优化　　通过预测患者生理状况优化抗生素转换

Magnus Ross[1*]、Nel Swanepoel[2]、Akish Luintel[3]、Emma McGuire[3]、
Ingemar J. Cox[1,4]、Steve Harris[2]、VasileiosLampos[1*]

[1]计算机科学系、人工智能中心，伦敦大学学院，英国[2]健康信息学研究所，伦敦大学学院，英国[3]医学微生物学系、感染科，伦敦大学学院医院，英国[4]计算机科学系，哥本哈根大学，丹麦

arXiv:2603.08242v1 [cs.LG] 9 Mar 2026

摘要

及时从静脉注射（静脉给药）转为口服抗生素治疗可缩短住院时间、减少导管相关感染并降低医疗保健成本，但在英格兰，每五名患者中仍有一名尽管符合转换标准，却继续接受静脉抗生素治疗。临床决策支持系统能够提高转换率，但从历史决策中学习的方法会重现常规实践中的延误与不一致。我们提出使用神经过程对生命体征轨迹进行概率建模，通过将预测结果与临床指南进行比较（而非学习过去的处置）来预测转换准备度，并对患者排序，以便优先进行临床评估。这一设计可生成可解释输出，无需重新训练即可适应更新后的指南，同时保留临床判断。在 MIMIC-IV（美国重症监护，6,333 条就诊记录）和 UCLH（一家大型城市英国学术医院集团，10,584 条就诊记录）上验证后，该系统识别出比随机选择多 2.2–3.2× 倍的相关患者。我们的结果表明，预测患者生理状况为抗生素管理中的决策支持提供了有原则的基础。

1 引言

及时将住院患者从静脉注射（IV）抗微生物治疗转为口服抗微生物治疗，即静脉转口服抗生素转换（IVOS），有益于患者和医院 [1]。从临床角度看，这种转换可以缩短住院时间并减少导管相关血流感染 [2, 3]；从运营角度看，则可释放护理时间并降低成本 [4]。然而，即便已有安全的口服替代方案，静脉治疗往往仍会持续。在英格兰，19% 的患者虽符合转换标准 [5]，仍继续使用静脉抗生素，有些患者则完全没有转换 [6]。静脉转口服抗生素转换有赖于及时审查不断变化的临床状况，但在繁忙的病房中很容易漏掉这项审查，尤其是在高级医护人员意见或微生物学建议延迟时。

静脉转口服抗生素治疗决策包含两个相互关联的部分 [7, 8]。第一部分是临床稳定性：即患者是否正在好转并可能保持稳定，这可通过生命体征（例如体温、心率、血压、呼吸频率、血氧饱和度）的变化趋势以及在可获得时结合实验室指标进行评估。第二部分是口服适用性：即口服治疗是否合适且可行，例如患者能够耐受经口摄入，并且有合适的口服药物可用。本文聚焦于临床稳定性。我们的任务是识别近期观测值表明其

