数 智 健 康 国 际 动 态

北京市卫生健康大数据与政策研究中心 2025.11.25

（十一）数智影像

计算技术和数据利用的快速发展加速了人工智能在临床研究和医疗领域的应用，特别是在临床医学影像领域。下面将从生成式人工智能在医学影像合成和临床数据分析两个角度来阐述其在提升诊断精度、实现疾病预测与个性化治疗方面展现出的巨大潜力、临床应用前景及面临的挑战，并提出未来发展的关键在于推动技术创新、制定行业标准并加强跨学科合作，以最终实现医疗质量的全面提升。

第一篇文章围绕生成式人工智能在医学图像合成中的潜力、应用与挑战展开系统探讨，旨在全面评估该技术对医学影像研究及临床实践的影响。自 2022 年以来，生成式人工智能已成为医学影像领域的变革力量，使得创建与现实世界数据极为相似的衍生合成数据集成为可能。在医学领域，这类技术已用于生成 X 光、MRI、CT 等影像，为数据增强、匿名化及疾病建模提供了新途径。多模态模型如 Med-Gemini、Med-Palm 等进一步拓展了其在诊断辅助、报告生成与视觉问答中的应用。生成模型分为基于物理的模型与统计模型两大类。前者依赖领域知识与物理规律构建高保真模拟，后者如 VAE、GAN、DDPM 则从数据中学习分布。两者各有优劣：VAE 采样快但质量较低，GAN 质量高但易发生模式崩溃，DDPM 质量与多样性俱佳但生成速度慢。图像质量评估是确保合成数据可用性的关键。除 SSIM、PSNR 等传统指标外，FID、IS 等统计指标被广泛使用。医学图像还需结合临床评估（如专家图灵测试）和文本-图像对齐指标（如 CLIP Score），以确保其解剖准确性与诊断价值。合成数据在医学影像中的主要应用——数据增强、隐私保护、图像补全与编辑、疾病建模与预测等。其中存在的挑战与伦理问题包括隐私风险、偏差问题、可解释性不足、数据来源不透明等。未来应建立临床导向的评估框架，推动混合模型（物理+统计）发展，并加强多学科合作以制定伦理与监管指南。FDA 已批准部分合成 MRI 技术，强调性能等效性与临床验证，为后续技术审批树立先例。生成式 AI 有望通过增强数据多样性、保护隐私及提升模型泛化能力，深刻改变医学影像的研究与临床范式，但其成

功依赖于持续的技术创新与严格的伦理监管。

第二篇文章系统描述了人工智能在临床影像数据分析中的应用现状、关键技术及未来发展趋势，重点聚焦于大型语言模型、基础模型、数字孪生等新兴技术在过敏等专科领域的实践与潜力。截止目前，人工智能已广泛应用于提升诊断准确性、医学影像分析、疾病风险预测、患者分层及实时健康监测。其中，传统机器学习模型（如随机森林、梯度提升）和深度学习方法（如 CNN、RNN、Transformer）在过敏性疾病如哮喘、特应性皮炎的诊断、表型分析和恶化预测中表现优异，AUC 常超过 0.8。可解释人工智能（XAI）方法如 SHAP、Grad-CAM 的引入，增强了模型透明度，助力临床信任与决策支持。新兴技术主要从以下三方面驱动临床 AI 变革——大型语言模型与多模态基础模型：能够整合电子健康记录、影像、基因组学和传感器数据，实现跨模态信息融合，推动精准诊断与个性化治疗。数字孪生：通过构建患者虚拟模型，模拟疾病进展与治疗反应，为动态干预和手术规划提供支持，其发展依赖于多模态数据集成与因果推断能力的提升。时间序列与生存分析：结合随机生存森林、Dynamic-DeepHit 等模型，实现对慢性病长期轨迹的预测与早期预警。过敏领域成为 AI 落地的重要场景，特别在诊断分类、急性加重预测和表型识别中成效显著。其他过敏性疾病如特应性皮炎、药物过敏和过敏性鼻炎的研究也逐渐增多，融合了可穿戴设备、传感器数据和影像分析，展现出 AI 在连续监测与个性化管理中的潜力。但是未来发展仍面临着一些挑战，包括数据隐私、质量异质性、模型泛化性及“黑箱”决策等问题，还有伦理与实施障碍、系统化发展路径、人才与协作等。人工智能正逐步重塑临床医学的研究与实践范式，其在过敏等专科领域的深入应用，有望实现更早诊断、动态预测和个性化干预，最终提升医疗质量与患者预后。

（徐健编辑）

译文一：

探索生成式人工智能在医学图像合成中的潜力：机遇、挑战与未来方向

Bardia Khosravi, Saptarshi Purkayastha, Bradley J Erickson, Hari M Trivedi, Judy W Gichoya, 徐健（译）

来源：The Lancet Digital Health.

时间：2025 年 11 月

链接：https://doi.org/10.1016/j.landig.2025.100890.

1. 简介

生成式人工智能是一类深度学习模型，能够创建与传统判定模型不同于以解释或决策为核心的内容。过去三年，生成式人工智能取得了快速进展，大型语言模型在 ChatGPT 的推出后获得了广泛关注。ChatGPT 是一类基于大量文本语料训练的模型，能够对用户问题做出连贯且真实的回答。大型语言模型在理解和生成自然语言方面展现出显著能力，为结合文本、视觉和语境理解的更高级多模态模型铺平了道路。这些大型多模态模型有潜力通过整合来自不同输入流的数据，助力包括医疗保健在内的多个领域。医学中大型语言模型的著名例子有 Med-PaLM 和 Med-Gemini，它们在回答医学问题、总结医疗文件以及基于患者症状和测试结果提出潜在鉴别诊断等任务中取得了令人期待的成果。此外，Med-Gemma 和 MedImageInsight 是基于不同类型医学图像训练的模型，包括放射学图像（如胸部 X 光、乳腺 X 光、CT）、皮肤科和眼科图像，这些图像允许用户通过语言和图像与模型交互（因此称为多模态基础模型）。这些多模态模型提供了非常规的视觉问答能力，并能从少数例子中学习以执行下游分类任务。

初步证据表明，生成式人工智能在视觉内容领域取得了显著进步，如 DALL-E、Stable Diffusion、Sora 和 Veo 等模型，这些模型在基于文本提示生成逼真的图像和视频方面表现出色。虽然这些模型主要以文本作为输入处理，有些模型还使用图像进行条件调节，但它们的主要重点是生成高质量图像。自 2022 年以来，医学影像

领域发表的开创性研究展示了生成式人工智能在创建逼真医学图像（合成数据）方面的潜力，提出了研究和临床应用的新方法。

本观点全面概述了医学影像中合成数据，并批判性分析了该领域的进展、应用与挑战。为此，本文探讨了各种图像生成范式，旨在评估这些生成技术如何改变医学影像研究的格局。探讨了这些模型及其衍生合成数据集的潜力，特别是它们增强和多样化医学研究资源的能力，以及它们在数据增强、匿名化和生物现象建模方面的益处。最后，讨论了使用合成数据的挑战，包括对严格评估指标和伦理考量的需求，并提出了可能显著促进医学影像领域的潜在研究方向。

2. 综合数据集

2.1 生成模型

合成数据领域仍处于萌芽阶段，尚未就一个普遍接受的统一定义达成共识。这种缺乏明确定义导致该术语在不同语境中使用和解释存在不一致，进而影响涉及合成数据研究的可重复性和透明度。英国皇家学会和艾伦·图灵研究所于 2022 年提出了合成数据的工作定义，指的是使用专门构建的数学模型或算法生成的数据，目的是解决一组数据科学任务。该定义强调合成数据的功能性和意图性，重点关注其在应对复杂科学挑战中的战略应用，而非仅仅模仿原始数据的统计属性。

生成式人工智能的发展引入了数据共享的新概念，我们称之为数据集模型。在这一概念中，生成模型学习并存储原始数据的模式和特征，包含在其内部参数（权重）中。这些训练权重包含了训练数据的关键特征和关系的压缩版本。与传统数据集共享涉及实际图像传输不同，共享模型权重提供了一种高效的替代方案，使他人能够生成具有与原始数据相似属性的新合成图像。这些合成数据集已被证明与原始数据极为相似，并捕捉了其分布，包括不同解剖特征的关系及其与不同病理过程的相关性。

生成模型主要分为两大类，能够生成合成数据集：基于物理的模型和统计模型。

基于物理的模型主要是基于规则的方法，通过数学方程和显式约束结合领域特定的知识和物理原理，生成真实且物理上合理的数据。这些模型不是直接从数据中学习模式，而是编码专家知识和已知的物理定律（如流体力学、组织生物力学或辐

射物理）来模拟生物现象。这些模型已成功应用于医学影像，模拟解剖结构（如股骨形状模型）、生理过程（如血管结构中的血流动态）以及医疗干预（如放疗规划中的辐射剂量分布模拟）。基于物理的模型提供高保真度和可解释性，但可能需要广泛的领域专业知识和计算资源。

与基于物理的模型不同，统计模型从数据模式和分布中学习（见图 1）。其中，变分自编码器（VAE）通过将数据压缩为低维表示，也称为潜空间，然后重建数据，从而有效捕捉数据分布。生成对抗网络（GAN）通过双网络系统运行，生成器创建数据样本，判定器评估这些数据样本并向生成器反馈。这种协同效应不断提升数据生成的质量和真实性。去噪扩散概率模型（DDPM）将噪声引入图像，并学习反转这一过程，从而产生高质量的采样。

