超越P值：将量化金融的风险与遗憾度量引入学习型卫生系统中的人工智能

Richik Chakraborty，理学硕士

乔治城大学

电子邮件：rc1586@georgetown.edu

关键词：医疗保健中的人工智能、学习型卫生系统、量化金融、风险度量、校准漂移、遗憾最小化、风险价值、条件风险价值、FDA监管、医疗人工智能安全、自适应算法、部署后监测

摘要

人工智能 (人工智能) 在临床环境中的部署日益增多，这对传统医学证据框架所依赖的基本假设提出了挑战。以随机对照试验、频率学派假设检验和静态置信区间为核心的经典统计方法，原本是为在稳定条件下评估固定干预措施而设计的。相比之下，人工智能驱动的临床系统会持续学习，随着时间推移调整自身行为，并在由不断变化的人群、实践和反馈效应所塑造的非平稳环境中运行。在这类系统中，临床伤害与其说源于平均错误率，不如说源于校准漂移、罕见但严重的故障，以及次优决策随时间累积。

在本文的观点中，我们认为，现行的统计显著性概念不足以表征学习型卫生系统中的证据与安全性。借鉴量化金融和在线决策理论中的风险理论概念，我们提出，应以时间索引校准稳定性、有界下行风险和受控累积遗憾为依据，重新构建针对自适应人工智能系统的医学证据。我们强调，这种方法并非要取代随机试验或因果推断，而是通过处理只有在部署后才会显现的风险与不确定性维度，对它们加以补充。该框架为在持续学习条件下评估人工智能驱动的临床系统提供了严谨的数学语言，并对临床实践、研究设计和监管监督具有启示意义。

1. 引言

人工智能系统日益融入临床决策，支持从诊断解读和预后判断，到治疗建议和资源分配等各类任务。与传统医疗干预不同，许多此类系统旨在部署后从新数据中学习，并随着临床环境的变化更新其内部表征。这些系统通常嵌入学习型卫生系统中，在那里，常规医疗过程中产生的数据会被反馈用于算法改进。

尽管这一范式有望提供更具响应性和个性化的护理，但它也对既有的医学证据框架提出了根本性挑战。随机对照试验和传统的

统计推断假设，接受评估的干预措施保持不变，且数据生成过程足够稳定，能够支持基于历史样本的推断。当这些假设应用于人工智能系统时，其可靠性正日益降低，因为这类系统的行为会随时间变化，且其部署本身会改变临床环境。

当前评估临床人工智能的方法，往往是在传统验证范式的基础上加以扩展，在单一时间点报告区分度指标、校准曲线和统计显著性。然而，这类基于快照的评估难以深入揭示人工智能系统在分布偏移下的行为、随着持续学习不确定性如何变化，以及误差如何在反复决策中累积。因此，这些方法未能捕捉临床风险的关键维度。

本文认为，在自适应人工智能的背景下，医学证据这一概念本身需要重新审视。具体而言，我们提出一种风险理论重构，将证据视为对校准、下行风险和随时间累积的遗憾所作的动态评估。这一视角借鉴了量化金融和在线学习的思想；这些领域长期以来一直在应对非平稳性、反馈循环，以及虽罕见但可能造成灾难性后果的失效。

2. 自适应系统中经典临床统计学的局限性

医学中的经典统计推断建立在一系列假设之上，这些假设适用于评估静态干预措施，却与自适应人工智能系统并不相称。其中有三个假设尤为重要：数据生成过程的平稳性、固定估计目标，以及一次性验证。

平稳性意味着预测变量与结局之间的关系会随时间保持稳定。然而，在真实的临床环境中，患者群体会不断变化，疾病患病率会发生波动，临床实践会改变，也会有新的治疗方法问世。持续学习的人工智能系统可能会通过改变临床医生的行为来放大这些动态变化，

以响应这些系统的输出，从而形成进一步破坏平稳性的反馈循环。

固定估计目标假定，所关注的量（例如治疗效应或预测性能指标）在整个评估期间都具有明确的定义。在自适应系统中，随着模型更新、患者群体变化或临床任务重新界定，估计目标本身也可能发生变化。在这种情况下，解读单个置信区间或 p值在概念上会变得棘手。

一次性验证假定，部署前评估的性能能够代表未来的表现。对于持续学习的人工智能系统而言，这一假定从设计上就不成立。部署后的性能可能与初始基准大相径庭，尤其是在分布偏移或概念漂移的情况下。