*电子邮件：magnus.ross@ucl.ac.uk, v.lampos@ucl.ac.uk

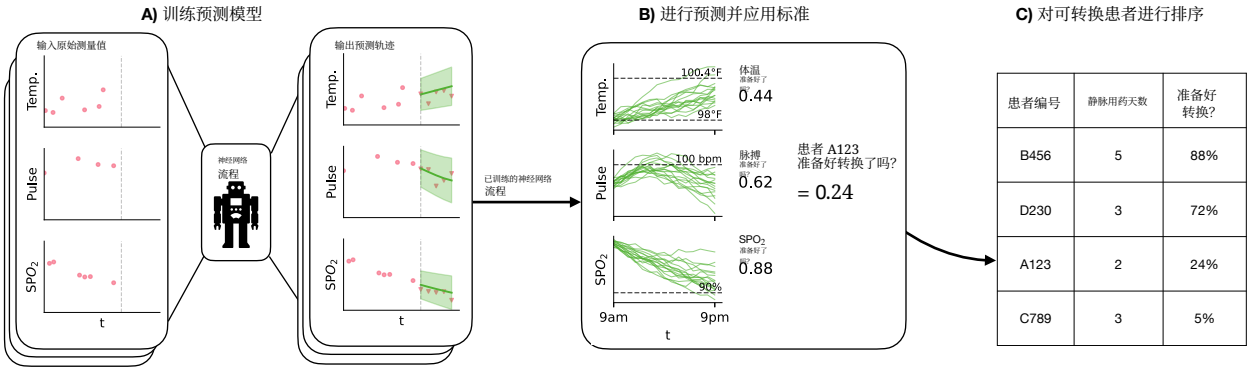


图 1：拟议的人工智能赋能的静脉转口服抗生素治疗系统。**A)** 训练概率神经过程模型，以预测患者轨迹。**B)** 使用训练好的模型生成未来一天可能出现的轨迹，并应用临床标准确定可转换患者所占的比例。**C)**将转换准备度概率以按优先级排序的患者列表形式呈现，供审查。

近期在正常范围内保持稳定的概率较高、因而可以优先进行审查的患者。

这一问题受到实际条件的制约 [9]。在英格兰，一家医院平均约有 550 名普通科、急症科及产科住院患者，且其中约三分之一正在接受抗生素治疗。这样的患者数量超出了典型抗菌药物管理团队（感染专科医师和专科药师）的工作能力。因此，决策支持必须帮助他们聚焦于最有可能从审查中获益的患者。静脉转口服抗生素治疗决策还取决于一些难以通过结构化电子健康记录数据可靠捕捉的信息（例如手术伤口愈合情况、感染源控制计划、不断变化的影像学检查结果或床旁评估）。因此，临床决策支持系统的目标并非自动执行静脉转口服抗生素治疗，而是筛选出患者，供专家审查并与主管团队讨论。

临床决策支持系统（CDSS）可以提高转换率并减少静脉治疗天数。以往基于规则的系统通过基于固定标准的警报减少了静脉给药 [10,11, 12, 13] （详细审查见补充材料 S1）。然而，固定规则难以判断患者是否正在改善，而非仅仅在某个时间点达到阈值。通过历史转换行为进行学习的 AI 方法面临着更深层的局限：如果训练信号是临床医生所做的，那么模型所能做到的最好程度也只是复现这种欠佳行为。它无法可靠地按照临床标准改善实践 [14, 15]。

我们提出一种替代方法：预测患者生理状况，而非过去的临床行动。我们使用神经过程对生命体征轨迹进行概率式建模，并将临床标准应用于这些预测，以判断患者的转换准备度。这样既避免学习历史延误，也无需从电子健康记录（EHR）中标注转换情况，并能生成临床医生可直接检查的输出（预测轨迹及其不确定性）。由于临床标准是在预测之后应用的，因此临床标准演变时，无需重新训练预测模型即可更新系统；系统还可适用于非典型病例（例如，对于正在使用退热药的患者，可忽略在具体情境中缺乏临床信息价值的变量，如发热）。系统根据预测的转换准备度对患者进行排序，帮助医疗团队确定审查优先级，同时将常规生命体征数据之外因素的临床判断权保留给临床医生。

图1展示了我们的系统架构。我们预测关键生命体征（**A**），应用临床标准估算转换准备度概率（**B**），并据此对患者进行排序，以供审查（**C**）。口服适用性和最终的静脉转口服抗生素治疗决策仍由临床医生作出；他们会综合考虑更广泛的临床病历，并在审查被标记的患者后与主管团队协调 [7, 8]。

统计数据	MIMIC	UCLH
入院人数	6,333	10,584
独立受试者	5,850	8,701
年龄（岁）	65.70 ± 15.29	55.98 ± 19.24
住院时长（天）	17.69 ± 17.07	14.23 ± 14.42
Sex		
女性	2,579 (40.7%)	5,564 (52.6%)
Male	3,754 (59.3%)	5,019 (47.4%)
前5种族/族裔		
	白人：3,772（59.6%）	英国白人：3,324（31.4%）
	未知：1,074（17.0%）	未知：3,033（28.7%）
	黑人：720（11.4%）	其他白人背景：1,204（11.4%）
	其他：292（4.6%）	其他族裔群体：671（6.3%）
	西班牙裔：247（3.9%）	非洲裔黑人：509（4.8%）
口服抗生素前五名		
	阿奇霉素：724（11.4%）	阿莫西林-克拉维酸：3,471（32.8%）
	甲硝唑：704（11.1%）	克拉霉素：797（7.5%）
	磺胺甲噁唑/甲氧苄啶：477（7.5%）	环丙沙星：771（7.3%）
	多西环素：371（5.9%）	阿莫西林：421（4.0%）
	万古霉素：370（5.8%）	多西环素：410（3.9%）
前五种静脉注射抗生素		
	万古霉素：4,158（65.7%）	阿莫西林-克拉维酸：4,920（46.5%）
	头孢吡肟：2,774 (43.8%)	哌拉西林-他唑巴坦：3,841 (36.3%)
	头孢曲松：2,629 (41.5%)	甲硝唑：2,849 (26.9%)
	头孢唑啉：2,249 (35.5%)	美罗培南：1,208 (11.4%)
	甲硝唑：1,449 (22.9%)	环丙沙星：882 (8.3%)
测量变量	每次住院平均次数	每次住院平均次数
心率	157.8	81.1
呼吸频率	154.0	78.4
SpO_2	154.9	80.8
收缩压	84.0	62.9
体温	36.1	61.9