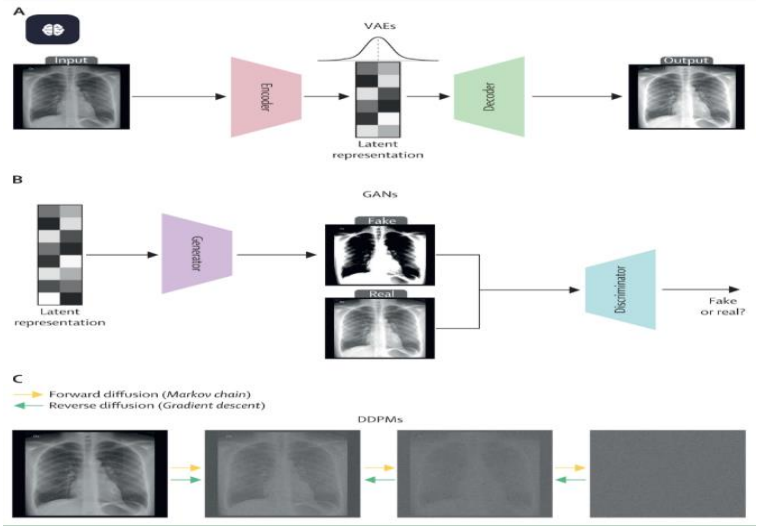

图 1 三种流行统计模型的架构及关键组成部分

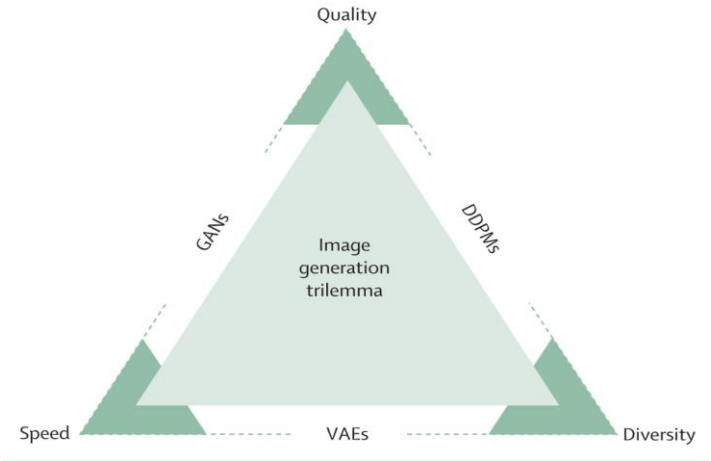

图 2 图像生成三难困境，代表生成模型的三个关键方面：多样性、质量和速度之间的权衡

统计模型面临生成式人工智能三难困境，涉及在高样本质量、全面模式覆盖和快速采样率之间取得平衡（见图 2）。VAE 以其快速采样能力著称，有时会导致采样质量较低。GAN 在生成高质量样本方面表现出色，但可能无法捕捉所有数据变异，导致模式覆盖率低，称为模式崩溃。DDPM 以其能够生成卓越质量和广泛模式覆盖的样本而著称，尽管采样速率较慢。最终用户选择与其感兴趣应用相匹配的生成模型，平衡所需的图像质量和速度。在数据集生成方面，优先事项通常转向确保高图像质量和全面模式覆盖，往往超过采样速度的担忧。

2.2 医学影像中的应用场景

生成模型及其合成数据集在医学影像领域有广泛应用（面板 1）。一个研究充分的用例是补充或替换真实数据，以训练深度学习模型，用于分类或分割等下游任务。生成图像可以基于类别标签（例如，有无肺炎）或描述性文本（例如，右中叶巩固）。研究表明，GAN 和 DDPM 生成的图像能显著提升下游病理分类器的性能。值得注意的是，随着向真实数据集添加更多合成数据，分类器性能会有所提升。在某些情况下，足够多的生成图像池可以匹配真实数据的性能优势，可能为数据共享开辟新途径，合成数据作为原始数据的替代。然而，在训练和评估生成模型时，需谨慎避免分布泄漏（即患者在训练和测试数据中均有代表性），这可能高估性能提升。值得注意的是，反复用其他生成模型的输出训练图像生成模型（通常超过三次迭代）存在模式崩溃的风险，从而降低最终模型的质量。

小组 1

合成成像数据集和图像生成模型在医学中的应用场景及其发现：

3D=三维。

AUROC=受试者工作特征曲线下的面积。

DDPMs=去噪扩散概率模型。

PSNR=峰值信噪比。

SSIM=结构相似指数。

我们使用以下术语组合进行了文献检索：“合成数据”或“VAE”或“GAN*”或“扩散模型*”以及“医学成像*”或“放射日志*”或“皮肤学*”或“病理学*”，并选取了符合各类合成数据的代表性论文纳入。

Chambon 等人（2022）：根据输入提示生成胸部 X 光片

•在合成与真实数据结合训练时，分类器性能提升了 5%

•仅在更大合成数据集上训练时，分类器性能提升了 3%

•经过微调后，气胸检测的文本编码表征提升了 25%

Pinaya 等（2022）：生成三维脑 MRI 并研究条件因素

•实现了可调年龄、性别和结构参数的真实 3D 脑 MRI 受控生成

•生成了 10 万张大脑图像的合成数据集供公众使用

Frid-Adar 等人（2018）：生成异常样本以应对 CT 扫描中肝病变检测中的类别不平衡

•使用合成数据增强，肝病灶检测灵敏度从 78.6%提升至 85.7%

•合成图像的特异性从 88.4%提升至 92.4%

•放射科医生在盲测中发现合成图像与真实图像无法区分

Khosravi 等（2024）：利用合成胸部 X 光片补充实像，扩展病理分类器的训练集

•在内部和外部测试集中，合成数据补充量的 10 倍，AUROC 提升了最多 0.02

•合成训练分类器在使用少 33% - 50%图像的情况下，性能可与真实数据模型匹敌

•结合真实数据和合成数据，将病理分类器的 AUROC 从 0.76 提升至 0.80，在跨源检测中

Ktena 等人（2024）：利用合成图像提升多模态下游分类器的公平性

•在合成和实像训练的胸部 X 光分类仪中，公平性差距减少了 44.6%

•病理切片间分布外预测准确率提升了 7.7%

•高风险皮肤镜敏感度提升 63.5%，公平性差距减少 7.5 倍

Conte 等（2021）：创建缺失的脑 MRI 序列以简化处理

•合成 MRI 序列下，肿瘤分割系数从 0.79 提升至 0.83

Rouzrokh 等人（2022）：引入和切除脑部 MRI 切片中的病灶

•有效彩绘（即根据标准选择性添加或移除特定图像部分，且不改变上下文）肿瘤成分、随机肿瘤和健康脑组织，使用 DDPM 进行

Khosravi 等（2024）：创建不同种族群体的反事实骨盆 X 光片，以评估大型影像数据集中的差异

•识别非裔美国患者与白人患者骨关节炎患病率的种族差异

•通过合成反事实骨盆 X 光片突出显示数据集尺度差异

Pérez-García 等人（2023）：通过创建反事实来对图像分类器进行压力测试，以评估可能

的捷径及其对模型性能的影响

•生成模拟获取、显现和种群变化的反事实数据集

•当 COVID-19 特征被移除后，COVID-19 分类器的准确率从 99.1%降至 5.5%

•人工拔除胸管后，气胸分类准确率从 93.3%降至 17.9%

Khosravi 等人（2023）：利用生成模型的内部特征进行标签高效的骨盆 X 光片分割

•使用生成模型特征，仅使用 20 个注释样本，提高了骨盆 X 线片分割准确度 0.30 - 0.32 点

Rouzrokh 等人（2024）：创建接受全髋关节置换术患者的合成术后图像

•合成术后髋关节 X 光片，髋臼平均角度为 39.9°（±4.6），99%处于安全区内

•合成 X 光片的效度（9.0±0.7）高于真实 X 光片（7.9±1.1）

袁等人（2024）：在阿尔茨海默病纵向研究中，基于过去或未来的扫描，将缺失的 3D 脑 MRI 归入

•实现了 0.895（颅骨切除）和 0.983（颅骨切除）的 SSIM，优于自编码器（颅骨切除 0.74，颅骨切除 0.91）和天真方法（骷髅切除 0.70，切除颅骨 0.89）

•体积误差率从 0.14（使用传统方法）降至 0.05

Kyung 等人（2024）：基于电子健康记录数据预测胸部 X 光形态

•在预测未来胸部 X 光病理方面，加权宏观 AUROC 为 0.72，优于仅表分类器和以往标签基线

•合成图像中性别（AUROC 0.96）和年龄（0.45）相关性保持

Liu 等（2025）：基于基线肿瘤特征和治疗方案预测肿瘤生长

•多参数 MRI 生成的 SSIM 为 0.92，PSNR 为 29.0，优于无治疗意识条件的基线模型

•生成的 MRI 质量在不同治疗日范围内保持较高，SSIM 根据治疗阶段范围为 0.88 至 0.94

•肿瘤生长预测在 4 个月内最为可靠，随着时间间隔从 0.5 个月延长至超过 24 个月，Dice 相似度系数从 0.85 降至 0.46。

生成模型在图像转换方面也表现出色。VAEs 和 GAN 长期以来实现了低剂量 CT 图像去噪，最终减少了患者的辐射暴露。近年来，加速 MRI 技术被用来将扫描时间缩短 30%。另一种图像到图像转换的应用场景是生成缺失的 MRI 序列，从而能够训练需要四个序列：T1、T2、显影后 T1 和 FLAIR 的下游算法。DDPM 使得修复成为可能，即根据标准选择性地添加或移除特定图像部分，而不改变上下文。例如，训练有素的扩散模型可以在健康的脑 MRI 中引入脑肿瘤病灶，或通过图像绘制图像来切除肿瘤区域。此类编辑可能丰富代表性不足的数据集，并引入罕见疾病，例如在阿

尔茨海默病患者中添加脑肿瘤。更高级的修复技术被开发出来，利用文字提示编辑胸部 X 光片的特定区域。编辑后的图像被用来对现有模型进行压力测试——例如，从气胸图像中去除胸腔引流管，以评估分类器的性能，且不涉及已知的混杂因素。

2.3 图像质量评估

评估生成图像的质量，决定了合成图像的使用方式，这一点至关重要。已提出多种指标以量化生成图像的质量，无论是否存在地面真实参考。这些指标大致可分为两类：图像指标和文本-图像指标（面板 2）。

第二组

基于医学图像生成应用场景的图像质量指标总结

BLIP=引导语言-图像预训练。

CAS=分类准确性得分。

CLIP=对比语言-图像预训练。

FID=弗雷谢起始距离。

IS=盗梦空间得分。

KID=核起始距离。

LLM=大型语言模型。

MMD=最大均值差异。

PSNR=峰值信噪比。

SSIM=结构相似指数。

图像超分辨率、去噪和修补

•SSIM：通过考虑亮度、对比度和结构，评估生成图像与参考图像之间的结构相似性

•PSNR：测量信号最大可能功率与生成图像与参考图像之间污染噪声功率的比值

类条件图像生成与无条件生成

•IS：利用 ImageNet 预训练的初始网络比较类别预测和生成样本的多样性

•FID：比较使用 ImageNet 预训练的初始网络从生成和目标分布中提取特征的均值和协方差

•KID：利用 ImageNet 预训练的初始网络计算生成分布和目标分布起始表示之间的 MMD 平方

域适应与类条件图像生成

•CAS：使用基于衍生医学图像训练的分类器，并评估真实图像的性能

感知**质量评估与现实性评估**

•人类图灵测试：医学专家区分真实图像与衍生图像

从文本描述生成图像

•基于分割的指标：对生成图像的不同器官进行体积分析，并与输入条件进行比较

•CLIPScore：计算文本描述嵌入与生成图像之间的余弦相似度

•BLIPScore：计算文本描述嵌入与生成图像之间的余弦相似度

•LLMScore：利用 LLM 在图像和不同对象层面创建详细说明，并将生成的说明与输入的文本描述进行比较。

图像指标

当有地面真实图像时——例如在超分辨率和去噪等任务中——可以使用传统指标如结构相似度指数和峰值信噪比来衡量生成图像与参考图像之间的相似度。然而，在缺乏基础真实信息的情况下——例如在类条件图像生成中——则需要替代的度量。例如，分类准确性评分基于衍生的医学数据训练分类模型，并评估其在实像上的表现，从而洞察生成模型的领域适应能力。

