

观点与分析

编辑日期: 2026年11月1日

改变了所预测未来的预测

> 警报本来告诉我们可能发生什么。我们据此采取行动。等到它警告的事情没有发生，我们又要求它为自己的真实性辩护，仿佛因为消防员进入的房屋没有烧毁，就要审判消防员。

有些病区听见了警报，另一些没有

2019 年 10 月至 2021 年 7 月，沙特国民警卫队卫生事务部所属五家医院开展了一项不能简单概括为“技术成功”的试验。45 个内外科病区分九批，从常规照护依次转入电子脓毒症早期筛查系统。脓毒症是可伴随感染出现的危险病理反应。这不是机器人秘密诊断，也不是语言模型撰写治疗建议；它更简单，因此对本文的问题更有用。

电子病历依据快速序贯器官衰竭评分（quick Sequential Organ Failure Assessment, qSOFA）的三项指标计算警报：收缩压降到某个阈值、呼吸频率升高或意识水平异常。规定时间窗内出现两项，系统便产生信号。在进入干预的病区，信号会呈现给照护团队并连接一条响应路径；其他病区则以不显示的模式用于比较。研究者因此能提出一个比计算准确性更重要的问题：人们得到一项可以采取行动的警报后，什么发生了变化？[1]

研究包括 60,055 名患者，其中电子筛查组 29,442 人，对照组 30,613 人。干预组在警报后接受乳酸检测和静脉补液的概率增加。但脱离语境看死亡表会令人困惑：90 天内院内死亡粗率在干预组为 3.2%，对照组为 3.1%，粗比较没有显示筛查带来清楚优势。按研究预设分析，对时间段、患者在病区与医院内的聚集以及新冠状态作调整后，调整相对风险估计为 0.85，95% 置信区间为 0.77 – 0.93。这是调整后的相对风险下降估计，不代表每一位患者都因项目获救，也不代表粗比较可以单独裁决结果。[1]

粗结果与调整结果为什么不同？因为这是分阶段转换试验，病区与时期不断变化，新冠疫情又改变了各阶段患者特征和风险。研究者说明，评价效果必须调整这些因素。对预测研究而言，这一点至关重要：即使拥有随机试验，自然也不会把一个现成因果答案放在一行数字里。我们仍要知道各组何时、怎样接触干预，以及周围条件如何变化。1

这并不能证明一个机器学习模型预测风险后，因为救下患者而失去准确性。沙特系统使用电子 qSOFA 规则，论文没有检验“机器学习模型部署后性能漂移”的说法。它提供的是本文回路的实验证据：风险信号改变团队行动，团队行动又能改变我们试图预测的事件。问题从这里开始变得不同于分类质量。

一项成功警报“错了”，究竟是什么意思？

设想一名患者出现恶化征象。风险系统给出未来 24 小时严重事件高概率，团队迅速检查并实施适当干预，事件没有发生。一个月后数据部门回看记录：“它预测了风险，结果没有发生。这不是误报吗？”

也许是，也许不是。

若模型给出的概率指“继续常规照护、没有新增警报时发生事件的概率”，就不能把它当作“我们响应警报后发生事件的概率”来检验。新信息改变了原本要预测的路径。若模型一开始就预测团队按预期响应后会发生什么，那么用该政策下的实际结果比较是合理问题，但预测目标已不同。

统计学中，常规政策下的事件风险可写为：$R_0(x)=P(Y(0)=1\mid X=x)$。警报与响应项目下的风险可写为：$R_1(x)=P(Y(1)=1\mid X=x)$。X 是行动前已知信息；$Y(0)$ 是没有新项目但仍有常规医疗与治疗时的潜在结果；$Y(1)$ 是实施警报项目时的潜在结果。二者差异是在特定设计假设下，政策对同类人可能产生的变化。

