）与因果推断**：
   - 推动 XAI 从技术可解释性向真正的“可致因性”（causability）发展，使临床医生能够理解 AI 决策背后的因果逻辑，而不仅仅是相关性。
   - 研究如何将因果发现与预测模型更紧密地结合，提升模型的鲁棒性和决策的可靠性。
   - 开发针对多模态输入和复杂 AI 架构（如基础模型）的有效 XAI 方法。
   - 建立 XAI 解释质量和临床实用性的评估标准。
5. **多模态 AI**：
   - 研究更强大的多模态数据融合技术，以有效整合异构数据源，同时解决由此带来的独特的隐私、安全和验证挑战。
   - 探索多模态 AI 在复杂疾病诊断、个性化治疗方案制定以及药物研发中的深层应用。

C. 对各利益相关方的建议

- **研究人员**：
  - 加强跨学科合作（如医学、计算机科学、伦理学、法学、社会学）。
  - 积极开发并遵循稳健的研究方法和透明的报告标准（如 CONSORT-AI、SPIRIT-AI）。
  - 优先攻关当前医学 AI 发展中的关键瓶颈问题，如自适应 AI 的验证、鲁棒 XAI 的开发等。
- **开发者与企业**：
  - 将“隐私融入设计”和“安全融入设计”作为产品开发的基本原则。
  - 投入资源进行超越监管最低要求的严格临床验证，确保产品的安全性和有效性。
  - 与临床医生紧密合作，确保 AI 工具的临床实用性和易用性。
  - 对模型的能力和局限性保持透明，避免过度宣传。
- **监管机构**：
  - 持续调整和完善监管框架，以适应 AI 技术的独特性（如自适应性、数据驱动性）。
  - 推动数据质量、偏倚缓解、上市后监测等方面的国际标准协调与互认。
  - 为行业提供关于 AI 医疗器械（特别是采用新技术如基础模型的产品）的清晰、可操作的指南。
- **医疗服务提供者/机构**：
  - 建立院内 AI 治理框架和伦理审查机制。
  - 投资于必要的基础设施建设和人员培训，为 AI 的安全有效部署做好准备。
  - 在采纳 AI 工具前进行审慎评估，并积极参与真实世界证据的生成与反馈。
- **患者与公众**：
  - 积极了解 AI 在医疗中的应用，维护自身的数据权利。
  - 参与关于医学 AI 伦理和社会影响的公共讨论，促进负责任的创新。

医学 AI 的未来发展，很可能从目前针对特定任务的“点状”应用，转向深度融入整个医疗服务流程的“线状”乃至“网状”整合。这意味着对隐私保护、安全性和临床验证的考量，也

必须从单一模型或工具的评估，扩展到对整个护理路径和系统的整体评估。例如，隐私保护需要覆盖跨系统的数据流转，安全性需要保护相互连接的各个组件，而临床验证则需要评估 AI 在复杂真实世界流程中的综合影响。

随着 AI 在医疗领域日益普及和强大，伦理考量，特别是公平性和普惠性，将占据更加核心的位置。解决数据和算法中的偏倚问题，不仅是技术验证的要求，更是贯穿隐私（如何处理弱势群体数据）、安全（如何防范可能对特定群体造成更大伤害的攻击）和临床验证（如何确保在不同人群中的公平表现）的核心伦理责任。这要求在 AI 的整个生命周期——从数据收集、模型开发到验证部署的每一个环节——都主动嵌入对公平和普惠的考量，可能需要新的审计框架和持续的社会对话来共同塑造一个更加公正、可信的医学 AI 未来。

VII. 参考文献

107

引用的著作

1. Ethical and legal considerations in healthcare AI: innovation and ..., 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12076083/
2. Secure AI-Powered Early Detection System for Medical Data Analysis & Diagnosis - InfoQ, 访问时间为 六月 11, 2025, https://www.infoq.com/articles/secure-ai-powered-early-detection-system/
3. AMA Principles for Augmented Intelligence Development, Deployment, and Use - American Medical Association, 访问时间为 六月 11, 2025, https://www.ama-assn.org/system/files/ama-ai-principles.pdf
4. Is Multimodal Better? A Systematic Review of Multimodal versus ..., 访问时间为 六月 11, 2025, https://www.medrxiv.org/content/10.1101/2025.03.12.25322656v1.full
5. Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine - CDC, 访问时间为 六月 11, 2025, https://www.cdc.gov/pcd/issues/2024/24_0245.htm
6. Europe Tightens Data Protection Rules for AI Models—And It's a Big Deal for Healthcare and Life Sciences - Petrie-Flom Center, 访问时间为 六月 11, 2025, https://petrieflom.law.harvard.edu/2025/02/24/europe-tightens-data-