另一个广泛采用的指标是初始评分，它利用在 ImageNet 上预训练的初始网络来评估一组生成样本的类别预测。Fréchet 起始距离（FID）比较 ImageNet 预训练起始网络提取特征在生成样本与真实样本之间的均值和协方差。通过考虑目标分布，FID 比起始评分更能估算图像多样性。FID 提出了多种变体和改进——例如，核起始距离是 FID 的一种变体，允许使用少量样本进行度量计算，而 FID 计算则需要生成大量样本且资源密集。这些指标的一个局限是依赖于预训练网络，且与自然图像不同，医学影像中没有普遍接受的特征提取模型。

人体评估

除了计算指标外，人工评估仍然是评估生成医学图像质量的黄金标准。人类图灵测试涉及领域专家，他们被要求区分真实的医学图像和衍生的医学图像。该评估为生成图像的感知质量和真实性提供了见解，这对于医学影像至关重要，因为准确性和真实度至关重要。然而，由于感知质量和真实性是主观衡量标准，图像评估过程应包含不同经验水平的广泛参与者。

文本-图像指标

虽然图像指标仅关注生成图像的视觉质量，但文本-图像指标旨在衡量输入文本与生成图像之间的对齐。这些指标在医学图像生成任务中尤为重要，因为生成的图像需要准确反映医学状况或解剖结构的文本描述。诸如对比语言-图像预训练评分（CLIPScore）和引导语言-图像预训练评分（BLIPScore）等指标衡量输入文本与生成图像之间的相似度，量化两种模态之间的对齐程度。

图像-文本匹配是评估生成医学图像与其对应文本描述对齐度的另一组关键指标。构图质量指标通过将文本和图像分解为单独组成部分并测量其对应关系来评估这种对齐性，通常使用对象检测技术。这些指标不仅限于整体视觉相似度，更注重对文本中提及的特定解剖结构、病理或医疗状况的准确表现。通过确保生成的图像准确传达预期的医学信息，构图质量指标在医学教育和研究中发挥关键作用。

医疗保健专用指标

评估合成医学图像需要针对医疗需求的指标，而不仅仅是结构相似指数（FID）等通用工具。目前正在努力将现有指标调整以适应医学情境。例如，研究人员开始用基于RadImageNet等医学数据集训练的网络替换FID中的Image Net预训练模型，以创建医疗FID，更好地捕捉放射图像的统计特性。然而，医疗相关指标仍是活跃的研究领域，因为疾病分类者可能更多依赖局部特征而非整体特征。同样，解剖学准确性也被优先考虑，开发利用分割工具的测量方法，确保合成图像中关键结构（如器官或病变）得以保存。这些调整旨在解决标准指标的局限性，这些指标往往无法反映临床相关性或诊断效用。

建议下一步是将临床验证与这些计算方法相结合。人类评估，如人类图灵测试，已经涉及专家区分真实图像与合成图像，从而为医疗用途所需的感知质量提供了见解。对于文本引导图像生成，诸如CLIPScore等指标正在通过使用BioMedClip等医学基础模型进行优化。在实际临床任务中测试合成图像，如训练疾病检测分类器，可以进一步凸显其实用性。结合这些努力，可以提供强有力的、针对医疗护理的评估，从而确保合成图像符合推动医学影像研究和实践的技术和临床标准。

3. 潜力与承诺

合成数据生成和图像生成模型对医学影像研究的未来充满了巨大希望。通过利用生成模型的力量，研究人员能够解锁前所未有的数据多样性、隐私保护和多功能，

改变数据集创建、利用和疾病建模的方式。

3.1 数据集规模和多样性的增加

通过统计模型生成数据的一个主要优势是能够增加数据集的规模和多样性。初步证据表明，生成模型可以训练以解开数据中的特定关联，从而创造出现实数据集中难以获得的新组合。例如，训练于脑部 MRI 扫描的模型可以生成不同程度萎缩或病灶负荷的图像，这些图像与年龄或性别等因素无关。这种解缠使训练模型能够检测特定病理，同时不干扰其他变量的影响。如前所述，通过生成图像补充增加的数据集规模，可以提升下游模型性能。此外，通过合成数据生成对少数群体或罕见病患者进行有针对性过度抽样，已被证明能缩小公平差距 40%。合成数据生成通过增加代表各子群体原始数据集分布的数据集规模，缩小了这一公平性差距。

3.2 隐私保护

合成数据集为医学研究中数据共享和利用面临的挑战提供了一种保护隐私的解决方案。生成式人工智能通过生成逼真的图像来匿名化敏感的患者信息，这些图像在视觉和模型特征空间中都模拟真实患者数据的生物特征，而无需直接复制原始数据。这种匿名化使得创建可共享和分析的数据集而不损害患者隐私，进一步开辟了协作研究的新途径，并促进了医学影像中稳健且符合隐私的人工智能模型的发展。

3.3 跨任务的多样性

图像生成模型，尤其是 DDPM 的另一个关键潜力在于其多功能性。基于医学图像训练的生成模型可以被改编和重新利用，用于除补充数据外的多种任务;例如，从无监督图像生成模型中学到的特征可用于少镜头图像分割，仅凭 20 个专家注释样本即可准确描绘解剖结构或病理。同一模型无需额外训练，也可以用于修复，生成多样化的训练样本。同样，无需初始训练后微调的生成模型可用于医学图像中的异常检测。这种多功能性扩展了合成数据集及其生成器模型的价值，因为单个模型可用于多个下游应用，简化了研究工作流程，减少了针对特定任务的数据收集和模型开发的需求。

3.4 复杂生物现象的建模

先进的生成模型可以通过训练过程内化复杂的生物现象，从而能够建模和模拟复杂的生理过程。这一内化的世界模型可以应用于超越前述下游任务的新颖应用。这一能力的一个显著例子是术后影像表现的预测。当这些模型在大量成对的基础成形术和关节置换术后盆腔 X 光片上训练时，能够生成高度逼真的术后 X 光片，模拟一场精心执行的手术。令人惊讶的是，领域专家评估生成的术后图像比真实图像更为稳健且解剖学准确，凸显了这些模型作为虚拟手术规划工具和教育资源的潜力。

这一内化世界模型的另一个重要应用是疾病进展的预测。例如，当获得初步脑 MRI 扫描和患者的治疗方案信息时，先进的 DDPM 可以生成一系列图像，描绘脑肿瘤随时间可能进展的情况。通过了解疾病特征、治疗效果与生物过程之间的复杂相互作用，这些模型能够为患者预后提供宝贵见解，并有助于临床决策。

4. 挑战与考虑

尽管衍生合成数据集和图像生成模型在医学影像研究中具有巨大潜力，但仍需解决若干挑战和伦理考量，以确保其负责任且有效的应用。第三组总结了这些挑战，并提出了一些未来的研究方向以缓解它们。

第三组

生成式人工智能和合成数据集在医学影像领域的挑战、考虑因素及未来研究方向总结

数据复制：

生成模型在复制与原始数据极为相似的图像时，可能会无意中暴露敏感的患者信息。

未来研究方向：

•创建衡量生成图像隐私风险的指标

•开发事后数据匿名化方法

•探讨图像质量与隐私保护之间的权衡

源数据集的识别：

识别用于训练生成模型的具体数据集具有挑战性，这会阻碍对生成数据中潜在偏差或局限性的评估。

未来研究方向：

•为合成医学影像数据集制定标准化报告指南

•生成模型中数据集指纹识别技术的发展

•为合成医学数据集创建可信的第三方验证服务

•探索逆向工程模型训练数据的方法

可解释性与可解释性：

生成模型的复杂性使得理解这些模型如何学习和生成数据变得具有挑战性。这种理解对于建立对模型输出的信任至关重要。

未来研究方向：

•在随机预测模型中实施不确定性量化方法

•创建临床相关的可解释性指标

•为临床医生开发交互式可视化工具，帮助他们探索模型决策

•研究领域知识与模型解释的整合

潜在偏差：

源数据集中的偏见可能在生成的数据中传播或放大，导致研究结果偏颇或应用歧视。

未来研究方向：

•为评估医学影像生成模型公平性制定基准

•建立多机构合作，创建符合人口统计学平衡的培训数据

•研究数据增强对偏倚减少的影响

4.1 患者隐私与数据复制

尽管合成数据集可以通过生成匿名数据来保护患者隐私，但关于潜在数据复制的担忧仍然存在。如果生成模型是在特定数据集上训练的，并且能够复制与原始数据极为相似的图像，那么模型可能会无意中泄露敏感的患者信息。复制是指数据集中存在多份图片或说明文字，这不仅需要仔细整理数据，但也引发了对训练数据匿名化程度及重新识别可能性的担忧。与表格数据不同，医学图像在像素值中嵌入患者身份信息，因此在匿名化方面带来了独特的挑战。例如，脑部 MRI 中的面部特征或 X 光片中独特的解剖标记，即使去除了明确的患者身份，也可能实现重新识别。

研究人员需要仔细评估数据复制的风险，并采取措施减轻这一问题，如采用差异性隐私技术或事后数据匿名化。过去四年合成数据隐私评估指标的进展，如会员

推断攻击和真实样本与生成样本的相似度评分，有助于量化隐私风险。此外，包括内容来源与真实性联盟（C2PA）和谷歌的 SynthID 在内的新兴合成内容来源标准也已制定，用于标记人工智能生成内容，同时解决透明度和知识产权问题。

4.2 来源数据集的识别与披露

对用于训练生成模型的源数据集保持透明，对于确保研究结果的完整性和可重复性至关重要。然而，识别具体的训练数据集可能具有挑战性，尤其是在模型在多个专有来源上训练，或研究人员使用预训练模型却未完全了解其训练数据时。这种透明度不足会阻碍评估生成数据中潜在偏差或局限性的能力。为弥补这一空白，研究人员应努力记录并披露训练过程中使用的所有源数据集，以便更好地理解和验证衍生数据。此外，应用于推理的特定超参数、特定类别或提示词条件，以及创建合成数据集的每一个后处理步骤，都应与模型或数据集发布一同发布，以确保后续工作的可重复性和适用性。应采纳数据集文档指南，如 2024 年发布的 STANDING Together 指南，用于合成数据生成模型。

4.3 可解释性与可解释性

随着生成模型日益复杂，其可解释性和解释性将变得更加困难。理解这些模型如何学习和生成数据，对于建立对其输出的信任，并确保其在医学影像研究中的安全可靠使用至关重要。尽管存在一些专门用于生成模型的解释方法，以确保对输入文本的正确理解或为数据集添加不确定性指标，但这些方法在医学影像中的适应和评估仍然受限。