同一患者的两个结局不能同时被观察。我们只看到实际选择路径中的结果，看不到同一患者处于另一个世界、团队没有听见警报的副本。这就是潜在结果（Potential Outcomes）问题，也让“什么都没有发生”远没有看上去那么有信息量。什么都没有发生，是在什么之后？经过哪种干预？如果沿着另一条路径继续，最可能发生什么？

2020 年，Juan Perdomo、Tijana Zrnic、Celestine Mendler-Dünner 与 Moritz Hardt 提出履行性预测（Performative Prediction）框架。模型支持决策时，可能影响未来将用于评价模型的数据分布；预测发布以后，模型所预测的世界不再等同于它生活的世界。研究讨论反复训练、预测与数据相互影响后的数学稳定性，并没有证明每一个准确性下降的警报都成功实现了预防。2

会改变人类行为的预言与什么也不改变的预言之间，不只是哲学差异。天气预报通常不会改变云层；患者恶化预测却会改变医生决定，学生可能掉队的预测会改变支持方式，工业事故预测会改变检修计划。在这里，人工智能不只是未来的镜子，也是制造未来的新变量。

一家从护士行为中提取风险信号的医院

我们还需要一个真正使用机器学习的案例。2025 年 4 月，《Nature Medicine》发表 CONCERN 患者恶化预警系统随机试验。系统并非主要依靠化验值，而是从护理记录与监测的时间模式中学习：护士何时增加测量、何时留下异常记录、何时在例行安排之外观察患者。团队行为的变化本身，成为身体可能恶化的先行信号。[3]

试验覆盖两个美国卫生系统随机分配的 74 个临床单元和 60,893 次住院。把 CONCERN 警报嵌入工作流程，与调整后院内死亡瞬时风险下降相关；风险比为 0.64，95% 置信区间为 0.53 – 0.78。0.64 不等于“64% 的患者因人工智能康复”，而是研究生存分析中的比较瞬时风险率。论文分析中，住院时长也有所下降。[3]

但同一张表里，意外转入重症监护的调整后瞬时风险上升，风险比为 1.25，置信区间 1.09 – 1.43。死亡下降，一个看似恶化的指标为何上升？一种可能是更早发现促使团队把更多患者转入更高级照护，但研究不能让我们判定每次转入都是成功救治。另一些情况下，意外转入仍可能是本应预防的恶化。需要知道时间、原因、临床适当性与结局变化，不能只看数字前面的加减号。[3]

管理报告因此有两张面孔。一个部门可以说系统改善了某项严重结局，另一个部门可以指出意外转入增加并要求解释。两者都在阅读真实结果。差异揭示更大事实：成功警报会改变事件版图；在理解产生新事件的干预路径前，不能先给它贴上成功或失败的含义。

研究者也公开了限制：试验在两个美国卫生系统开展，时期受新冠影响；重症单元组合、记录实践不同，部分结果可能受未测因素影响。因此，它比营销故事更有证据力，却远不足以保证所有医院与所有警报系统都会得到相同结果。[3]

学校里也一样：风险信号下降，最终指标没有动

为了不把悖论困在重症监护室，可以看一项美国高中研究。2014 – 2015 学年，后来由美国教育部发布的研究测试了面向有毕业风险学生的 EWIMS 预警、干预与监测系统。73 所高中被随机分配：37 所实施项目，36 所在评价年度继续常规流程。研究覆盖高中第一、二学年的 37,671 名学生。它不是语言模型或现代机器学习系统试验，而是早期指标、跟进与干预体系。[4]

一年后，项目学校长期缺勤率为 10%，对照学校为 14%；至少一门课程不及格的学生比例为 21% 对 26%，两项差异均具有统计显著性。但未获得足够学分、无法按四年进度毕业的学生，在两组都是 14%；一年后，研究未发现该学业进度指标存在可检测效应。研究也报告，项目七个步骤实施困难，许多学校总体执行较弱。[4]