protection-rules-for-ai-models-and-its-a-big-deal-for-healthcare-and-life-sciences/
7. AI in healthcare privacy: Enhancing security or introducing new risks? - Paubox, 访问时间为 六月 11, 2025, https://www.paubox.com/blog/ai-in-healthcare-privacy-enhancing-security-or-introducing-new-risks
8. Adversarial Attacks and Defense Mechanisms in Generative AI - [x]cube LABS, 访问时间为 六月 11, 2025, https://www.xcubelabs.com/blog/adversarial-attacks-and-defense-mechanisms-in-generative-ai/
9. What Is Adversarial AI in Machine Learning? - Palo Alto Networks, 访问时间为 六月 11, 2025, https://www.paloaltonetworks.com/cyberpedia/what-are-adversarial-attacks-on-AI-Machine-Learning
10. Adversarial Attacks on Medical Imaging AI: Risks and Mitigation Strategies - ResearchGate, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/390113972_Adversarial_Attacks_on_Medical_Imaging_AI_Risks_and_Mitigation_Strategies
11. Validation of Medical Devices Incorporating Artificial Intelligence - SQS, 访问时间为 六月 11, 2025, https://www.sqs.es/validation-of-medical-devices-incorporating-artificial-intelligence/?lang=en
12. The FDA Released Final Guidelines for AI-Enabled Medical Device ..., 访问时间为 六月 11, 2025, https://www.fmai-hub.com/the-fda-released-final-guidelines-for-ai-enabled-medical-device-development/
13. Finding the Best AI Algorithms: Is FDA Approval Enough? - Mayo Clinic Platform, 访问时间为 六月 11, 2025, https://www.mayoclinicplatform.org/2025/04/17/finding-the-best-ai-algorithms-is-fda-approval-enough/
14. The SPIRIT-AI and CONSORT-AI initiative - The Alan Turing Institute, 访问时间为 六月 11, 2025, https://www.turing.ac.uk/research/research-projects/spirit-ai-and-consort-ai-initiative
15. AI Reporting Guidelines: How to Select the Best One for Your Research - RSNA Journals, 访问时间为 六月 11, 2025, https://pubs.rsna.org/doi/pdf/10.1148/ryai.230055
16. Keeping Devices Safe After Launch with Post-Market Surveillance: A New Age Perspective, 访问时间为 六月 11, 2025, https://www.freyrsolutions.com/blog/keeping-devices-safe-after-launch-with-post-market-surveillance-a-new-age-perspective
17. A general framework for governing marketed AI/ML medical devices - PubMed, 访问时间为 六月 11, 2025, https://pubmed.ncbi.nlm.nih.gov/40450160/