4.4 潜在偏差

使用合成数据集和生成模型引发了重要的偏见考虑。源数据集中的偏见可能在生成的数据中传播或放大，是关键问题。如果训练数据偏向某些人口统计学、病理或影像协议，生成的数据可能会延续这些偏见，导致研究结果偏颇或歧视性应用。例如，历史上许多医学影像数据集中少数群体代表性不足，导致人工智能系统在不同人口统计群体间的性能可能存在差异。当代表性较低时，生成模型可能难以真实

捕捉这些代表性不足群体的分布。然而，2023 年的一项研究表明，当整体数据集足够大以捕捉高层特征时，更新的生成模型甚至可以从少至 20 个样本中获得有意义的表示。此时的缓解策略包括训练时的多样性感知采样、对抗性去偏见技术、模型目标中的明确公平约束，以及利用新型生成模型的少数样本微调能力。研究人员需要积极评估和减轻源数据中的潜在偏见，并定期审核生成数据的公平性和代表性。

5. 未来方向

生成式人工智能在医学影像领域正在快速发展，多个关键研发领域有望推动合成数据集和图像生成模型的能力和应用。一个关键方向是开发更健全、标准化的评估框架，以考虑医学影像的独特挑战和需求，包括建立临床相关的指标、基准数据集，以及不同生成模型的比较分析和验证所面临的挑战。

另一个重要方向是探索新型架构和训练策略，例如结合物理学和统计方法，并融入领域特定知识和约束的混合模型。将生成模型与其他人工智能技术（如强化和主动学习）结合，有望创建个性化且针对患者的数据集，用于精准医疗和有针对性治疗计划。

解决合成数据集和图像生成模型使用相关的伦理和监管挑战对于实现其全部潜力至关重要，这需要研究人员、临床医生、伦理学家和政策制定者合作，制定负责任使用、数据隐私、同意和问责制的指南和最佳实践。包括美国食品药品监督管理局（FDA）和欧洲药品管理局在内的监管机构，将在建立临床应用合成数据验证和批准框架中发挥关键作用。合成医学影像评估框架已经在形成中，FDA 已批准合成 MRI 技术。这些技术被监管为图像处理软件，而非完全新颖的技术，FDA 要求大量临床验证以证明放射科医生在使用合成图像与传统图像时的诊断表现相当。这一监管先例为未来合成数据技术提供了一条路径：标准化诊断任务的性能等效性证明、多读取器进行严格的临床验证，以及市场后监测承诺以监测临床结局的任何差异。

总之，衍生合成数据集和图像生成模型有潜力改变医学影像研究和临床实践。解决相关挑战、建立最佳实践以及投资于研究和创新，有助于充分发挥生成式人工智能在改善患者护理、推动科学发现和改变医学影像格局方面的全部潜力。

译文二：

临床数据分析中的人工智能：大型语言模型、基础模型、数字孪生及过敏应用综述

Yutaro Fuse, Shawn N. Murphy, Hisahiro Ikari, Akiko Takahashi, Kenshiro Fuse, Eiryo Kawakami, 徐健（译）

来源：Allergology International.

时间：2025 年 10 月

链接：https://doi.org/10.1016/j.alit.2025.06.005.

1. 介绍

现代医学和医疗保健的总体目标是通过预防措施和健康促进，解放人类免于疾病，促进整体福祉。为此，直接从活体人体获得的临床数据对医学研究人员来说是一笔丰富的“宝藏”。近年来计算能力的进步加速了人工智能（AI）在包括临床数据分析在内的多个学科中的应用（见表 1）。

表 1 临床数据分析中代表性的人工智能模型及其关键特性

方法	数据兼容性					型号特征		主要解释
	表格	图像	发短信	时间序列	模态	高过拟合风险	处理非线性	
基于规则的系统	✓					✓		手动编码的专家规则
逻辑回归	✓					✓		变量与对数赔率之间的线性关系
脊/套索回归	✓							带有 L2/L1 正则化的线性模型
天真贝叶斯	✓		✓					假设特征独立性的贝叶斯分类
支撑向量机（SVM）	✓		✓				✓	最大化类别间的间距;非线性数据的核技巧
决策树	✓					✓	✓	特征空间的递归划分
随机森林	✓						✓	决策树集合
梯度增强机	✓						✓	顺序集合学习
卷积神经网络 CNN	✓	✓					✓	空间特征提取的卷积
循环神经网络 RNN			✓	✓			✓	通过循环回路获得顺序信息
变压器	✓	✓	✓	✓	✓		✓	注意力机制;并行序列处理

方法	数据兼容性					型号特征		主要解释
	表格	图像	发短信	时间序列	模态	高过拟合风险	处理非线性	
视觉变换器（ViT）		✓			✓		✓	将变换器应用于图像补丁
自然语言处理			✓				✓	基于变压器的预训练+微调

表格中出现勾选标记表示常见或代表性的用例和特性。

这一趋势进一步受到电子健康记录（EHR）和医疗设备大规模数据集加速积累，以及大型语言模型（LLM）和多模态基础模型的出现推动。因此，将人工智能技术融入临床实践的趋势迅速加快。

在这样的背景下，出现了两个核心问题：人工智能目前在临床数据分析中能取得哪些成就？我们预计近期将有哪些进展？在本综述中，我们聚焦于数据临床医生经常遇到的临床和研究问题，并探讨可用的人工智能技术来应对这些问题。本综述旨在探索适用于临床医学的更广泛人工智能方法，并通过选定实例突出其与过敏相关的潜在相关性，而非仅限于现有过敏特异性研究。此外，我们还概述了近期研究成果及机器学习在过敏领域的技术基础，强调当前挑战及克服这些挑战的潜在策略。最后，通过引入值得关注的新兴概念和技术，我们旨在为临床医生和研究人员提供开展人工智能研究及其在临床环境中成果实施的路线图（见图 1）。因此，本综述旨在为临床医生、研究人员和政策制定者提供全面的指南，特别强调与过敏研究相关的进展。

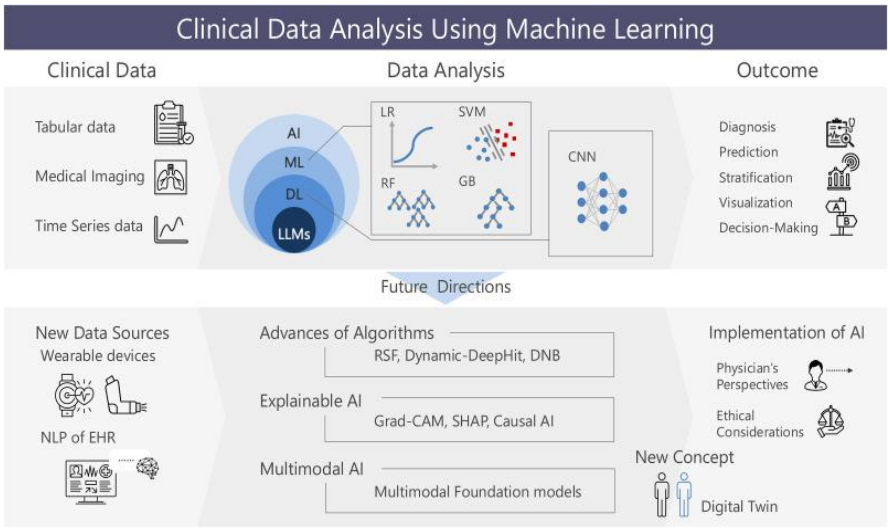

图 1 利用机器学习进行临床数据分析的图形摘要

临床数据分析中人工智能当前应用及未来方向的示意总结。通过机器学习、深度学习和大型语言模型分析多样化的临床数据源——包括表格数据、时间序列数据和

医学影像，这些都属于更广泛的人工智能（AI）领域。代表性的机器学习技术包括逻辑回归、支持向量机、随机森林和梯度增强，而卷积神经网络则代表深度学习方法。这些方法实现了多种临床结果，包括提高诊断准确性、疾病发病和进展预测、患者分层、健康与疾病状态可视化以及临床决策支持。新数据源，如可穿戴设备和电子健康记录的自然语言处理，也正作为关键输入出现。分析方法论的未来方向包括算法开发、可解释人工智能以及基于基础模型的多模态人工智能的进展。数字孪生的概念被强调为个性化医疗中的一种新范式。此外，医生的视角和伦理考量是人工智能在临床实践中成功应用的关键组成部分。人工智能（AI）;机器学习（机器学习）;深度学习（DL）;大型语言模型（LLM）;逻辑回归（LR）;支持向量机（SVM）;随机森林（RF）;梯度增强（GB）;卷积神经网络（CNN）;自然语言处理（NLP）;电子健康记录（EHR）;随机生存森林（RSF）;动态网络生物标志物（DNB）;梯度加权类别激活映射（Grad-CAM）;SHAPLEY 加法解释。

2. 人工智能技术在临床数据分析中的应用领域

2.1 提升诊断准确性

临床诊断涉及确定疾病是否存在并分类其特定类型。临床数据通常包含许多变量，可能作为诊断线索。机器学习的一个关键优势是能够全面分析这些变量，从而适当选择、权重和分类诊断所需的信息。

传统诊断方法通常依赖统计方法，如计算比值比或风险比，这些作为先验概率，并通过多变量分析将这些指标结合，以获得更准确的诊断支持。然而，这类统计模型通常依赖线性或逻辑回归，这限制了可行纳入的变量数量。此外，随着变量数量的增加，模型不稳定性增加，捕捉非线性关系的能力下降，这使得对临床数据进行真正全面的分析变得困难。

为解决这些问题，机器学习模型已被广泛应用于多个过敏相关研究领域的诊断应用，包括哮喘，嗜酸性食管炎，以及阻塞性呼吸疾病的分化。其中，基于树的集成模型尤其受到关注。两个显著的例子是随机森林，其中大量决策树通过多数投票和梯度提升随机生成和整合，它在学习过程中迭代完善决策树。在系统性红斑狼疮（SLE）的诊断中——这一疾病传统上需要整体评估身体表现及血液和尿液检测结果——随

机森林在通过 50 多个不同特征识别 SLE 时，报告准确率达 92.89%。随机森林能够同时考虑多个变量并捕捉非线性关系，这很可能是实现如此高准确性的关键。梯度提升模型还通过在学习过程中顺序改进决策树的构建，有效地为变量赋予权重，从而解决非线性关系。例如，Maintz 等人基于 130 项患者背景因素和血液检测项目评估特应性皮炎严重程度的研究，在 ROC（AUC-ROC）下实现了 0.71 的区域。这些发现凸显了机器学习在临床诊断中日益增长的应用价值。