教训是什么？构成风险的行为可能先于最终目标变化。长期缺勤是一项信号，课程不及格是一项信号，毕业进度则更接近远期目标。若新系统用后续干预减少部分缺勤或不及格后的记录来训练，再把每位改善的学生称作“风险误判”，就混淆了“谁需要帮助”与“谁在帮助后仍然失败”。反过来，如果出勤指标改善而毕业进度没有变化，也不能宣布预防已经实现最终目标。

学校与医院使用的技术不同，成功定义也不同。共同点只有一条：风险标记被用于改变人的行为后，随后的结果成为机构响应的一部分，不再是对标记预测能力的中立考试。

找到高风险者还不够

这种区分不能被用作任何弱模型的舒适借口。在美国新泽西州卡姆登，一项强化照护协调项目针对频繁住院且需求复杂的患者。运营假设很有吸引力：既然知道谁反复入院，就提供协调的医疗与社会支持，降低再次入院。

但一项纳入 800 名患者的随机试验没有发现 180 天再入院率具有统计显著下降。项目组为 62.3%，常规照护组为 61.7%；调整分析差值为 0.82 个百分点，置信区间从下降 5.97 点到上升 7.61 点。更重要的是，仅比较项目组干预前后，会显示近 38 个百分点的大幅下降，因为患者在异常高使用期后进入试验，随后两组使用都下降。不能把这种下降自动归功于项目。5

卡姆登不是人工智能试验，却对一个常见混淆给出关键实证反驳：找到最容易发生事件的人，不等于找到能从现有干预中获得最大预防收益的人。最高风险者未必是特定项目预期收益最大者；风险原因可能不响应项目、需要项目没有提供的资源，或本来就是短期状态。

纸面上，两个模型可以按再入院概率给患者排序，并一致选出最危险者。但若问谁会从特定干预中受益最多，排序可能完全不同。一个问题是风险水平，另一个是风险可改变性。混在一起，会让高预测准确性在没有证据时看上去像卓越预防政策。

同样难题出现在 2026 年 2 月一项心脏病区研究中：10,422 次住院接受人工智能恶化预测的被动展示试验，没有强制规定临床响应。研究未发现主要结局改善；医生把约 11% 患者在显示与不显示病床之间转移，削弱了实际随机化的部分性质。这不是所有预警模型失败的结论，而是提醒：知道风险不会自动产生能够降低风险的行动。6

于是有两个不能分开的部分：警报可能正确，干预也可能救下患者；警报也可能不准，或干预没有效果。区分它们需要因果设计，而不是对算法抱有善意。

我们需要的结果，不是手里已有的结果

Nan van Geloven 等在 2020 年论文中强调，当风险测量后开始治疗时，必须明确模型预测的对象。他们借鉴临床试验中目标效应量的命名，提出 Predictimand：我们预测不增加治疗时会发生什么，还是训练数据中既有治疗实践继续时会发生什么，或是在治疗开始前事件是否发生？这些问题有不同答案、估计方法与假设。7

2025 年 10 月，Christopher B. Boyer 等进一步研究反事实预测模型的估计与评价。他们解释，不能用一种治疗政策下观察到的数据，直接测量另一种假设政策下的预测质量，并提出在反事实世界中估计损失、校准和 ROC 曲线下面积等性质的方法，同时列出统计识别条件与假设。案例是未开始他汀治疗者的心血管风险模型。这些是有条件的方法学结果，不是让未观察结局变得可见的配方。8

这里必须区分三个量：

1. 基线风险：在参考政策下发生伤害的概率，例如没有新警报系统的常规照护。
2. 干预后风险：按照已知政策使用警报、人员作出实际响应时发生伤害的概率。
3. 政策收益：从第一种政策转向第二种后减少的伤害，并计入副作用、成本、资源与其他影响。