这些局限并非只是技术层面的。它们对患者安全有直接影响。部署时达到统计显著性，并不能保证人工智能系统在条件变化时仍保持可靠，也无法量化罕见但严重的故障所造成的潜在伤害，或多次决策中小误差累积带来的潜在伤害。

3. 证据作为一个随时间变化的过程

在自适应临床人工智能系统中，结局用随机过程来表示，比用独立样本更为自然。预测、决策和结局按时间顺序逐步展开，每项行动

都会影响未来的数据和模型更新。从这一视角来看，证据并非模型的一项静态属性，而是对其行为随时间演变的刻画。

在单一时间点评估证据，会掩盖时间依赖性和路径依赖风险。例如，某个模型的平均表现可能良好，但在分布偏移后，仍可能出现持续一段时间的校准失准。这类情况即使在整体性能指标仍处于可接受范围时，也可能导致患者伤害集中发生。

证据的这一时间维度要求我们将关注重点从点估计转向变化轨迹。指标必须能够捕捉不确定性、误差和风险如何随着持续使用而演变。这一认识促使我们寻找能够动态而非静态地呈现证据的替代数学框架。

4. 校准漂移：一项核心临床风险

在模型性能的诸多维度中，校准在临床决策中占据核心地位。经过良好校准的预测有助于临床医生正确解读概率，从而为检测、治疗和监测相关决策提供依据。反之，校准失准会损害信任，并可能导致系统性治疗过度或不足。

重要的是，校准并非模型的一项固定属性。实证研究表明，患者群体、疾病患病率、测量实践或临床工作流程的变化，都可能导致校准随时间推移而下降 (Davis 等，2019年；Davis 等，2020年；Guo 等，2021年；Nestor 等人，2019)。在学习型卫生系统中，临床医生会根据算法输出调整自身行为，由此产生的反馈效应可能进一步加剧这种漂移。

我们认为，随时间变化的校准误差应被视为临床风险的首要度量单位。与其报告单条校准曲线或一个汇总统计量，不如通过证据描述校准如何演变，识别不稳定时期，并量化其持续时间和严重程度。

Box 1：时间索引校准误差

令 $\hat{P}t(Y=1|X)$ 表示模型在时间 *t* 的预测概率。时间索引校准误差可定义为：

$$\text{期望校准误差}(t) = E[|\hat{P}t(Y=1|X) - P(Y=1|X,t)|]$$

随时间监测期望校准误差(t)，即使区分度保持稳定，也能检测到校准漂移。

这可以类比于金融市场的波动性：相对平静的时期与不稳定性加剧的阶段交替出现 (Bollerslev, 1986)。尽管健康数据与金融时间序列存在根本差异，这种概念上的相似性凸显了随时间监测稳定性的重要性，而非假定历史表现会持续不变。

---

5. 临床示例详解

为使这些概念立足于临床现实，我们介绍三类问题，在这些问题中，风险理论评估揭示了传统指标无法发现的失效。

示例 1：脓毒症预测模型中的校准漂移

以 COVID-19 疫情前部署在医院病房的脓毒症预警系统为例。许多此类模型在部署时表现出可接受的区分度和校准度。然而，疫情期间，患者构成、呼吸衰竭患病率、实验室检测模式以及临床工作流程的变化，显著改变了数据生成过程。回顾性分析表明，数个已部署的脓毒症模型在此期间出现明显的校准漂移，系统性地低估了具有非典型炎症表现患者的风险 (Davis 等，2020；Finlayson 等人，2025)。

重要的是，标准的部署前验证指标并未揭示这一失效模式。平均曲线下面积值保持相对稳定，掩盖了这样一个事实：预测概率已不再对应观察到的事件发生率。从临床角度看，伤害并非均匀分布：在表现超出疫情前训练分布的患者中，延迟升级处置的情况尤为常见 (Nestor 等人，2019)。

在时间索引证据框架下，这种失效会表现为校准误差随时间持续上升，从而触发加强人工监督或进行模型再训练，以免伤害广泛累积。

示例 2：重症监护病房死亡率预测中的尾部风险

重症监护病房死亡率预测模型通常平均表现优异，但在罕见临床情境下可能遭遇灾难性失效，例如患者患有罕见合并症或生理变化轨迹异常 (Ovadia 等人，2019)。在这些情况下，少数预测可能出现极端误差，预测死亡概率与观察到的结局大幅偏离。