2.2 医学图像分析与诊断支持

医学图像涵盖了广泛的格式，从静态图像（如计算机断层扫描、磁共振成像、PET 和病理切片）到通过内镜和腹腔镜手术获得的动态视频数据。传统上，这些图像通过提取特征值并应用经典机器学习算法（如支持向量机）来处理。然而，自 2010 年代末以来，深度学习的快速发展使卷积神经网络（CNN）成为医学图像分析的主流方法，基于 CNN 的方法现已被广泛应用于临床研究和应用中。

CNN 在临床图像分析中的应用在训练数据丰富领域进展最快。在 Esteva 等人（2017）中，CNN 模型在超过 12 万张图像上训练——包括皮肤癌、良性肿瘤和炎症性皮肤疾病病例——在识别恶性病变方面达到了皮肤科医生级别的准确性。值得注意的是，数据集包含了视觉特征重叠的炎症性疾病，凸显了 CNN 在复杂诊断任务中的实用性。反映这一成功，CNN 现已应用于过敏相关皮肤病。例如，一个用 4740 张临床图像训练的 CNN 模型在区分银屑病、湿疹和特应性皮炎方面达到了 95.8%的准确率，强调基于图像的诊断在过敏护理中的潜力。

近年来，视觉转换器（ViT）已成为图像分析中 CNN 的有前景替代方案。Transformer 架构最初为自然语言处理任务开发，采用注意力机制捕捉输入特征之间的长距离依赖关系，从而实现对图像内容的更全局理解。在图像分析中，ViT 在多个基准测试中表现优于传统 CNN 模型。然而，ViT 的高准确性通常依赖于大量数据，这在过敏领域存在挑战，因为该领域数据稀缺较为常见。

为解决这一限制，基金会模型成为一个显著的概念。该方案由 Bommasani 等人于 2021 年左右提出，其特点是一个在庞大数据集上训练的大规模 AI 框架，能够在多种下游任务中实现稳健的性能。在医学影像中，关键要求是对关注区域的准确分割。医疗支援人员，一个专为医学图像分割设计的基础模型，因其广泛的适用性和在各

类影像模态中的通用性而备受关注。MedSAM 基于超过 157 万张医学图像开发，在 86 个内部验证任务和 60 个外部验证任务中，展示了比专门模型更优的准确性和鲁棒性。在病理学中，UNI 通过超过一亿张图像训练，不仅实现了分割，还实现了投资回报率（ROI）的检索和癌症亚型分类。Prov-GigaPath 利用 13 亿张图像对九种主要癌症类型进行了分类。与此同时，在眼科领域，RETFound 由 160 万张视网膜图像构建，支持多种眼疾病的诊断。这些例子凸显了 Foundation 模型在医学影像领域的快速演变。未来几年，基金会模型开发的持续进展预计将加速医学影像的发展，显著提高诊断准确性，即使是过敏领域常见的罕见疾病。

2.3 识别与预测疾病风险

医学的主要目标之一是准确预测疾病的发病、进展和死亡风险。当能够准确预测时，临床医生可以缩小所需的诊断测试范围并实施早期干预，从而显著提高预后改善的可能性。

尽管传统临床评分和指标在过敏及其他领域被广泛应用，但在风险预测中存在显著局限性。湿疹面积和严重程度指数等工具，对于评估当前疾病严重程度非常有用，但本质上具有描述性，不适合预测。即使是预测性指数，比如哮喘预测指数，努力捕捉个体间疾病的完整复杂性和异质性。此外，这些模型通常难以适应不断变化的临床环境——包括地理、人口和时间变化——且主要基于静态数据，限制了其对动态疾病进展的解释能力。随着临床数据集日益庞大和多元化，这些限制凸显了对更灵活、可扩展和数据驱动的疾病风险预测方法的紧迫需求。

作为回应，机器学习作为解决这些问题的途径逐渐受到关注。通过通过脊回归或套索回归等正则化方法扩展传统统计方法，系统性识别相关变量而不必忽视关键预测变量。视觉化方法，如计字图，进一步促进了对风险评估的清晰直观理解。除了这些方法外，非线性模型（例如支持向量机（SVM），随机森林和梯度增强）相比线性模型展现出更优越的预测性能，这使得患者症状多样性能够更细致地呈现。此外，时间序列和动态建模方法（见下文介绍）使得能够适应不断变化的患者数据并随时间更新的模型成为可能。

然而，这些机器学习技术仍面临持续挑战：其推理过程难以阐明。这种所谓的“黑匣子”特性一直是基于人工智能预测模型广泛临床应用的主要障碍。

近年来，可解释人工智能（XAI）出现以应对这一挑战。XAI 涵盖了清晰阐述人工智能模型如何得出结果或决策的框架。该概念通过国防高级研究计划局（DARPA）XAI 项目获得了广泛认可，该项目自 2017 年起为期四年。一种著名的 XAI 方法是 SHapley 加法解释（SHAP），它基于博弈论计算 Shapley 值，以量化每个输入变量对最终预测的贡献。SHAP 的一个显著特点是其具备局部和全局可解释性的能力——能够突出每个变量在特定实例中以及整个模型中的重要性。这种双重解释性使其在过敏领域得到成功应用，SHAP 被用来提升青霉素过敏预测模型的透明度。最终的可视化清晰展示了每个变量对预测的贡献，使模型输出更直观、更易理解。另一项研究中，Yonehara 等人报告了利用 XAI 技术分析前段裂隙灯图像，诊断过敏性结膜疾病。其他 XAI 方法，包括自解释神经网络，可解释的 SVM，局部可解释的模型无关解释，梯度加权类激活映射，分数加权职业激活映射，使用概念激活载体进行测试，自动化基于概念的解释，图神经网络解释器，以及可解释图神经网络，也被开发用于增强机器学习模型的透明度和信任度（见图 2、图 3）。这些进展凸显了人工智能模型的新时代，既实现了预测准确性又可解释性，从而加速其在现实临床实践中的应用。

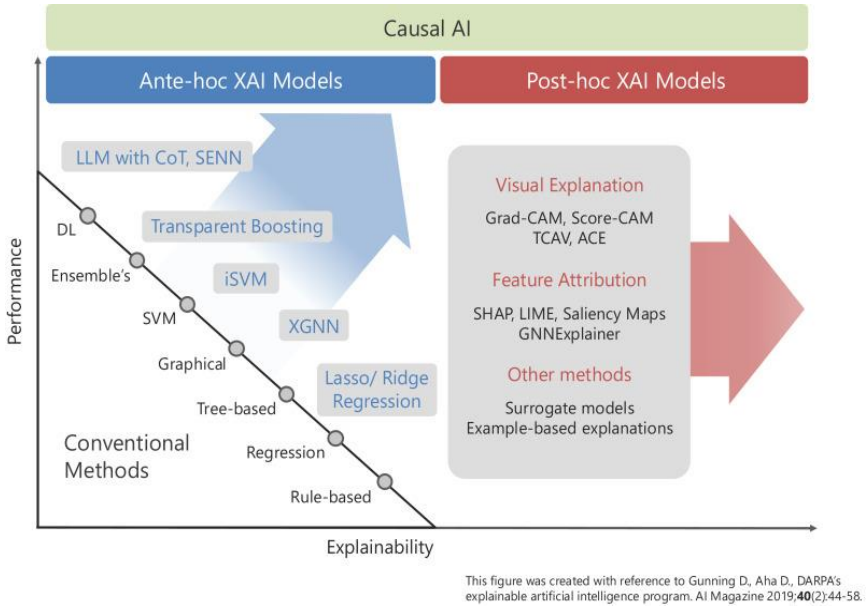

图 2 可解释人工智能（XAI）方法在模型性能和可解释性方面的情况

展示模型性能与可解释性在各种机器学习方法中的权衡。对角线代表传统模型，涵盖从基于规则的系统到深度学习，反映了它们在预测准确性和内在可解释性方面的相对位置。现代机器学习的一个关键挑战是同时提升性能和可解释性——这是可解释人工智能（XAI）发挥关键作用的领域。XAI 方法大致分为两类：先行模式和事

后模式。前置模型（左侧）本质上是可解释的。这些包括带有思维链提示的大型语言模型、自解释神经网络、透明增强、可解释的支持向量机以及可解释图神经网络。这些模型旨在确保透明度，同时显著牺牲预测性能。相比之下，事后模型（右侧）在模型训练后应用，以解释模型行为。这些方法进一步分为：（i） 视觉解释方法，如梯度加权类别激活映射（Grad-CAM）、分数加权 CAM、概念激活向量测试以及自动化概念解释;（ii） 特征归因方法，包括 SHapley 加法解释、局部可解释模型无关解释、显著性图和图神经网络解释器;（iii） 其他方法，如代理建模和基于实例的解释。最后，因果人工智能被视为一个有前景的新方向，旨在实现高预测性能和稳健的可解释性。人工智能、人工智能;XAI，可解释人工智能;深度学习;大型语言模型（LLM）;CoT，思维链;SENN，自我解释神经网络;SVM，支撑向量机;iSVM，可解释的支持向量机;XGNN，可解释图神经网络;Grad-CAM，梯度加权类别激活映射;评分-CAM，分数加权班级激活映射;TCAV，概念激活载体测试;ACE，自动化概念解释;SHAP，SHAPLEY 补充解释;LIME，局部可解释的模型无关解释;GNNExplainer，图神经网络解释器。

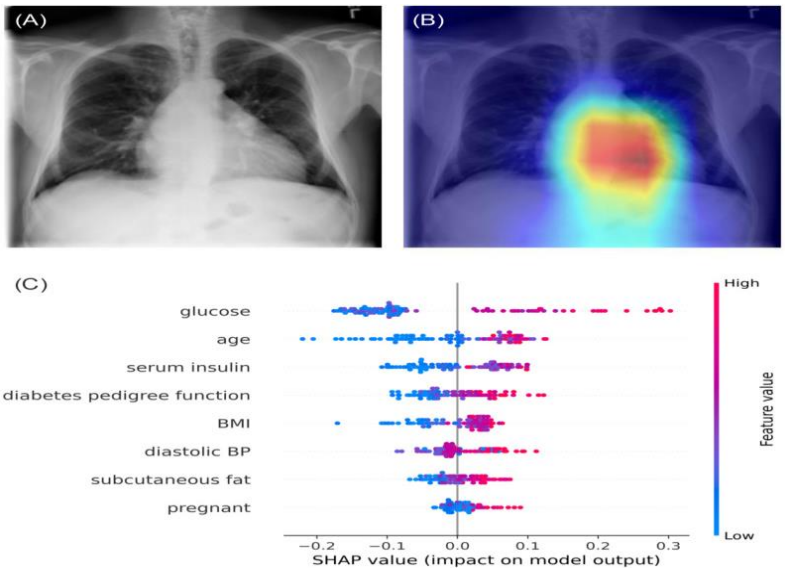

图 3 应用于临床数据的可解释人工智能模型示例

(A) 来自 NIH 胸部 X 光数据集的胸部 X 光图像，作为卷积神经网络（CNN）模型的输入，用于预测心脏肥大。 **(b)** 为 CNN 模型做梯度加权类激活映射（Grad-CAM）可视化，突出对心脏肥大预测贡献最大的关键区域（红色），从而提供可视化解释性。该模型是使用 TorchXRayVision 库开发和分析的。**(C)**SHapley 加法分析（SHAP）总结图，展示了利用皮马印第安人糖尿病数据库预测糖尿病状态的随机森林模型特