基线风险可能很高，干预后风险较低，政策收益很大；也可能基线风险同样高，干预却无效，收益很小；干预还可能伤害某些群体。一个“风险分数”不能替我们完成选择。

只作数值说明：假设甲在常规政策下事件风险为 30%，某项干预下为 10%；乙的风险从 20% 降到 19%。估计收益差分别为 20 个百分点与 1 个百分点。这些是解释用假设数字，不是研究结果；每个人只观察一次结果，不能据此知道已实现的个体收益。我们只能在合适设计和假设下，尝试估计群体平均效应。

系统定期用自身使用后产生的数据再训练时，问题更复杂。高分者经治疗、转院或支持后未发生事件，模型会从这些新关系中学习。它可能更善于预测当前政策下会发生什么，却更不适合估计未实施政策时会发生什么。再训练并非总是错误，但问题定义必须稳定。不能悄悄更换预测对象，再把它叫作版本更新。

为什么准确率会误导?

公司宣布模型 ROC 曲线下面积很高，最多说明它在特定测试集中的区分能力，不能单独回答使用对患者、学生或机构的影响。成功干预部署后预测指标下降，也不能在不知道原目标和政策性质时直接解释。事件分布可能因干预改变，人群特征可能变化，模型质量也可能真的下降。

有三场不同审判，不能让其中一场吞掉另外两场：

第一，预测审判：在明确参考世界中，风险估计是否可靠？包括校准（Calibration）、区分（Discrimination）、数据正确性以及时期与人群适配。

第二，响应审判：信息是否改变决策者行为？从警报到检查或处置用了多久？行动是否适当，目标人群是否响应?

第三，结果审判：与常规照护或替代政策相比，政策有什么影响？伤害是否下降，是否出现其他伤害、成本、不必要转诊，或人员与资源耗竭?

三场审判不能压成一个比例。沙特试验中，检测与补液增加是团队响应，死亡是另一项结局。CONCERN 中一种风险下降，重症转入增加。EWIMS 学校里一些指标改善，一年后毕业进度未见效应。卡姆登虽找到高使用群体，项目对主要目标没有证明收益。机制不同，科学判断都要求把预测、行动与影响分开。

这也不同于某些人根本没有机会表现结果，导致标签缺失的情形。预防性预测中的结果可能完整记录，却是在政策干预后发生的结果。记录里没有事件，不能独自说明没有干预时是否会发生。这不是表格少一列，而是世界缺少另一条路径。

如何检验会阻止部分预测事件的模型？

在伦理与机构条件允许时，最直接的实验道路是：所有人继续获得常规照护或服务，把尚未证明有效的新项目附加功能随机分配给相近单位或分阶段部署；在揭示预测前记录预测、揭示时间、人员行动与后续结果。这不是剥夺必要治疗或让个人暴露于已知风险。敏感场景还必须经专业与伦理治理批准。

比较期间，某些系统可继续在隐藏模式计算风险分数，同时保护信息并尊重常规照护边界。研究的是向病区或群体提供警报的效果，不是假装知道活下来的人中谁原本会死亡。分阶段转换与集群随机适合知识会在同一单位员工间泄漏的情形，但也面临时期与公共事件挑战，沙特医院试验遭遇新冠正说明这一点。

不能随机试验时，可以用观察数据模拟目标试验（Target Trial Emulation），或使用处理混杂调整与反事实效应估计方法。但假设不能隐藏：必须识别影响决定与结局的混杂因素，让个体具有合理机会接受不同选择，以一致方式定义干预，并处理单位间相互影响。并非每个环境都满足这些假设。

若一人的干预占用了另一人需要的重症床位，问题就不再是孤立个体效应，而是资源配置、拥堵与相互干扰（Interference）。一项政策可能对几十人有效，推广到全国病区后结果却不同。不能检验模型准确率以后，就假设警报翻倍等于救治翻倍。响应能力有限，过度警报产生成本，团队行为也会随时间改变。