表1：MIMIC-IV和UCLH数据集评估集队列的比较。分类变量以计数（入院记录百分比）表示。连续变量以均值 ± 标准差表示。MIMIC-IV记录种族，而UCLH记录族裔，这反映了各源数据库所使用的编码系统；“其他白人背景”表示其他白人背景。由于患者在一次入院期间可能接受多种抗生素，抗生素频次总和可能超过100%。底部部分显示了每个变量在每次入院记录中记录的生命体征测量值的均值。缩写：Pip.-Taz.（哌拉西林-他唑巴坦）、Sulfameth/Trimeth（磺胺甲噁唑-甲氧苄啶）、IV（静脉注射）、SpO_2 （血氧饱和度）、BP（血压）。

数据集	模型	变量				
		HR (每分钟心跳次数)	**RR** (每分钟心跳次数)	**SpO_2** (%)	**SBP** (mmHg)	**体温** (°F)
UCLH	**重复**	8.79 [8.76-8.82]	1.97 [1.96-1.98]	1.51 [1.51-1.52]	12.86 [12.81-12.91]	0.69 [0.69-0.70]
	GBDT	7.94 [7.92-7.96]	1.76 [1.76-1.77]	**1.31 [1.30-1.31]**	10.86 [10.83-10.89]	0.58 [0.58-0.58]
	NP	**7.58 [7.56-7.60]**	**1.71 [1.70-1.72]**	**1.30 [1.30-1.31]**	**10.71 [10.67-10.75]**	**0.57 [0.57-0.57]**
MIMIC	**重复**	9.17 [9.14-9.19]	4.09 [4.08-4.11]	2.05 [2.04-2.05]	14.29 [14.24-14.34]	0.63 [0.62-0.63]
	GBDT	8.92 [8.90-8.94]	3.49 [3.48-3.50]	1.79 [1.78-1.79]	12.27 [12.24-12.30]	0.55 [0.55-0.55]
	NP	**8.32 [8.29-8.34]**	**3.40 [3.39-3.41]**	**1.76 [1.76-1.77]**	**11.97 [11.92-12.01]**	**0.54 [0.54-0.54]**

表2：12小时预测范围内模型预测的平均绝对误差。图中显示了95%置信区间，该区间通过对测试集各窗口的平均误差进行自助法重采样计算得出。粗体值表示表现最佳的模型（$p \geq 0.05$）。变量缩写：HR（心率）、RR（呼吸频率）、SpO_2（血氧饱和度）、SBP（收缩压）、Temp.（体温）。模型缩写：GBDT（梯度提升决策树）、NP（神经过程）。

2 结果

我们使用两个不同的电子健康记录数据集验证了该方法：MIMIC-IV v3.1 [16], ，其中包含来自美国波士顿 Beth Israel Deaconess Medical Center 的重症监护病房入院病例；以及来自英国伦敦综合医院集团 University College London Hospitals (UCLH) 的非公开数据集。这些队列的人群和临床实践存在显著差异，为检验模型的泛化能力提供了有力依据（队列的详细特征见表1）。

我们的临床决策支持系统的核心任务是通过预测患者的生理稳定性，为每日静脉转口服抗生素转换审查提供支持。除非另有说明，我们预测未来12小时内的五项关键生命体征——心率、呼吸频率、血氧饱和度、收缩压和体温。如果患者的所有预测生命体征在整个时间窗口内均保持在临床正常范围内，则定义为“具备转换条件”。我们使用神经过程（NPs） [17], 对这些变化轨迹进行建模。神经过程是一类概率模型，因其能够处理临床时间序列数据固有的不规则采样和不确定性而被选用。模型架构和训练流程的详细信息见第4.3节。