征归因。每个点代表一个单独的患者，颜色表示特征值（蓝色：低;红色：高）。特征按其对模型输出的贡献进行排序，从而为预测提供可解释性。SHAP 分析使用 SHAP 库进行。BMI，身体质量指数;血压、血压。

2.4 使用时间序列数据进行疾病进展建模与患者分层

临床人工智能的一个关键目标是通过预测疾病的发展以及哪些治疗对特定患者最有效，从而实现精准医疗。时间序列建模在这一工作中起着核心作用，尤其是在具有动态轨迹的疾病中。本节将概述主要的时间序列分析方法，主要参考普通医学中此类技术已成熟的实例。随后我们讨论了它们对过敏性疾病的潜在适用性，这些疾病由于慢性和复发的病程、季节性加重以及环境和个体因素驱动的变异性，常呈现出独特的时间模式。

现有时间序列分析的大部分研究主要集中在急性状况上，使用如 MIMIC-III 等数据集，其中包含高频测量数据，如实验室数据和呼吸机设置。在此背景下，循环神经网络——尤其是长短期记忆（LSTM）网络——因其能够捕捉时间依赖关系并减轻梯度消失而被广泛采用。LSTM 在重症监护室再入院和败血症预测等应用中已证明其实用性。同时，使用膨胀因果卷积的时间卷积网络（TCN）在训练稳定性、并行性和计算效率方面具有优势。例如，Chen 等人基于 MIMIC-III 的 17 个变量开发了基于 TCN 的模型，预测 48 小时死亡风险时 AUC-ROC 为 0.837。在观察值在死亡等关键结果发生前发生剧烈变化的情境中，这些深度学习方法被证明特别有效。

除了预测外，时间序列数据还通过聚类实现疾病分层。时间序列聚类可以识别具有相似进展模式的患者亚组，从而洞察疾病异质性。例如，术后骨密度轨迹的聚类已被用来区分有骨质疏松相关并发症风险的患者。该技术有潜力帮助临床医生识别高风险人群进行早期干预，并突出疾病亚群的异质性。

时间序列数据还可以用于捕捉健康中潜在且不可观察的方面，如疾病进展和生理状态。隐马尔可夫模型特别适合估计离散潜态之间的转变，并已被应用于患者健康状况的模型变化。连续时间隐马尔可夫模型（CT-HMM）通过整合状态转变的时序来扩展这一框架，使其非常适合处理不规则采样的医疗数据。CT-HMM 已被用于阐明多发性硬化症患者的疾病阶段进展及残疾获得机制。相比之下，状态空间模型在模拟由连续值表示的潜在状态动态变化时更具优势。这些模型将观测到的多变量时间

序列数据分解为观测方程和状态方程，从而能够估计随时间演变的潜变量。这使患者能够持续跟踪临床轨迹和风险波动，在每个时间点提供动态且个性化的风险评估。此外，状态空间模型与神经网络架构的整合为分析高维医疗数据开辟了新可能。

过敏研究中的时间序列分析具有显著的未来应用潜力。过敏性疾病的特征是慢性复发、季节性加重以及由环境和个体因素引发的波动。这些特征需要在时间尺度和分辨率上与急性状况分析方法不同的分析方法。然而，随着纵向时间序列数据在更长时间内的日益公开，将这些先进分析方法应用于过敏研究将变得更加可行。例如，SLE 中，纵向背景因子和自身抗体谱的聚类，使得患者分层为四个具有明显 10 年结局的患者亚组。这表明时间序列聚类在分层异质患者群体中的价值。此外，时间序列专用基础模型的发展，如 Chronos，为预测建模开辟了新途径。这些模型训练于多种时间模式，并且越来越能处理稀疏和不规则的时间序列数据。通过整合静态特征（如患者人口统计和基线特征）以及动态变量（如实验室检测结果、物理发现和疾病活动），基础模型最终有望实现对过敏领域疾病轨迹和患者分层的高度准确预测。

2.5 物理状况和疾病的监测与可视化

移动健康（m-Health）的快速扩展和可穿戴设备的演变正在改变慢性病的监测方式。传统上，临床医生依赖间歇性评估——如门诊血液检测和肺功能测试——限制了持续监测患者状况的能力。相比之下，可穿戴设备和基于智能手机的健康平台现在支持实时、持续的生理数据采集。

可穿戴设备生成高维且常常噪声较大的时间序列数据，包括活动水平、心率和血糖。基于人工智能的技术被用于提取临床有意义的特征并建立所谓的“数字生物标志物”。在实际作中，心电图（ECG）贴片能够检测心房颤动，而智能手表则被用来识别心律失常和心房颤动。连续血糖监测（CGM）还支持预测低血糖或高血糖等关键事件，以及预测未来血糖水平。通过这种方式，可穿戴设备被广泛采用，能够收集与传统临床金标准相当的信号（如心电图、血糖），从而直接指导诊断或预后。

尽管可穿戴设备在过敏管理中的应用仍然有限，但新兴研究显示出一些令人鼓舞的进展。在哮喘护理中，关键挑战之一是准确检测喘息。一项最新研究表明，基于 CNN 的模型分析胸部加速度计数据，在检测喘息方面达到了 95%的准确率。此外，

研究人员还致力于通过机器学习技术预测通过配备传感器的吸入器收集的临床数据——如呼吸功能和吸入器使用模式。基于这些进展，AI Asthma Guard 集成了大型语言模型和传感器数据，提供个性化管理建议，结合严重程度特征提取、预测建模和患者个性化反馈。

随着这些应用的成熟，如何在过敏背景下建立数字生物标志物的问题变得越来越紧迫。一种方法是利用最初为某一疾病开发的治疗方法，用于不同的适应症。例如，低血糖期间心电图信号的变化已被用来结合 CNN 和 LSTM 检测夜间低血糖，而癫痫发作则通过心电图得出的心率变异性预测。同样，通过心电图或光电容积描记法识别呼吸速率有助于监测哮喘。为了将这些信号转化为稳健的临床生物标志物，开放资源和明确的验证框架都是必不可少的。

为了充分发挥可穿戴设备捕捉的生物信号潜力，基础模型预计将发挥关键作用。通过利用大规模的既有数据集，这些模型可以解决个体间的变异性，提高预测准确性。其优势包括减少再培训需求、保护隐私的本地运行，以及在多任务中实现中等计算成本的可重用性。例如，基于 Transformer 的基础模型 GluFormer 基于 CGM 数据训练，已成功预测血糖水平，展示了纵向生物信号建模的实用性。当收集到类似的连续过敏信号时，构建针对过敏疾病监测优化的基础模型可能变得可行。

尽管过敏性疾病常伴有器官特异性症状，但它们共同的免疫学基础表明存在统一的建模方法。可穿戴设备捕捉的持续生理信号因此可能成为新型数字生物标志物和针对过敏的基础模型的基础，最终实现更精准的预测和个性化疾病管理。

利用电子健康记录数据进行早期疾病检测、治疗决策和工作流程支持

电子健康记录（EHR）数据在医疗领域的应用显著推动了早期疾病检测、治疗决策和临床工作流程支持等领域的发展。在早期疾病检测中，电子健康记录数据通过预测分析实现主动监测，能够在症状出现前识别潜在疾病状态。例如，基于 EHR 记录训练的机器学习模型，可能通过追踪生命体征、实验室结果和处方的变化，检测出败血症、癌症或自身免疫疾病等疾病发展中的异常或模式。

在治疗决策中，电子健康记录（EHR）提供了患者的全面病史视图，帮助临床医生更有效地定制治疗方案。机器学习算法可以从电子健康记录中的历史数据中识别各种治疗与患者结局之间的相关性，通过基于个性化医疗原则建议最有效的干预措施，帮助决策过程。例如，在决定在癌症治疗中使用免疫疗法还是化疗时，模型可

能会预测哪种方案更可能为特定患者带来更好的结果。

在工作流程支持方面，电子健康记录通过自动提醒和通知促进护理流程的简化，帮助医疗提供者高效管理职责。这些系统可以标记潜在的药物相互作用或禁忌症，基于临床指南建议合适的护理路径，或根据 EHR 数据识别的患者风险因素提醒从业者所需的预防措施。例如，集成电子健康记录的机器学习模型可以预测肺炎球菌肺炎的高风险人群，并自动预约疫苗接种。此外，这些系统还可以通过根据历史趋势或从电子健康记录中大量数据集中推导出的预测分析，帮助资源分配。

总之，将电子健康记录数据整合进机器学习不仅提升了护理提供的速度和准确性，也使医疗专业人员能够更明智地做出早期发现和个性化治疗方案的决策。然而，隐私、数据质量和伦理考量等挑战必须被认真解决，以确保这些益处能有效且安全地为患者实现。

2.6 人工智能在过敏领域的当前应用

哮喘是人工智能应用研究最广泛的领域。现有研究已涵盖若干临床重要任务，包括诊断支持/分类，加重预测，及表型分析（表 2）。虽然电子健康记录（EHR）被广泛使用，但也采用了多种方式，包括肺功能测试，支气管挑战测试， CT 影像，以及吸入器中的传感器数据。常用的人工智能算法包括射频（RF），GB， NN，以及无监督聚类方法。在许多研究中，这些模型的 AUC 得分超过 0.8。哮喘研究数量较多，可能归因于其高患病率以及相对容易收集相关临床数据以进行诊断和严重程度评估。

表 2 临床数据分析中哮喘的代表性人工智能应用

任务	输入数据	人工智能算法	样本量	预测表现	研究设计	主要参考文献
诊断支持/分类	电子记录、肺功能测试、支气管挑战测试	LR、SVM、DNN	566	DNN 计分：0.98	前瞻性队列的再分析	Tomita K et al. (2019)
	音频	NN	585	PPA：0.97，NPA：0.91	前瞻性多中心	Porter P et al. (2019)
	电子健康记录数据	基于卷积神经网络的模型	14,697	AP：0.825	回顾性，单一中心	Yu G et al. (2021)
	甲酰胆碱挑战测试	射频、远程、虚拟监测、XGB、非恩	1501	射频 AUC：0.950，AUPRC：0.909	回顾性单一中心	Kang N et al. (2024)