尽管这些事件可能不到病例的 5%，但其临床后果十分严重，会影响升级护理、治疗目标讨论和资源分配等决策。传统评估指标会对这些失误取平均值，实际上将其视为统计噪声。

尾部风险视角则明确关注最坏情况下的误差分布。对最严重预测失误所造成的预期伤害设定上限，更符合临床安全优先事项；在这些情境中，预防灾难性错误比小幅提升平均准确率更为重要。

示例 3：肿瘤学决策支持中的累积遗憾

设想一个临床决策支持系统，它会针对惰性恶性肿瘤，建议采取监测还是早期干预。即使该系统的平均判断正确，反复给出保守建议，也可能导致部分病情意外进展的患者延迟启动治疗。

由此产生的伤害会不断累积：每一次延迟决策都会加重未来风险，导致确诊时分期更差、治疗选择减少。以单个决策点为评估对象的标准性能指标，无法捕捉这种累积效应 (Subbaswamy 和 Saria, 2019)。

基于遗憾的评估将这种伤害重新界定为：在模型建议下实现的结局与事后看来最优策略本可带来的结局之间的累积差异。这一视角凸显了基于快照的验证无法揭示的长期机会成本。

6. 尾部风险与临床伤害的集中性

平均性能指标掩盖了临床伤害往往集中于罕见但严重事件这一事实。少数灾难性错误可能主导患者结局，尤其是在重症监护、肿瘤学或急诊医学等高风险场景中。

量化金融中发展的风险理论度量，为刻画这类现象提供了一套术语 (Artzner等人，1999；Rockafellar与Uryasev，2000)。风险价值描述了一个阈值，损失一旦超过该阈值便不可接受；条件风险价值则衡量损失超过该阈值后的预期严重程度。将这些概念转用于临床语境，它们对应于对最坏情况下的预测误差或决策失败设定界限，并量化不利情景下的预期伤害。

框 2：下行风险与临床条件风险价值

令*L*表示具有临床意义的损失函数（例如，干预延迟伤害）。水平为α 时的条件风险价值定义如下：

$$\mathbf{CVaR(L) = E[L \mid L \geq VaR]}$$

在临床环境中，CVaR0.95代表预测失误中最严重的5%所造成的预期伤害，使评估与安全关键结局保持一致。

我们强调，目标并非将金融模型原封不动地引入医学，而是采用一种明确承认并量化下行风险的框架。这样做更符合临床对安全性的直觉：与平均表现的微小提升相比，临床更优先避免罕见但后果严重的结局。

定量示例

为说明这些概念，请看一个模拟的风险预测模型：该模型部署了12个月，其结局发生率逐渐发生变化。部署期间，曲线下面积始终稳定在约0.83；但第4个月之后，校准误差持续上升，从0.02增至0.12。与此同时，CVaR0.95从0.08升至0.28，表明尾部风险不断增加（见图1）。

部署时进行基于p值的验证，会因该系统的区分度达到可接受水平而批准其使用。然而，持续监测ECE(t)和CV

aR0.95 会在第6个月将模型标记为不安全，此时两项指标均超过预先设定的安全阈值。这个简化示例展示了风险理论指标如何揭示传统评估未能发现、但与临床相关的失效；这些模式与分布偏移期间已部署的脓毒症预测系统中观察到的性能退化相一致 (Davis 等人，2020)。



图 ***1:*** 模拟部署中风险度量的时间演变。***(A)*** 曲线下面积在整个部署期间保持稳定，掩盖了潜在的失效。***(B)*** 期望校准误差 ***(ECE)*** 在第***4***个月后单调上升。***(C)*** 条件风险价值 ***(CVaR0.95)*** 显示尾部风险不断增加，并在第***6***个月超过安全阈值。红色虚线表示预先设定的安全阈值。部署时的传统验证会批准该系统，而持续风险监测会在第***6***个月将其标记为不安全。

7. 遗憾作为累积的临床机会成本

在自适应决策中，次优行动的后果会随时间累积。遗憾是在线学习的核心概念（(Lattimore & Szepesvári, 2020年；Bubeck & Cesa-Bianchi, 2012)），它将这种累积形式化为：特定策略所取得的结局，与事后看来采用最优策略本可取得的结局之间的差异。