2.1 预测

我们首先旨在评估所提出的神经过程模型的预测性能。表2展示了所提出的神经过程模型（NP）和两种基线方法的点预测性能。模型的任务是预测紧接预测时间之后12小时内发生的任意测量指标的值；预测时间定义为已知患者病史的截止时点。重复基线仅将预测窗口之前最近一次测量的值作为预测。梯度提升决策树则更为复杂，会利用根据既往测量数据生成的一组统计特征 [18]。有关基线构建的讨论，请参见第4.4节。除 SpO_2，外，NP 模型在所有测量指标上的表现均显著优于两种基线模型，不过梯度提升决策树的得分仅略逊一筹。与梯度提升决策树相比，NP 模型在各变量上的平均得分提升幅度，在 UCLH 数据中为 2.25%，在 MIMIC 数据中为 2.99%。与 MIMIC 数据相比，UCLH 数据更容易预测，这体现在所有模型对每项测量指标的误差都更低。这可能是因为 MIMIC 重症监护病房队列中的患者病情更严重、状况也更不稳定。

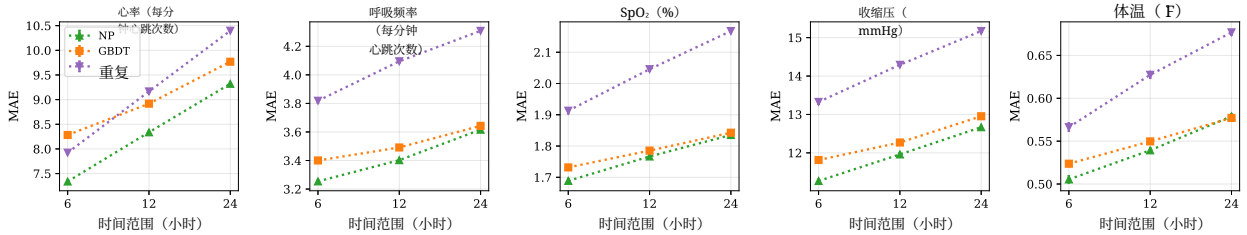


图 2：对于神经过程模型和基线方法，随着预测时域增加，各变量在 MIMIC 数据上的预测性能。请注意，横轴采用对数刻度。误差线表示 95%置信度。变量缩写：HR（心率）、RR（呼吸频率）、SpO_2（血氧饱和度）、SBP（收缩压）、Temp.（体温）。模型缩写：GBDT（梯度提升决策树）、NP（神经过程）。

尽管对于我们的临床情境而言，12小时的预测时域最为合适，但不同医疗机构的工作流程各异，因此适用时长可能有所不同。因此，我们还在另外两个时域——6小时和1天——上验证了模型性能。图2展示了这项实验在 MIMIC 数据上的结果。正如预期，随着预测时域延长，平均绝对误差也随之增大；对更远未来时点的预测准确率较低。在较短时域内，NP 在各变量上的表现均显著优于所有基线方法。随着时域延长，梯度提升决策树与 NP 的性能逐渐趋同，在1天时域下，两种模型的表现相近。需要指出的是，NP 模型是概率式模型，因此会针对可能的测量值生成预测分布，而不像基线模型那样只给出单一点估计。这种概率式输出使其预测结果能够用于下游任务，例如对患者的切换准备状态进行排序，正如我们将在下一节中看到的那样。表2未评估模型的这一概率式特性，仅量化了点预测性能。

2.2 切换准备度预测

所提出的临床决策支持系统中的关键机器学习任务，是预测患者在预测窗口内是否已准备好进行转换。我们采用一组临床标准，将关键生命体征测量值分别限定在正常范围内，以此作为切换准备就绪的替代指标。如果患者的测量值在整个预测窗口内均处于正常范围，我们便将其标记为当天已准备好进行转换。这两个数据集的类别分布存在明显差异，反映了患者群体的不同：在 UCLH 数据中，患者在 40.2% 的测试日已准备好进行转换；相比之下，在病情更严重的 MIMIC 重症监护病房患者群体中，这一比例仅为 12.2%。