任务	输入数据	人工智能算法	样本量	预测表现	研究设计	主要参考文献
	成像	特别空间管理	95	急性加值：0.80，F1：0.81（哮喘/慢性阻塞性肺病分化）	回顾性单一中心	Moslemi A et al. (2022)
	医疗管理数据库	多项回归、GB、RNN	178,962	GB 中 RNN F1：0.76	回顾性研究	Joumaa H et al. (2022)
加重的预测	电子健康记录数据	长距离、射频、轻型导弹	60,302	AUC：0.71（OCS 爆发）AUC：0.88（急诊）AUC：0.85（住院）	回顾性单一中心	Zein J et al. (2021)
	电子健康记录数据、社会标志物、病毒载量	DT、套索、射频、GB	29,392	GB AUC：0.84（住院）	回顾性多中心	
	吸入器传感器数据	GBTree	360	AUC：0.83	开放标签前瞻性研究	
	每日呼气峰值流量与症状评分	左侧、DT、NB、感知机	2010	LR AUC：0.85（3 天加重期）	RCT 数据分析	Patel SJ et al. (2018)
	电子健康记录数据（人口统计、合并症、用药等）	弹性网 LR、射频、GB	3057	英国 AUC：0.74	回顾性队列数据库	Lugogo NL et al. (2022)
	电子健康记录数据，社会标志物	射频，特别信号系统	3678	右外野评分：0.66（再接纳）	回顾性队列数据库	Zhang O et al. (2021)
表型/内型	临床、生理、炎症变量	聚类	378	6 个与严重程度/早发相关的群聚	前瞻性队列的回顾性分析	Inselman JW et al. (2023)
	临床、生理、炎症、人口统计变量	聚类	346	4 个伴有皮质类固醇反应的簇	前瞻性队列的回顾性分析	Shin EK et al. (2018)
	皮肤/IgE 检测	嗯，聚类	1053	5 级车型，“多重早期”运转：29.3	基于人群的出生队列研究	Wu W et al. (2014)
	电子监控设备	聚类	220	3 依附性群	RCT 数据分析	Wu W et al. (2019)

人工智能;电子健康记录（EHR）;逻辑推理，逻辑回归;SVM，支撑向量机;DNN，深度神经网络;准确性;神经网络 NN;PPA，正百分比协议;不良贷款，负百分比协议;卷积神经网络 CNN;AP，平均精度;右外野，随机森林;XGB，极限梯度增强;曲线面积 AUC;AUPRC，精度-回忆曲线下的面积;F1，F1 积分;GB，梯度增强;RNN，循环神经网络;LGBM，光梯度增强机;口服皮质类固醇 OCS;急诊科急诊部;拉索，LASSO 回归;DT，决策树;天真贝叶斯;嗯，隐马尔可夫模型;OR，赔率比;随机对照试验，随机对照试验。

相比之下，针对其他过敏性疾病的 AI 研究较少（见表 3）。对于特应性皮炎和湿疹，当前 AI 应用主要集中在表型分析和患者分层方面，还有更多研究探讨可穿戴传感器数据的应用。在针对食物和药物过敏的研究中，基于 EHR 的方法已被广泛采用。关于过敏性鼻炎和结膜炎的研究包括利用临床数据进行表型分析，基于全国调查的分析，以及基于影像的模型。这些应用预计将通过整合组学数据和生物信息学方法，以及传统临床数据分析而进一步发展。

表 3　非哮喘过敏性疾病临床数据分析中的代表性人工智能应用

疾病	任务	输入数据	人工智能算法	样本量	预测表现	研究设计	主要参考文献
特应性皮炎/湿疹	表型/亚型	纵向问卷和病历数据	贝叶斯	9801	8 个潜在类别，“特应性三月”特征：3.1%	基于人群的出生队列研究	Belgrave，DCM 等（2014）
	患者分层	临床数据	LR，英国	367	GBAUC：0.71（严重程度）	前瞻性队列横断面分析	MaintzL 等人（2021）
	湿疹严重程度的预测	自我报告的症状评分	贝叶斯	393	概率水平预测提升了 60%	对两个已发表队列的回顾性分析	HuraultG 等（2020）
	夜间抓挠行为的检测	可穿戴传感器数据（加速度计）	SVM、PCA、DT、RF	32	r=0.76（多导睡眠图）r=0.82（视频）	前瞻性	MahadevanN 等（2021）
食物/药物过敏	诊断与风险分层	电子健康记录数据	LR，RF	4077	射频 AUC：0.80	回顾性数据库分析	LandauT 等人（2024）
	预测	电子健康记录数据	LR、NB、SVM、K-NN、MLP、RF、CNN	219,902	CNNF1：0.956（过敏性休克预测）	回顾性，单一中心	Segura-BedmarI 等（2018）
		电子健康记录数据	LR，英国	4777	GBAUC：0.67（青霉素过敏预测）	回顾性、多中心研究	Gonzalez-Estrada 等人（2024）
过敏性鼻炎/结膜炎	表型/内型	病史、皮肤测试、鼻腔细胞学	LCA	168	两类：中性粒细胞和嗜酸性粒细胞优势	横断面研究	MaliziaV 等人（2022）
	自杀相关行为的预测	来自全国性调查的临床数据	射频	300,301	AUC：0.90	全国队列分析	LeeH 等（2024）
	影像学诊断	成像	U-Net++	452	检测睑板腺结构变化	前瞻性单中心	WeiJ 等（2024）

人工智能;逻辑推理，逻辑回归;GB，梯度增强;曲线面积 AUC;SVM，支撑向量机;PCA，主成分分析;DT，决策树;右外野，随机森林;电子健康记录（EHR）;k-NN，k-最近邻;MLP，多层感知器;卷积神经网络 CNN;F1，F1 积分;LCA，潜在类分析。

3. 未来方向

3.1 数学与人工智能分析技术的发展

数学和人工智能分析方法的开发与维护是从临床数据中阐明生物现象和预测疾病诊断的关键基础。尽管基于状态空间模型的数据同化方法等时间序列预测模型已被用于短期预测事件，但预测更远的结果仍然具有挑战性。由于慢性疾病的发作和加重往往在初始数据收集后数月至数年内发生，因此采用能够考虑长期影响的分析技术至关重要。

一个显著的例子是生存分析，这是一种统计方法，用于分析和解释从数据获取到死亡等终点的时间。近年来，机器学习越来越多地与生存分析相结合，基于多种预测因素预测个体的疾病发病或复发情况。

随机生存森林（RSF）是一个结合机器学习与生存分析的框架。RSF 以生存计数为响应变量拟合多重决策树，从而实现从多种协变量中预测事件。与标准随机森林一样，RSF 能够评估多个因素的联合影响，并处理多种变量类型。在一项利用 RSF 预测早期乳腺癌手术干预后转移复发的研究中，结合包括 Ki67 表达和表皮生长因子受体表达在内的五个因子，预测准确率优于 Cox 回归。

Dynamic-DeepHit 是一种将深度学习与生存分析结合的算法，从而从纵向测量生成动态预后模型。与假设特定概率分布的传统方法不同，Dynamic-DeepHit 直接学习不同风险因素与疾病发生时间之间的关系。在约 6000 名囊性纤维化患者的纵向数据集中，它准确预测了个体的病情加重，并识别出不同竞争风险的关键因素。

此外，物理学和复杂系统科学中的变点检测方法也被应用于开发数学框架，用于识别不可逆变化（如疾病发作或恶化）的早期预警信号。进一步发展这一观点，动态网络生物标志物（DNB）理论提供了一种机制，用于检测临界点附近的预警信号，因为该点的状态可能迅速变化，难以恢复基线。DNB 将多个生物标志物或临床参数之间的相互作用建模为网络，并利用网络波动作为早期预警信号。它已被用于多种疾病，包括急性肺损伤、乙型肝炎相关肝癌、B 细胞淋巴瘤和 1 型糖尿病。

3.2 利用 LLM 进行多模态临床数据：未来与 LLM 的基础模型

在多模态临床数据分析中利用大型语言模型（LLMs）为进步提供了多条途径。随着我们朝向更全面的医疗方法发展，LLM 可以成为跨多样数据集集成的基石。LLM 可以分析多模态临床数据（如基因组学、影像学和患者报告结果），以优化诊断准确性。通过同时处理这些复杂数据集，模型可以捕捉不同类型数据之间的相互依赖关系，从而更全面地理解疾病状态。LLM 可以提供个性化的治疗建议。将临床记录、实验室结果、影像研究和可穿戴设备数据与大型语言模型整合，可以实现针对个别患者需求的动态治疗方案，不仅考虑当前状态，还考虑过去的健康轨迹。

通过在纵向多模态数据集（包括电子健康记录、基因组序列和生活方式数据）上预训练 LLMs，并随后在特定临床预测任务上进行微调，这些模型相比传统方法能够实现更高的疾病预报准确率。将环境暴露、遗传倾向、患者行为模式和实时监测数据等多种因素纳入由大型语言模型训练的机器学习模型，可以为慢性疾病或疾病（如过敏、癌症）提供更准确的风险评估。或者神经系统疾病。

一个跨多种模态大规模训练的基础模型，可以为不同数据类型之间的集成建立共同基础。这种标准化使未来模型能够利用来自类似数据集的预训练知识，而无需大量重新训练。例如，在训练时提取的特征或在某种类型的数据（如基因组学）上学到的权重，可以作为分析不同模态（如成像数据）的有力起点。通过重复使用组件获得的效率减少了所需的计算资源，加快了模型开发周期。这在临床数据集获取成本高昂或需要特定领域专业知识的情况下尤为有利。

挑战包括确保多样化数据源——每个数据源都有其独特格式和质量——被正确格式化和表示，需要细致的预处理和特征工程，这可能复杂且资源密集。尽管可解释人工智能（XAI）取得了进步，LLM 在理解其多模态数据处理方式时仍是一个“黑箱”。在保持预测能力的同时提升可解释性仍是一个持续的挑战。最后，处理敏感患者数据需要严格的隐私和安全措施，同时还要权衡使用大型数据集训练 LLM 的优势。

总之，大语言模型在多模态临床数据中的整合，有望带来精准医疗的变革性进展。随着技术成熟和伦理框架的建立以支持其使用，我们可以期待诊断准确性、个性化治疗计划、预测分析和风险评估模型的显著提升。这些发展凸显了大型语言模型推

动更全面医疗系统的潜力，该系统能够考虑患者健康和疾病进展的多个方面。

3.3 数字孪生

数字孪生日益被视为精准医疗领域的变革性概念。最初在航空航天行业开发，它指的是创建物理物体的虚拟复制品，实时模拟其行为。该概念由格里夫于 2002 年提出，包含三个要素：物理实体、其数字对应物以及它们之间的数据链路。在制造业中，数字孪生被广泛用于生命周期管理，通过集成包括时间序列信号在内的多种数据流。得益于高性能计算和大数据基础设施的进步，数字孪生现在在临床环境中越来越可行。