在医学中，遗憾对应于有效疗法的延迟启动、发病时间延长、可避免的并发症，或错失预防机会。这些伤害在横断面分析中可能并不明显，但在纵向评估决策时便会显现。

专栏3：临床决策中的累积遗憾

在时间范围 *T* 内，遗憾定义为：

$$R(T) = \Sigma_{t=1}^{T} [l(a_t) - l(a_t)]^*$$

其中，a_t 表示所采取的行动，而 a_t^* 表示事后看来最优的行动。安全约束学习旨在限制暴露于不可接受的风险，同时约束 R(T) 。

传统统计框架缺乏量化这类累积效应的机制。相比之下，基于遗憾的度量明确考虑了不确定性下反复决策的长期后果 (Sutton & Barto, 2018)。将遗憾纳入证据的定义，会使关注点从孤立的正确性转向持续的临床表现。

对遗憾最小化会助长不道德实验的担忧是合理的，必须加以应对。在临床情境中，探索必须受到非对称损失函数、伦理考量和安全要求的约束 (Auer et al., 2002)。近期关于保守型学习和风险感知型学习的研究，为调和遗憾最小化与临床伦理提供了潜在路径。

8. 医学证据的风险理论定义

结合这些脉络，我们提出一种适用于人工智能驱动的临床系统的医学证据风险理论定义。在此框架下，证据由四个相互关联的组成部分构成：

1. **后验信念更新**，反映动态不确定性
2. **随时间推移的校准稳定性**，确保概率可靠性
3. **有界下行风险**，限制最坏情况下的患者伤害
4. **受控累积遗憾**，约束长期机会成本

这一定义保留了概率推断的优势，同时纳入了仅在自适应的真实世界部署中才凸显的风险维度。它拓展了证据视角，不再局限于部署前验证，从而补充了随机试验和因果推断。

表 1 将这一框架与传统方法进行比较，凸显了它在捕捉时间动态和集中性伤害方面的独特能力。

表 *1*：临床人工智能证据框架比较

框架	主要指标	时间性的认知	尾部风险	累积 Harm	部署阶段
随机对照试验 + p值	平均处理效应	None	隐含在置信区间中	未测量	Pre-部署 only
AUC/Brier 评分	区分度/校准	快照	平均值	未测量	Pre-部署 + 验证
贝叶斯更新	后验概率	连续的	通过后验	未明确	可以是连续的
提议：风险理论的	**ECE(t)、CVaR、遗憾**	**明确的**	**有界的**	**受监控的**	**连续的 部署后 部署**

9. 对实践、研究和监管的启示

临床实践

这一观点提出，风险的静态验证转向连续根据风险阈值动态调整，在不放弃自动化的同时提升安全性。当 ECE(t) 超出预设范围，或条件风险价值表明尾部

风险升高时，临床工作流程可升级处理，要求增加人工审核，或暂时暂停算法建议。

与风险阈值相匹配的动态监督机制，为不加批判地实现自动化与全盘拒绝人工智能辅助之间提供了一条中间路径。将风险视为连续变量而非二元状态，临床医生即可在受益于算法支持的同时，保持适当的态势感知。

研究

该框架推动了自适应试验和部署后研究采用新的终点，在传统疗效指标的基础上纳入遗憾和尾部风险 (Lu et al., 2019)。比较评估应优先关注稳定性和安全性，而非区分度的边际提升。开发自适应人工智能系统的研究人员不仅应报告部署时的曲线下面积和校准情况，还应报告在现实部署情景下校准漂移和累积遗憾的预计界限。

此外，研究基础设施必须不断发展，以支持纵向评估。具有持续随访的前瞻性队列对于描述时间风险特征至关重要，这要求机构承诺持续投入，超出通常的验证研究时间线。

监管

近期监管框架已认识到这些挑战。美国食品药品监督管理局的预定变更控制计划 (PCCP) 和良好机器学习规范 (GMLP) 指南承认，自适应人工智能系统需要在初始批准之后继续接受监督 (US FDA, 2021, 2023; IMDRF, 2021)。然而，现行指南尚未明确如何随着时间推移量化安全性和有效性 (Vokinger et al., 2021)。

风险理论证据框架提供了操作性指标，例如校准稳定性、下行风险界限和累积遗憾，这些指标可直接与监管机构对持续性能监测的要求相衔接。监管机构可以不再采用一次性批准，而是要求持续证明校准、下行风险和遗憾始终处于可接受范围内。

这种**便捷**称监督成为可能：风险界限更严格且已证明具有稳定性的系统，可能无需如此频繁地审查；而接近安全阈值运行的系统则需要强化监测。FUTUREAI国际共识指南同样强调了持续监测框架的必要性，但尚未明确规定风险理论指标 (FUTUREAI联盟, 2025)。