神经过程模型会生成预测测量值的分布。我们利用这一分布，确定患者在预测窗口内满足相关标准的概率，进而确定其已准备好进行转换的概率。切换准备度概率较高的患者，在呈现给临床医生的列表中排名更高，因此会被优先审查。该两步流程首先预测生命体征，然后应用相关标准，从而生成临床医生可用于辅助决策的可解释输出，因为预测概率所依据的患者状态预测是清晰易懂的。

另一种可解释性较弱的方法是使用监督分类模型直接预测切换准备状态标签。我们采用直接预测方法作为基线。作为基线模型的输入，我们使用根据过去测量值生成的一组统计特征。我们分别采用逻辑回归和梯度提升决策树。对于神经过程模型，我们使用两个变体：上文所述的标准神经过程，以及一种经过调优、旨在生成

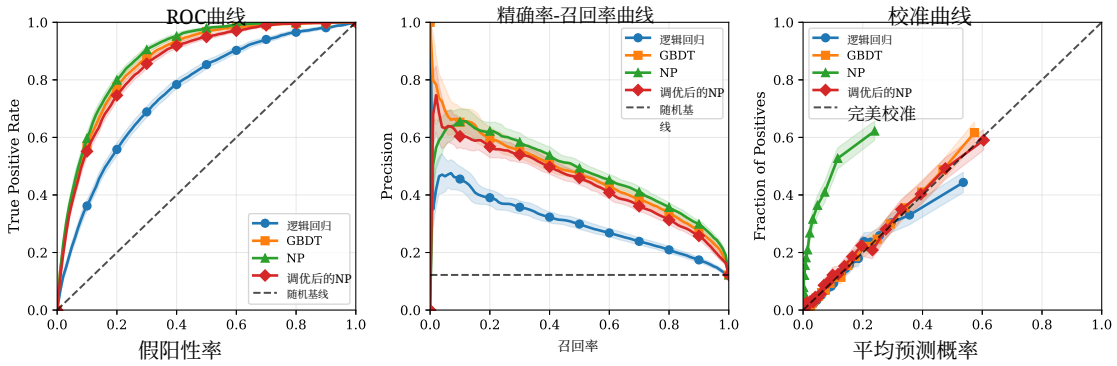
数据集	模型	指标			
		AUROC	AP	布里尔分数	前5项精确率
UCLH	逻辑回归	0.827 [0.823-0.831]	0.740 [0.732-0.747]	0.167 [0.165-0.169]	0.836 [0.823-0.850]
	GBDT	0.862 [0.858-0.865]	**0.794 [0.787-0.800]**	0.149 [0.147-0.151]	**0.899 [0.888-0.910]**
	NP	0.858 [0.854-0.861]	0.785 [0.779-0.792]	0.216 [0.212-0.219]	0.883 [0.871-0.894]
	调优后的NP	**0.867 [0.864-0.871]**	**0.800 [0.793-0.807]**	**0.146 [0.144-0.148]**	**0.901 [0.890-0.912]**
MIMIC	逻辑回归	0.759 [0.749-0.767]	0.300 [0.284-0.315]	0.098 [0.095-0.100]	0.280 [0.263-0.297]
	GBDT	0.869 [0.863-0.875]	**0.471 [0.453-0.490]**	**0.081 [0.079-0.084]**	**0.381 [0.363-0.400]**
	NP	**0.882 [0.877-0.887]**	**0.477 [0.459-0.497]**	0.106 [0.102-0.110]	**0.393 [0.373-0.413]**
	调优后的NP	0.862 [0.856-0.868]	0.453 [0.433-0.472]	**0.083 [0.080-0.085]**	0.369 [0.350-0.387]

表3：切换准备度预测的模型性能比较。除布里尔分数外，所有指标的值越高越好。粗体数值表示性能最佳的模型（p≥ 0.05）。指标缩写：AUROC（受试者工作特征曲线下面积）、平均精度（AP）。模型缩写：GBDT（梯度提升决策树）、NP（神经过程）。