18. The Good, the Bad, and the Ugly of AI in Medical Imaging - EMJ, 访问时间为 六月 11, 2025, https://www.emjreviews.com/radiology/article/the-good-the-bad-and-the-ugly-of-ai-in-medical-imaging-j140125/
19. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata - Diagnostic and Interventional Radiology, 访问时间为 六月 11, 2025, https://dirjournal.org/pdf/beb8919b-f013-4ea1-b1c8-40332e840fe1/articles/dir.2024.242631/DIR-2024-242631-AYDA.pdf
20. The Impact of Adversarial Attacks on Medical Imaging AI Systems - ResearchGate, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/381356977_The_Impact_of_Adversarial_Attacks_on_Medical_Imaging_AI_Systems
21. What Is the Role of Explainability in Medical Artificial Intelligence? A ..., 访问时间为 六月 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12025101/
22. Explainable AI in Clinical Decision Support Systems for Precision Medicine - ResearchGate, 访问时间为 六月 10, 2025, https://www.researchgate.net/publication/391662889_Explainable_AI_in_Clinical_Decision_Support_Systems_for_Precision_Medicine
23. Causability and explainability of artificial intelligence in medicine ..., 访问时间为 六月 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7017860/
24. Explainable Artificial Intelligence in Medical Decision Support Systems - IET Digital Library, 访问时间为 六月 10, 2025, https://digital-library.theiet.org/doi/book/10.1049/pbhe050e
25. Regulations of AI in healthcare knowledge management | Market ..., 访问时间为 六月 11, 2025, https://marketlogicsoftware.com/blog/ai-regulations-in-healthcare-knowledge-management/
26. AI has come a long way in healthcare, still has a long way to go: Research recap, 访问时间为 六月 10, 2025, https://aiin.healthcare/topics/artificial-intelligence/ai-has-come-long-way-healthcare-still-has-long-way-go-research-recap
27. AI In Healthcare Highlights & Milestones 2021 - AI Time Journal - Artificial Intelligence, Automation, Work and Business, 访问时间为 六月 10, 2025, https://www.aitimejournal.com/ai-in-healthcare-highlights-milestones-in-2021/33128/
28. How AI Achieves 94% Accuracy in Early Disease Detection: New Research Findings, 访问时间为 六月 11, 2025, https://globalrph.com/2025/04/how-ai-achieves-94-accuracy-in-early-disease-detection-new-research-findings/
29. An Economic Overview of the Value of AI in Radiology - Aidoc, 访问时间为 六月 11, 2025, https://www.aidoc.com/learn/blog/value-artificial-intelligence-radiology/

30. (PDF) Comprehensive Review of Multimodal Medical Data Analysis ..., 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/366601357_Comprehensive_Review_of_Multimodal_Medical_Data_Analysis_Open_Issues_and_Future_Research_Directions
31. Advancements in Medical Radiology Through Multimodal Machine Learning: A Comprehensive Overview - PMC, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12108733/
32. Multimodal AI: Revolutionising the Healthcare Industry, 访问时间为 六月 11, 2025, https://www.octopusintelligence.com/what-are-the-most-promising-applications-of-multimodal-ai-in-healthcare/
33. 基于层次注意力的多模态数据融合算法在疾病诊断与癌症亚型预测中 ..., 访问时间为 六月 11, 2025, https://www.ebiotrade.com/newsf/2025-5/20250522115925959.htm
34. 多模态大数据智能应用引领精准医疗时代到来，神州医疗南方医院项目二期启动, 访问时间为 六月 11, 2025, http://www.dhctech.com/gongsixinwen/397.html
35. A multi-modal deep learning solution for precise pneumonia diagnosis: the PneumoFusion-Net model - PMC, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11937601/
36. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use, 访问时间为 六月 11, 2025, https://royalsocietypublishing.org/doi/10.1098/rsos.241873
37. The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union - Frontiers, 访问时间为 六月 11, 2025, https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.927721/full
38. Prospective Clinical Implementation of Paige Prostate Detect ..., 访问时间为 六月 11, 2025, https://pubmed.ncbi.nlm.nih.gov/40036728/
39. Federated Learning with Homomorphic Encryption for Ensuring Privacy in Medical Data, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/382340489_Federated_Learning_with_Homomorphic_Encryption_for_Ensuring_Privacy_in_Medical_Data
40. A Selective Homomorphic Encryption Approach for Faster Privacy-Preserving Federated Learning - arXiv, 访问时间为 六月 11, 2025, https://arxiv.org/html/2501.12911v4
41. Federated learning in healthcare: the future of collaborative clinical ..., 访问时间为 六月 11, 2025, https://www.owkin.com/blogs-case-studies/federated-learning-