在医学中，数字孪生代表了一个个体患者的虚拟模型，能够基于实时和历史的临床数据模拟未来的生理状态（见图 4）。为了有效，这样的模型必须准确复现个人的生理状况，并支持动态、个性化的预测。虽然建模整个人体在技术上仍具挑战性，但针对器官或症状的数字孪生正逐渐变得可行。早期的例子包括基于 CNN 的“心脏孪生”系统，用于分类缺血性与非缺血性心脏疾病，心脏功能模拟，并根据呼吸机数据模拟呼吸功能。可穿戴设备的日益普及使得高分辨率生物信号的持续获取成为可能，从而能够构建针对特定临床任务（如疾病监测、分类或干预计划）量身定制的虚拟模型。

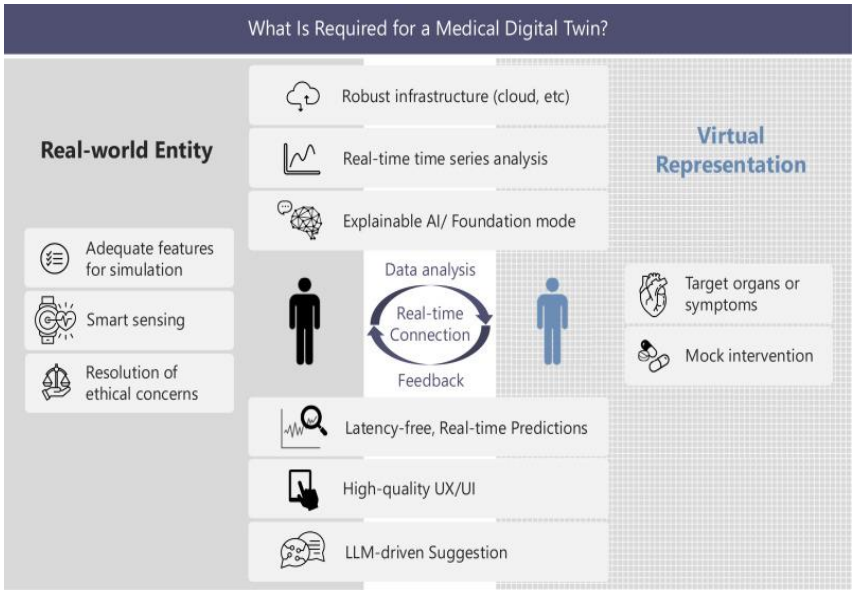

图 4 构建医疗数字孪生的关键组件

构建医疗数字孪生的关键要素示例，围绕三大主要领域构建：现实世界实体;虚

拟表示;以及他们之间的实时连接。人工智能、人工智能;用户体验/界面、用户体验/用户界面;大型语言模型（LLM）。

展望未来，数字孪生在临床环境中的实际应用必须满足三个关键条件：（1）提升单个模型的准确性并建立多模态数据集成平台;（2）结合因果推理、反馈学习和动态适应性;以及（3）提供高度可解释的接口。

首先，提升模型准确性需要采用能够识别关键模拟相关特征并获取全面替代人类健康和疾病状态的智能传感技术。此外，建立一个集成分析平台，通过结合生物信号、组学数据、医学图像和可穿戴设备的电子健康记录等多样化数据源，持续学习和更新个性化模型也至关重要。基于大规模多模态数据集训练的基础模型将成为支持这一集成平台的关键组成部分。

其次，整合因果推断和动态适应性至关重要。数字孪生应在实际应用前模拟虚拟干预（如治疗变更）。引入动态因果模型以估算干预效果并迭代评估虚拟情景至关重要。此外，临床状况可能迅速变化，如过敏性休克等急性事件，生命体征波动剧烈。因此，需要高度的实时仿真适应性。应对这些挑战需要临床医生、工程师和数据科学家之间的协作，并依靠确保伦理、治理和方法论一致性的团队结构支持。

最后，提供直观且可解释的界面至关重要。预测结果和模拟干预结果应通过用户友好的系统向临床医生和患者传达，这些系统由可解释人工智能（XAI）框架和大型语言模型（LLMs）支持。这些界面增强了反馈循环和共享决策，促进了信任和临床采用。

最终，随着这些系统的演进，数字孪生有望从器官专用工具发展为全面的终身模型，指导个性化的预防护理和疾病管理，贯穿个人整个生命周期。

临床医生与人工智能共存的策略：我们应如何应对人工智能？

近年来，越来越多的研究探讨了人工智能在临床数据分析中的应用。然而，尽管取得了这些进步，人工智能在实际临床实践中的整合仍然有限。随着对准确性和临床实用性需求的不断提升，临床医生如何有效参与人工智能模型的开发、验证、实施和应用，这一关键挑战最终确保这些技术为患者带来切实益处。

随着近期技术的进步，临床医生现在拥有比以往更多的机会参与人工智能开发，因为高性能计算的普及和大型语言模型的进步显著降低了技术壁垒。然而，一个普遍的担忧依然存在：担心人工智能将取代医生，导致一些临床医生对人工智能相关

项目产生疏远。然而，通过直接利用现有临床数据创建和验证人工智能模型，医生可以更深入地理解人工智能的能力和局限性。这种亲身体验强化了一个关键认识：临床医生不再将人工智能视为威胁，而是将其视为通过数据驱动洞察提升临床专业知识的决策支持工具。

阻碍人工智能在临床应用中的关键因素是人工智能模型开发与现实临床实用之间的差距。作为领域专家，临床医生最有能力识别人工智能应解决的具体问题，并必须在引导人工智能发展中发挥重要作用。除了模型准确性外，临床人工智能应用还需要严格的验证，以确保可解释性、可推广性以及与多模态临床数据的整合。建立健全的临床试验框架对于系统评估人工智能对临床结果的影响以及确定其在现实环境中的真实效果至关重要。此外，AI 开发应优先考虑用户体验和实用性，以确保与临床工作流程无缝整合。低可用性和繁琐的流程常常阻碍了人工智能的实施。为了克服这些障碍，临床医生与人工智能研究人员之间的积极合作从模型设计的最初阶段就至关重要。

通过明确人工智能在临床实践中的适当角色并积极监督其部署，临床医生可以在保持其在医疗决策中核心作用的同时，充分发挥其潜力。临床医生与人工智能之间均衡的合作关系将为更先进、个性化和以患者为中心的医疗服务铺平道路。

3.4 伦理问题与患者意识

随着人工智能越来越多地融入临床数据分析，必须关注其伦理问题。一项 2021 年的研究表明，患者对人工智能的使用存在重大担忧，包括安全、数据隐私和安全。对此，诸如三脚架+人工智能等指导方针。这些都旨在促进基于人工智能模型的透明度。在过敏领域，2022 年提出了负责任使用的人工智能框架，随后于 2023 年发布了环保意识指南。这些举措共同旨在突出潜在风险，并确保人工智能技术的伦理部署。

尽管现有指南解决了许多已知问题，但新兴技术如基金会模型和数字孪生带来了前所未有的伦理挑战，可能超越现有框架。基础模型通常需要庞大的计算能力，并受益于大规模数据集。这可能加大资金充足的企业与学术或临床研究机构之间的差距，导致对私营实体的过度依赖，提供必要的公共卫生基础设施。此外，如果数据收集和模型开发变得地理集中，可能会引入偏见，使特定人群或地区处于不利地位。

数字孪生技术也提出了关于个体自主权的重要问题。随着这些模型逐步模拟一个人从出生到死亡的健康轨迹，患者可能会觉得必须依赖于临床症状出现前就已做出的预测。在这样的未来，围绕人工智能生成预测的必然感可能会侵犯个人的自决权和独立决策权。

3.5 迈向针对过敏的基础模型工作流程及其实际挑战

针对过敏的基础模型工作流程开发包含若干关键且连续的步骤，每个步骤都伴随着实际挑战，并需要多方利益相关者的参与（见图 5）。第一步是建立多机构基础设施，用于获取多模态临床数据。实现这一目标需要机构领导和一线医疗专业人员的号召，拥抱新技术，促进各地点的标准化数据收集。第二步涉及安全的数据共享、存储和维护。这里的主要障碍包括跨机构的通用性问题以及由隐私和监管问题驱动的数据共享限制。为克服这些问题，政策制定者必须制定灵活的立法框架、伦理指南和数据治理政策，理想情况下应与临床医生和技术人员密切合作，以确保可行性和合规性。第三步是构建基础模型本身，这最好通过学术与产业的合作来实现。最后，单个基础模型的每个下游任务结果应无缝集成到现有的电子健康记录系统中。这一最后阶段需要电子健康记录供应商的积极配合以及临床医生的反复反馈。这一概念性工作流程共同确保了 AI 输出具有临床解释性、可作性，并与现实环境保持一致。

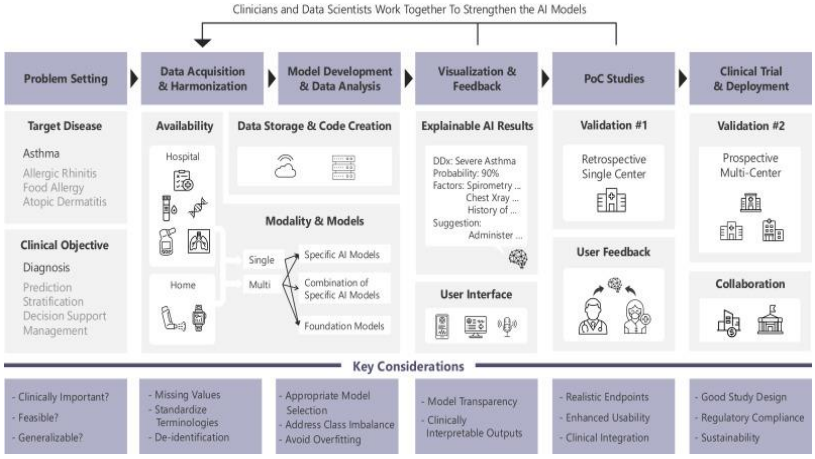

图 5 以哮喘为例的过敏特异性 AI 模型的概念性工作流程

开发和实施针对过敏的人工智能（AI）模型的概念性工作流程示意图，哮喘为代表案例。该工作流程概述了从问题设置、数据采集与协调、模型开发与分析、可视化与反馈、概念验证研究、临床试验到实际部署的关键阶段。贯穿全程强调临床医

生与数据科学家之间的协作，以确保临床相关性和可用性。在后期阶段，与行业合作伙伴和监管机构的紧密合作对于成功实施至关重要。XAI，可解释人工智能;DDx，鉴别诊断。

4. 结论

随着计算能力的显著进步和临床数据的广泛可用性，人工智能从根本上改变临床医学的时代已经到来。在本综述中，我们概述了迄今为止如何利用临床医生常用的临床数据进行人工智能研究。我们还提供了适用于具体数据集的实用指导。此外，我们还考察了人工智能研究的新兴趋势，讨论了未来面临的关键挑战，包括伦理问题以及基础模型和数字孪生等先进技术的发展。