通过为“可接受风险”提供定量基准，该框架为监管机构提供了具体的决策标准，以平衡创新与患者安全——这一直是临床人工智能快速发展领域面临的持续挑战。

10. 局限性与未来方向

这一框架面临着诸多重要挑战，包括稀疏且经删失的临床数据、实时风险估计方面的计算约束，以及风险在不同人群之间分布不均所引发的公平性问题。

将金融风险概念拓展至临床环境并非毫无争议。医疗保健数据与金融时间序列具有根本不同的特性：结局稀疏且经过删失，

并具有因果结构；人群具有异质性；伦理约束也限制了可进行的探索 (Ashukha et al., 2020)。这些差异要求谨慎调整方法，而非直接照搬。

此外，持续风险监测所带来的计算开销，可能会给资源受限的环境造成实际障碍。实时计算时间索引校准误差和尾部风险，需要流式数据基础设施和稳健的统计流程，而这些条件并非所有临床环境都具备。边缘计算和高效近似算法或许有助于克服这些障碍，但仍有待进一步的方法学开发。

最后，如果不同患者亚组之间的风险分布不均，基于风险的监测可能会无意中加剧差异，对历史上处于边缘地位的人群不成比例地启动强化监督 (Zink & Rose, 2020)。要解决这些公平性问题，就必须在风险理论框架中明确纳入公平性约束，确保安全机制不会沦为偏见的工具。

对这一框架在真实世界部署中的实证验证至关重要。通过前瞻性研究，在多个临床领域和机构环境中比较传统验证与风险理论监测，将为支持或质疑这些提议提供关键证据。此类研究应明确衡量的不仅是技术性能，还包括临床医生信任、工作流程整合和健康公平结果。

应对这些挑战需要方法创新、计算资源投入，并仔细考量伦理影响。尽管如此，明确问题空间是迈向更适合自适应临床人工智能时代的证据标准的必要第一步。

11. 结论

随着人工智能系统日益成为临床护理不可或缺的一部分，我们必须重新审视何为医学证据。在学习型卫生系统中，证据不再是对疗效性的静态断言，而是对不确定性、风险和遗憾的动态评估。若没有能够捕捉这些动态变化的风险理论框架，人工智能的发展将继续快于用于规范其部署的数学工具。以这种方式重新界定证据，对于使创新与患者安全相协调至关重要。

脓毒症预测漂移、重症监护病房尾部风险和肿瘤学遗憾累积的案例表明，这些并非假想的担忧，而是已部署系统中可观察到的失效模式。随着医学日益依赖实时学习的系统，问题不再是某项干预措施是否“有效”，而是在条件变化时，其不确定性、风险和遗憾是否仍处于临床可接受范围。

这一观点并不否定随机试验、因果推断或传统统计学，因为这些仍然是医学证据的基础。相反，它主张扩展证据工具箱，以应对在非平稳性条件下运行的自适应系统所特有的挑战。通过借鉴那些在管理时间风险方面拥有更丰富经验的领域中的概念框架，临床人工智能可以建立更成熟的安全与监督方法。

未来的发展之路需要临床医学、机器学习、定量风险分析、生物伦理学和监管科学领域的通力合作。唯有通过这样的跨学科对话，我们才能制定出与所部署系统相适应的证据标准。

致谢

作者承认使用了 Anthropic 的 Claude 协助构建数学结构，并使用 Grammarly 进行拼写检查。所有智识贡献、解读和结论均完全归作者所有。本研究未获得外部资助。