预测—干预账：一项可被推翻的建议

本文建议机构建立“预测—干预账”（Forecast – Intervention Ledger）。这是编辑性提案，不是公认标准，也不是已证实绩效的系统。它不是为了给仪表盘增加字段，而是把三件常被混淆的事分开：干预前知道什么、因它采取什么行动、此后发生什么。

对每个符合条件的案例，记录事前信息收集时间、模型版本、分数所预测风险的定义、预测期限、警报发出时间、是否向决策者显示、随后响应与时间、案例结局和可能伤害。独立分析层再对可比群体估计因果效应，并保存不确定性范围与假设。不能给“如果没有警报，这名患者会怎样”填写一个虚构个体数字，因为那个未来没有被观察。

该账可以形成三张分别审查的面板：

- 预测面板：参考政策下的模型校准、警报敏感性，以及风险分数对目标群体的适配。
- 干预面板：引发适当响应的警报比例、时间间隔、负担、转诊与伤害。
- 影响面板：与明确对照政策相比的结局差异，连同置信区间、成本与资源分配影响，而不只是被警报者后来发生什么。

推翻系统价值的实验也必须清楚：若良好比较下重要结局没有显著差异，警报伤害超过收益，或找到最高风险群体并不改变干预有效性，“我们预防了不可见事件”的修辞也不能挽救它。我们既允许模型在结果出来前存在成功可能，也必须允许它最终没有改善世界。

这本账最重要的规则是，答案永远不能是：“安装系统后案例减少，所以系统救下了它们。”案例可能因人群、季节、照护变化或异常高峰回落而下降。卡姆登的教训很严厉：项目组前后比较显示巨大改善，随机对照组却经历了相似路径。

医院董事会、教育部门或公共安全机构真正要检验的，不是系统发出多少警报，也不只是灾难多少天没有发生，而是：这项特定政策改变了多少伤害、影响了谁、付出什么成本，我们对推断有多大把握？

当第二个未来成为科学问责的一部分

我们不要求人工智能预知未知，也不要求研究者把没有发生的未来显示在屏幕上。需要承认的是，预防决策制造了传统分类测试无法独自解决的悖论：有些良好结果可能来自成功干预，另一些来自过度警报、运气或自然变化。算法不能只挑一个最讨喜的故事来解释自己。

因此，不能因为模型害怕一切而奖励它，也不能只因它害怕的事情没有发生就惩罚它。不能把患者风险与可治疗性混为一谈，也不能把学生可能失败与得到教师帮助后能否改善混为一谈，更不能把项目后所有下降都归功于项目。

履行性预测要求机构完成的最深转变，是从一个看似科学的问题“你预测到了后来发生的事吗”，走向更难也更公平的问题：“如果我们没有因你的预测行动，会发生什么；在没有假装看见两个未来的情况下，怎样检验差异？”

到这里，成功的含义会改变。最好的警报机器不是让最多警告成为现实，因为风险实现不是预防目标；也不必然是部署后所有事件都消失，因为消失未必由它造成。最好的系统，是我们能用实验或严谨因果推断证明，它帮助人类降低了坏结果的发生概率，同时不隐藏代价与替代伤害。

我们已经学会检验计算机是否知道关于未来的某些事情。接下来还要学会检验：它的知识是否让未来变得更好。

参考文献

1. Arabi 等，JAMA，电子脓毒症筛查（2024 年在线、2025 年印刷）。
2. Perdomo 等，Performative Prediction，ICML 2020。
3. Rossetti 等，CONCERN 随机对照试验，Nature Medicine 2025。
4. 美国教育科学研究所，EWIMS 影响研究，2017 年 4 月。
5. Finkelstein 等，医疗高使用者项目随机试验，NEJM 2020。
6. Keim-Malpass 等，人工智能分析随机试验，Scientific Reports 2026。
7. van Geloven 等，Predictimands，European Journal of Epidemiology 2020。
8. Boyer 等，反事实预测评价，Statistics in Medicine 2025。