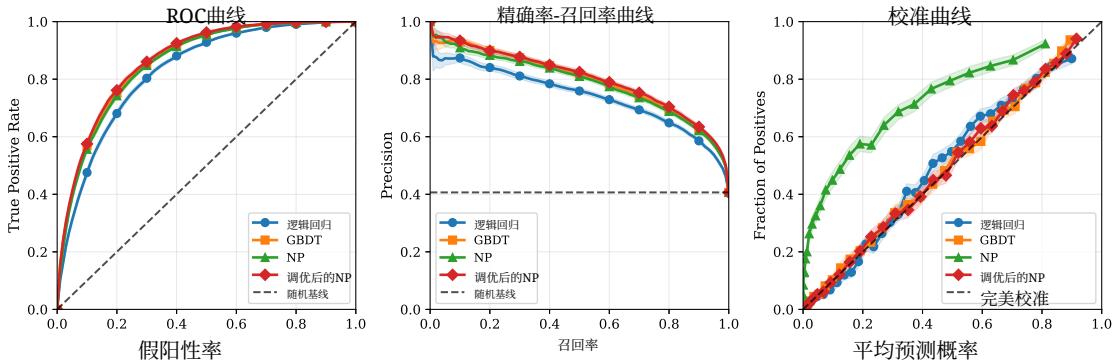
校准良好的概率的变体（调优后的NP）。该变体使用一个额外的分类头，对预测模型最终层生成的编码进行处理，以直接预测切换准备就绪状态。补充材料第S2节将进一步讨论校准问题。

为了评估模型预测患者是否已准备好转换的能力，我们采用四项指标：AUROC、平均精度（AP）、布里尔分数和前5项精确率（见补充材料S6中的详细定义）。选择这些指标是为了与模型的预期部署方式相契合：生成一份排序列表，列出最有可能已准备好接受审查的患者。AUROC从整体上衡量模型区分正类和负类的能力，表示随机选取一名正类患者的排名高于随机选取一名负类患者的概率。我们也报告平均精度，因为与AUROC不同，它对模型在正类（具备转换条件）上的表现尤为敏感，而正类所占比例较低。较高的平均精度分数表明，模型预测概率排序列表的前列聚集了较多真正例。纳入布里尔分数是为了评估概率分数本身的准确性和校准度；较低的布里尔分数表示模型预测概率能够可靠地反映事件发生的真实概率 [19]。前5项精确率指标表示测试集中每一天模型排名前5的患者中，实际已准备好转换的患者所占比例，并对所有日期取平均。该指标旨在反映模型的部署场景：每天定期向临床医生呈现一份排序列表。

表3展示了各模型在UCLH和MIMIC数据测试集上的性能；统计显著性通过自助法重采样确定（见补充材料S8）。这些结果表明，梯度提升决策树基线和神经过程模型均能够在两个数据集上准确地对患者进行排序。关于NP方法优势的讨论见第3节。对于随机模型，我们预期AUROC为0.5，平均精度和前5项精确率等于正例所占比例。前5项精确率是与预期临床部署最为契合的指标，表明神经过程模型在UCLH数据上会比随机选择选出 $2.2\times$ 更多相关患者，在MIMIC数据上则 $3.2\times$ 。在UCLH数据上，NP调优模型在所有指标上得分最高，不过就平均精度和前5项精确率而言，梯度提升决策树的表现在统计上无法区分。在MIMIC数据上，离散程度更大：标准神经过程



(a) MIMIC



(b) UCLH

图3：两个数据集的模型性能曲线对比。从左到右依次为受试者工作特征曲线、精确率-召回率曲线和校准曲线。前两者中的虚线表示随机模型的性能，最后一者中的虚线表示完美模型的性能。对于校准曲线，聚合分箱是预测概率分布的分位数；MIMIC 中正例占比较小，因此曲线未覆盖图表的完整范围。阴影区域表示 95%置信度。模型缩写：GBDT（梯度提升决策树）、NP（神经过程）。

模型的AUROC显著更优，而在其他指标（平均精确率、前5项精确率、布里尔分数）上，梯度提升决策树与神经过程的表现相近。标准神经过程模型在两个数据集上校准较差，表现为布里尔分数偏高；该问题在补充材料S2中有进一步讨论。与预测情形相同，我们发现UCLH数据更具可预测性。所有模型在筛选相关患者方面的优异表现验证了总体方法的有效性，即通过生理指标预测并应用临床标准来对患者进行IVOS审查排序，而不是从历史切换决策中学习。