in-healthcare-the-future-of-collaborative-clinical-and-biomedical-research
42. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture - PubMed Central, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10897620/
43. Privacy-Preserving Federated Learning Framework for Multi-Source Electronic Health Records Prognosis Prediction - MDPI, 访问时间为 六月 11, 2025, https://www.mdpi.com/1424-8220/25/8/2374
44. Differential Privacy and Federated Learning for Medical Data | Towards Data Science, 访问时间为 六月 11, 2025, https://towardsdatascience.com/differential-privacy-and-federated-learning-for-medical-data-0f2437d6ece9/
45. Privacy-preserving machine learning: a review of federated learning techniques and applications - ResearchGate, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/388822437_Privacy-preserving_machine_learning_a_review_of_federated_learning_techniques_and_applications
46. (PDF) Comprehensive Review on Privacy-Preserving Machine Learning Techniques for Exploring Federated Learning - ResearchGate, 访问时间为 六月 11, 2025, https://www.researchgate.net/publication/383847661_Comprehensive_Review_on_Privacy-Preserving_Machine_Learning_Techniques_for_Exploring_Federated_Learning
47. [2501.12911] A Selective Homomorphic Encryption Approach for Faster Privacy-Preserving Federated Learning - arXiv, 访问时间为 六月 11, 2025, https://arxiv.org/abs/2501.12911
48. Aidoc | Clinical AI Company | Rapid Responses, Smarter Care, 访问时间为 六月 11, 2025, https://www.aidoc.com/
49. What are the limitations of Explainable AI? - Milvus, 访问时间为 六月 10, 2025, https://milvus.io/ai-quick-reference/what-are-the-limitations-of-explainable-ai
50. Research Report 2024: Global Causal AI Market Set to Soar - GlobeNewswire, 访问时间为 六月 10, 2025, https://www.globenewswire.com/news-release/2025/01/08/3006403/28124/en/Research-Report-2024-Global-Causal-AI-Market-Set-to-Soar-to-USD-456-8-Million-by-2030-Propelled-by-Demand-in-Decision-Making-Tools.html
51. 6 Key Adversarial Attacks and Their Consequences - MindGard AI, 访问时间为 六月 11, 2025, https://mindgard.ai/blog/ai-under-attack-six-key-adversarial-attacks-and-their-consequences
52. Real- World Adversarial Attacks on AI Systems - IGI Global, 访问时间为 六月 11,

2025, https://www.igi-global.com/ViewTitle.aspx?TitleId=382260&isxn=9798337322001
53. Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems - Analytics for Cyber Defense (ACyD) Lab, 访问时间为 六月 11, 2025, https://acyd.fiu.edu/wp-content/uploads/GLOBECOM_Adverserial_Attacks_to_Health_Systemss.pdf
54. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions, 访问时间为 六月 10, 2025, https://www.medrxiv.org/content/10.1101/2024.08.10.24311735v2.full-text
55. The Evolving FDA Regulatory Landscape of Artificial Intelligence - Gardner Law, 访问时间为 六月 11, 2025, https://gardner.law/news/theevolving-fda-regulatory-landscape-of-artificial-intelligence
56. FDA has approved over 1,000 clinical AI applications, with most aimed at radiology, 访问时间为 六月 11, 2025, https://healthimaging.com/topics/artificial-intelligence/fda-has-approved-over-1000-clinical-ai-applications-most-aimed-radiology
57. FDA Approvals Surge for AI-Enabled Medical Devices in 2024 | Insights & Resources, 访问时间为 六月 11, 2025, https://www.goodwinlaw.com/en/insights/publications/2024/11/insights-technology-aiml-fda-approvals-of-ai-medical-devices
58. Artificial Intelligence in Medical Technology Myths vs. Facts, 访问时间为 六月 11, 2025, https://www.advamed.org/wp-content/uploads/2024/02/AI-Myths-vs.-Facts.pdf
59. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension | EQUATOR Network, 访问时间为 六月 11, 2025, https://www.equator-network.org/reporting-guidelines/consort-artificial-intelligence/
60. Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models - PubMed Central, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12138723/
61. Reasoning Beyond Accuracy: Expert Evaluation of Large Language Models in Diagnostic Pathology | medRxiv, 访问时间为 六月 11, 2025, https://www.medrxiv.org/content/10.1101/2025.04.11.25325686v1.full-text
62. Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review - PubMed Central, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11700560/
63. Evaluating the Efficacy of AI Physicians in Medical Dialogues - Bioengineer.org, 访问时间为 六月 11, 2025, https://bioengineer.org/evaluating-the-efficacy-of-ai-