利益冲突

作者声明不存在利益冲突。

参考文献

1. Artzner P、Delbaen F、Eber JM、Heath D。一致风险度量。数学金融。1999年；9(3):203-228。doi:10.1111/1467-9965.00068
2. Ashukha A、Lyzhov A、Molchanov D、Vetrov D。深度学习中域内不确定性估计和集成的陷阱。arXiv。预印本于2020年2月28日在线发布。arXiv:2002.06470v2
3. Auer P, Cesa-Bianchi N, Fischer P. 多臂老虎机问题的有限时间分析。机器学习。2002年；47:235-256。doi:10.1023/A:1013689704352
4. Bollerslev T. 广义自回归条件异方差。计量经济学杂志。1986；31(3):307-327。doi:10.1016/0304-4076(86)90063-1
5. Bubeck S, Cesa-Bianchi N. 随机与非随机多臂老虎机问题的遗憾分析。机器学习前沿与趋势。2012；5(1):1-122。doi:10.1561/2200000024
6. Davis SE, Greevy RA Jr, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. 一种用于纠正临床预测模型漂移的非参数更新方法。美国医学信息学会杂志。2019年；26(12):1448-1457。doi:10.1093/jamia/ocz127
7. Davis SE, Greevy RA Jr, Lasko TA, Walsh CG, Matheny ME. 检测临床预测模型中的校准漂移，以指导模型更新。生物医学信息学杂志。2020年；112:103611。doi:10.1016/j.jbi.2020.103611
8. Finlayson SG, Beam AL, Subbaswamy A. 利用机器学习近似医疗数据中的反事实。arXiv。预印本于2019年12月18日在线发布。arXiv:1912.08544v1
9. Finlayson SG、Subbaswamy A、McDermott MBA等。检测并修正有害数据漂移，以负责任地部署临床人工智能模型。*JAMA Netw Open*。2025年；8(6)：e2513872。doi:10.1001/jamanetworkopen.2025.13872
10. FUTURE-AI联盟。FUTURE-AI联盟共识指南：保健中值得信赖且可部署的人工智能。英国医学杂志。2025年；388:e081554。doi:10.1136/bmj-2024-081554
11. Guo LL、Pfohl SR、Fries J等。临床医学中存在时间数据集漂移时保持机器学习性能的方法系统综述。*Appl Clin Inform*。2021年；12(4):808-815。doi:10.1055/s-0041-1735183
12. 国际医疗器械监管机构论坛。医疗器械开发良好机器学习实践：指导原则。发布于2021年10月。更新于2025年1月29日。访问日期

2026年1月3日。https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles13. Lattimore T、Szepesvári C。《多臂老虎机算法》。剑桥大学出版社；2020年。doi:10.1017/978110857140114. Lu J、Liu A、Dong F、Gu F、Gama J、Zhang G。在概念漂移下学习：综述。《IEEE知识与数据工程汇刊》。2019年；31(12):2346-2363。doi:10.1109/TKDE.2018.287685615. Nestor B、McDermott MBA、Boag W等。非平稳健康记录中的特征鲁棒性：常见临床机器学习任务中可部署模型性能的注意事项。《机器学习研究论文集》。2019年；106:381-405。16. Nestor B、McDermott MBA、Chauhan G等。重新思考临床预测：为什么机器学习必须考虑诊疗年份和特征聚合。arXiv。预印本在线发布于2018年11月30日。arXiv:1811.12583v117. Ovadia Y、Fertig E、Ren J等。你能信任模型的不确定性吗？评估数据集偏移下的预测不确定性。《神经信息处理系统进展》。2019年；32。18. Rockafellar RT、Uryasev S。一般损失分布的条件风险价值。《银行与金融杂志》。2002年；26(7):1443-1471。doi:10.1016/S0378-4266(02)00271-619. Rockafellar RT、Uryasev S。条件风险价值优化。《风险杂志》。2000年；2(2):1-42。doi:10.21314/JOR.2000.03820. Subbaswamy A、Saria S。从开发到部署：机器学习中的数据集偏移、因果性和偏移类型。arXiv。预印本在线发布于2019年11月6日。arXiv:1911.02511v121. Sutton RS、Barto AG。《强化学习：导论》。第2版。麻省理工学院出版社；2018年。22. 美国食品药品监督管理局。人工智能/机器学习(人工智能/机器学习)-驱动的医疗器械软件(医疗器械软件) 行动计划。发布于2021年1月。更新于2025年3月25日。访问日期为2026年1月3日。https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device23. 美国食品药品监督管理局。医疗器械开发良好机器学习实践：指导原则。发布于2021年10月。更新于12月19日

，2025年。访问日期：2026年1月3日。https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles24. 美国食品药品监督管理局。机器学习医疗器械的预定变更控制计划：指导原则。发表于2023年4月。访问日期：2026年1月3日。https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device25. Vokinger KN、Feuerriegel S、Kesselheim AS。医疗器械中的持续学习：美国食品药品监督管理局的行动计划

n

及其他内容。《柳叶刀·数字健康》。2021年；3(6)：e346-e347。doi:10.1016/S2589-7500(21)00076-426. Zink A、Rose S。医疗支出的公平回归。《生物统计学》。2020年；76(3):973-982。doi:10.1111/biom.13206