physicians-in-medical-dialogues/
64. Why Is My AI Model's Performance Degrading? How to Solve Model Drift - 314e, 访问时间为 六月 11, 2025, https://www.314e.com/engineering-hub/why-is-my-ai-model-s-performance-degrading-how-to-solve-model-drift/
65. Strategies to prevent AI model data shifts in hospitals - healthcare-in-europe.com, 访问时间为 六月 11, 2025, https://healthcare-in-europe.com/en/news/prevent-ai-model-data-shift-hospitals.html
66. Model Drift: Best Practices to Improve ML Model Performance - Encord, 访问时间为 六月 11, 2025, https://encord.com/blog/model-drift-best-practices/
67. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions, 访问时间为 六月 10, 2025, https://www.medrxiv.org/content/10.1101/2024.08.10.24311735v2
68. Toward explainable AI in radiology: Ensemble-CAM for effective thoracic disease localization in chest X-ray images using weak supervised learning - Frontiers, 访问时间为 六月 10, 2025, https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1366415/full
69. Assessment of commercially available artificial intelligence software for differentiating hemorrhage from contrast on head CT following thrombolysis for ischemic stroke - PMC, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11915465/
70. Aidoc Secures Landmark FDA Clearance for Foundation Model AI, 访问时间为 六月 11, 2025, https://www.aidoc.com/about/news/aidoc-secures-landmark-fda-clearance/
71. Integrating predictive modeling and causal inference for advancing medical science, 访问时间为 六月 10, 2025, http://chikd.org/journal/view.php?id=10.3339/ckd.24.018
72. pmc.ncbi.nlm.nih.gov, 访问时间为 六月 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12025101/#:~:text=Compared%20with%20non%2Dexplainable%20AI,that%20supports%20clinical%20decision%20making.
73. Causal Deep Learning with Applications in Healthcare - Apollo, 访问时间为 六月 10, 2025, https://www.repository.cam.ac.uk/items/459d5932-b63b-42c6-a562-c0c8cbf9ec02
74. Causal ML: What is it and what is its importance? - Plain Concepts, 访问时间为 六月 10, 2025, https://www.plainconcepts.com/causal-ml/
75. What is the significance of causal inference in Explainable AI? - Milvus, 访问时间为 六月 10, 2025, https://milvus.io/ai-quick-reference/what-is-the-significance-of-

causal-inference-in-explainable-ai
76. A Review of Challenges and Opportunities in Machine Learning for Health - PMC, 访问时间为 六月 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7233077/
77. pmc.ncbi.nlm.nih.gov, 访问时间为 六月 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9300826/#:~:text=Causal%20inference%20is%20a%20broad,outcomes%20using%20many%20data%20types.
78. MedMimic: A Physician-Inspired Multimodal Fusion Framework for Early Diagnosing Fever of Unknown Origin - arXiv, 访问时间为 六月 11, 2025, https://arxiv.org/html/2502.04794v1
79. 1On Pearl's Hierarchy and the Foundations of Causal Inference - Elias Bareinboim, 访问时间为 六月 10, 2025, https://causalai.net/r60.pdf
80. Causal AI: How cause and effect will change artificial intelligence - S&P Global, 访问时间为 六月 10, 2025, https://www.spglobal.com/en/research-insights/special-reports/causal-ai-how-cause-and-effect-will-change-artificial-intelligence
81. The Role of Causal Inference in Drug Discovery and Development - ResearchGate, 访问时间为 六月 10, 2025, https://www.researchgate.net/publication/390284490_The_Role_of_Causal_Inference_in_Drug_Discovery_and_Development
82. Diagnostic AI — Paige.ai, 访问时间为 六月 11, 2025, https://www.paige.ai/diagnostic-ai
83. Paige PanCancer Detect Earns FDA Designation for Multi-Tissue Cancer Detection, 访问时间为 六月 11, 2025, https://www.targetedonc.com/view/paige-pancancer-detect-earns-fda-designation-for-multi-tissue-cancer-detection
84. Judea Pearl - Wikipedia, 访问时间为 六月 10, 2025, https://en.wikipedia.org/wiki/Judea_Pearl
85. An Introduction to Causal Inference by Judea Pearl (English) Paperback Book - eBay, 访问时间为 六月 10, 2025, https://www.ebay.com/itm/387644692494
86. Causal Inference in Statistics by Judea Pearl - Porchlight Book Company, 访问时间为 六月 10, 2025, https://www.porchlightbooks.com/products/causal-inference-in-statistics-judea-pearl-9781119186847
87. A deployment safety case for AI-assisted prostate cancer diagnosis - PMC - PubMed Central, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12131227/
88. Inka Health Launches Global AI-Oncology Consortium to Advance ..., 访问时间为 六月 11, 2025, https://trial.medpath.com/news/655d43ab7d111ffe/inka-health-launches-global-ai-oncology-consortium-to-advance-predictive-cancer-

research
89. Multimodal Intelligence Surges: From AI Assistants to Precision Medicine and Global Logistics, 访问时间为 六月 11, 2025, https://www.geekynews.org/?newsletter=multimodal
90. AI Tech and Services - Paige.ai, 访问时间为 六月 11, 2025, https://www.paige.ai/ai-technology-and-services
91. Paige.ai, 访问时间为 六月 10, 2025, https://paige.ai/
92. Peak AI performance without calibration or training - Paige.ai, 访问时间为 六月 11, 2025, https://www.paige.ai/publications/blog-post-title-one-y25lx
93. Viz.ai, the proven AI-powered care coordination platform, 访问时间为 六月 10, 2025, https://www.viz.ai/
94. Four New Studies Demonstrate that Viz.ai Finds New Patients with Hypertrophic Cardiomyopathy Earlier When Embedded into the Clinical Workflow, 访问时间为 六月 11, 2025, https://www.businesswire.com/news/home/20250326744258/en/Four-New-Studies-Demonstrate-that-Viz.ai-Finds-New-Patients-with-Hypertrophic-Cardiomyopathy-Earlier-When-Embedded-into-the-Clinical-Workflow
95. Two New Clinical Studies Demonstrate Accuracy of Viz.ai Algorithm for Intracranial Hemorrhage Volume Measurements to Advance Timely, Critical Treatment Decisions, 访问时间为 六月 11, 2025, https://www.viz.ai/news/clinical-studies-demonstrate-accuracy-of-viz-ai-algorithm-for-intracranial-hemorrhage
96. Latest clinical studies and publications | Viz.ai, 访问时间为 六月 11, 2025, https://www.viz.ai/publications
97. Real-world evaluation of the accuracy of the Viz.AI automated ..., 访问时间为 六月 11, 2025, https://pubmed.ncbi.nlm.nih.gov/39832899/
98. Automated Emergent Large Vessel Occlusion Detection Using Viz.ai Software and Its Impact on Stroke Workflow Metrics and Patient Outcomes in Stroke Centers: A Systematic Review and Meta-analysis - PubMed, 访问时间为 六月 11, 2025, https://pubmed.ncbi.nlm.nih.gov/40335883/
99. 访问时间为 一月 1, 1970, https.www.aidoc.com/about/news/aidoc-secures-landmark-fda-clearance/
100. Independent Evaluation of a Commercial AI Software for Incidental ..., 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11991795/
101. Tempus Introduces Loop, an AI-Powered Target Discovery and Validation Platform, 访问时间为 六月 11, 2025, https://www.tempus.com/news/tempus-introduces-loop-an-ai-powered-target-discovery-and-validation-platform/
102. Tempus Introduces Fuses, A Program Designed to Transform ..., 访问时间为 六月

11, 2025, https://investors.tempus.com/news-releases/news-release-details/tempus-introduces-fuses-program-designed-transform-therapeutic
103. SOPHiA DDM™ for Multimodal - SOPHiA GENETICS, 访问时间为 六月 11, 2025, https://www.sophiagenetics.com/sophia-ddm-for-multimodal/
104. Causaly: The Most Complete AI Platform For Life Sciences, 访问时间为 六月 10, 2025, https://www.causaly.com/
105. PathAI Showcases AI-Powered Pathology Innovations at the United States, 访问时间为 六月 11, 2025, https://www.pathai.com/resources/pathai-showcases-ai-powered-pathology-innovations-at-the-united-states-and-canadian-academy-of-pathology-uscap-114th-annual-meeting/
106. Precision for Medicine and PathAI Announce Strategic Collaboration ..., 访问时间为 六月 11, 2025, https://www.prnewswire.com/news-releases/precision-for-medicine-and-pathai-announce-strategic-collaboration-to-advance-ai-powered-clinical-trial-services-and-biospecimen-products-302438551.html
107. Comprehensive Review of Multimodal Medical Data Analysis: Open Issues and Future Research Directions - Acta Informatica Pragensia, 访问时间为 六月 11, 2025, https://aip.vse.cz/pdfs/aip/2022/03/09.pdf
108. A systematic review of challenges and proposed solutions in modeling multimodal data, 访问时间为 六月 11, 2025, https://arxiv.org/html/2505.06945v1
109. A systematic review of challenges and proposed solutions in modeling multimodal data - arXiv, 访问时间为 六月 11, 2025, https://arxiv.org/pdf/2505.06945
110. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century - PMC, 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11047988/
111. AMFGNN: an adaptive multi-view fusion graph neural network ..., 访问时间为 六月 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12066569/
112. Medical Heterogeneous Graph Transformer for Disease Diagnosis - Engineering Letters, 访问时间为 六月 11, 2025, https://www.engineeringletters.com/issues_v32/issue_12/EL_32_12_11.pdf
113. MULTIMODAL LARGE LANGUAGE MODELS FOR AUTOMATED DIAGNOSIS AND CLINICAL DECISION SUPPORT, 访问时间为 六月 11, 2025, https://aircconline.com/csit/papers/vol15/csit150515.pdf
114. Multimodal Data Fusion for Bioinformatics Artificial Intelligence 9781394269938, 访问时间为 六月 11, 2025, https://dokumen.pub/multimodal-data-fusion-for-bioinformatics-artificial-intelligence-9781394269938.html
115. AI Driven Healthcare: Leveraging Multimodal Data for Precision Health - IQVIA,

访问时间为 六月 11, 2025， https://www.iqvia.com/locations/middle-east-and-africa/blogs/2025/04/ai-driven-healthcare-leveraging-multimodal-data-for-precision-health

116. Causal inference in drug discovery and development | CoLab, 访问时间为 六月 10, 2025， https://colab.ws/articles/10.1016/j.drudis.2023.103737
117. 10 Causal AI Companies to Watch in 2025 | StartUs Insights, 访问时间为 六月 10, 2025， https://www.startus-insights.com/innovators-guide/causal-ai-companies/
118. Top Explainable AI (XAI) Companies - TopDevelopers.co, 访问时间为 六月 10, 2025， https://www.topdevelopers.co/directory/research/explainable-ai-companies/
119. Causal Inference Collaboratory - Institute for Public Health Practice, Research and Policy, 访问时间为 六月 10, 2025, https://iphprp.org/opportunities/faculty/collaboratories/causal-inference/
120. Top AI Medical Imaging tools for Healthcare Analytics - 2025 | Factspan, 访问时间为 六月 10, 2025， https://www.factspan.com/blogs/ai-medical-imaging-tools-transforming-healthcare-analytics-in-2025/
121. How 'causal' AI can improve your decision-making - IMD Business School, 访问时间为 六月 10, 2025， https://www.imd.org/ibyimd/artificial-intelligence/how-causal-ai-can-improve-your-decision-making/
122. Causality and scientific explanation of artificial intelligence systems in biomedicine - PMC, 访问时间为 六月 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11958387/
123. Maximizing the Value of Real-World Data: The Critical Role of Causal Inference in Pharmaceutical Innovation - Target RWE, 访问时间为 六月 10, 2025, https://www.targetrwe.com/research/news/maximizing-the-value-of-real-world-data-the-critical-role-of-causal-inference-in-pharmaceutical-innovation/
124. Milestones in the History and Development of Electronic Health Record Systems, 访问时间为 六月 11, 2025， https://www.sprypt.com/blog/emr-history
125. FDA Review of Radiologic AI Algorithms: Process and Challenges - RSNA Journals, 访问时间为 六月 11, 2025, https://pubs.rsna.org/doi/abs/10.1148/radiol.230242
126. 访问时间为 一月 1, 1970, https.www.businesswire.com/news/home/20250326744258/en/Four-New-Studies-Demonstrate-that-Viz.ai-Finds-New-Patients-with-Hypertrophic-Cardiomyopathy-Earlier-When-Embedded-into-the-Clinical-Workflow
127. 访问时间为 一月 1, 1970， https.www.paige.ai/diagnostic-ai