图3展示了各模型在两个数据集上的性能曲线。这些曲线再次印证了这一发现：梯度提升决策树和神经过程模型在两个数据集上对患者的排序性能相近，而神经过程模型在 MIMIC 数据集上略有提升。校准曲线清楚地显示，标准神经过程的校准效果较差，该模型始终低估概率。包括经过调优的神经过程模型在内的所有其他模型，在两个数据集上都具有良好的校准效果。我们还验证了神经过程模型和基线方法在不同预测时域上的排序性能，结果如图4所示。在不同预测时间范围内，各模型的排序保持一致，神经过程模型和梯度提升决策树均表现出色。请注意

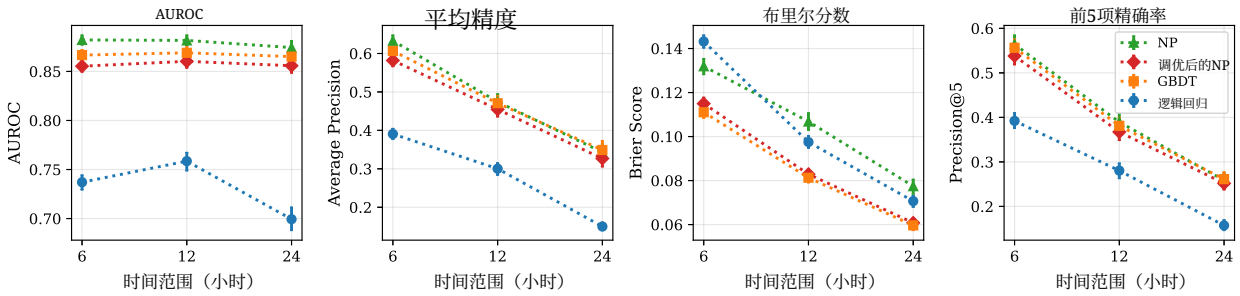


图4：随着NP模型和基线方法的预测时域不断增加，各变量在MIMIC数据上的排序性能。横轴采用对数刻度。误差线表示95%置信度。请注意，随着预测时域增加，各类别变得更加不平衡：在6、12和24小时时域中，正例分别占实例总数的20.1%、12.2%和7.8%。模型缩写：GBDT（梯度提升决策树），NP（神经过程）。

尽管问题难度增加，布里尔分数仍会随着预测时域的增加而降低，这是因为正例所占比例减少了。

我们的主要关注点是模型在将患者排序、找出最适合转换的患者这一任务上的表现，因为这与模型预期的部署方式最为一致。排序不需要设定二分类决策阈值，而且我们无意让模型直接决定是否转换。尽管如此，比较模型预测这一二分类类别的能力仍可能具有参考价值。表4列出了用于评估模型二分类预测表现的多项指标。我们通过计算验证集上F1分数最高时对应的值，为每个模型设定决策阈值。在UCLH数据上，梯度提升决策树和NP调优模型的准确率和F1分数均表现最佳。在MIMIC数据上，标准神经过程模型的F1分数最高，而Repeat和神经过程模型的召回率均表现最佳。梯度提升决策树基线的精确率通常显著更高。与预测和排序结果一致，模型在MIMIC数据集上的表现较差；正类患病率较低，导致所有模型的精确率均显著下降。

3 讨论

为说明所提出的系统在临床实践中的运行方式，我们考察了测试集中的一条患者就诊记录。图5展示了这次就诊的进展情况。补充材料S7.2中还提供了另外两个示例。在约两周的时间里，患者通过多张独立处方接受了五种不同的抗生素药物。随着就诊进展，患者的生命体征持续改善，尤其是心率、体温和血氧饱和度。模型预测的切换准备度概率与这一临床变化轨迹高度吻合：在初期患者病情较重时，该概率保持较低；随着生理稳定性恢复，该概率则大幅上升。值得注意的是，尽管患者实际上从未转换为口服治疗，模型仍判定患者在就诊结束前已做好转换准备，这表明若部署了所提出的系统，患者本可以安全地更早转换治疗。

虽然这一示例展示了模型的良好表现，但也揭示了系统必须应对的静脉转口服抗生素治疗问题本身固有的复杂性。生命体征的采样时间不规律，存在较长时间的数据缺失情况和频繁出现的瞬时尖峰，因此必须以稳健的方式加以处理。

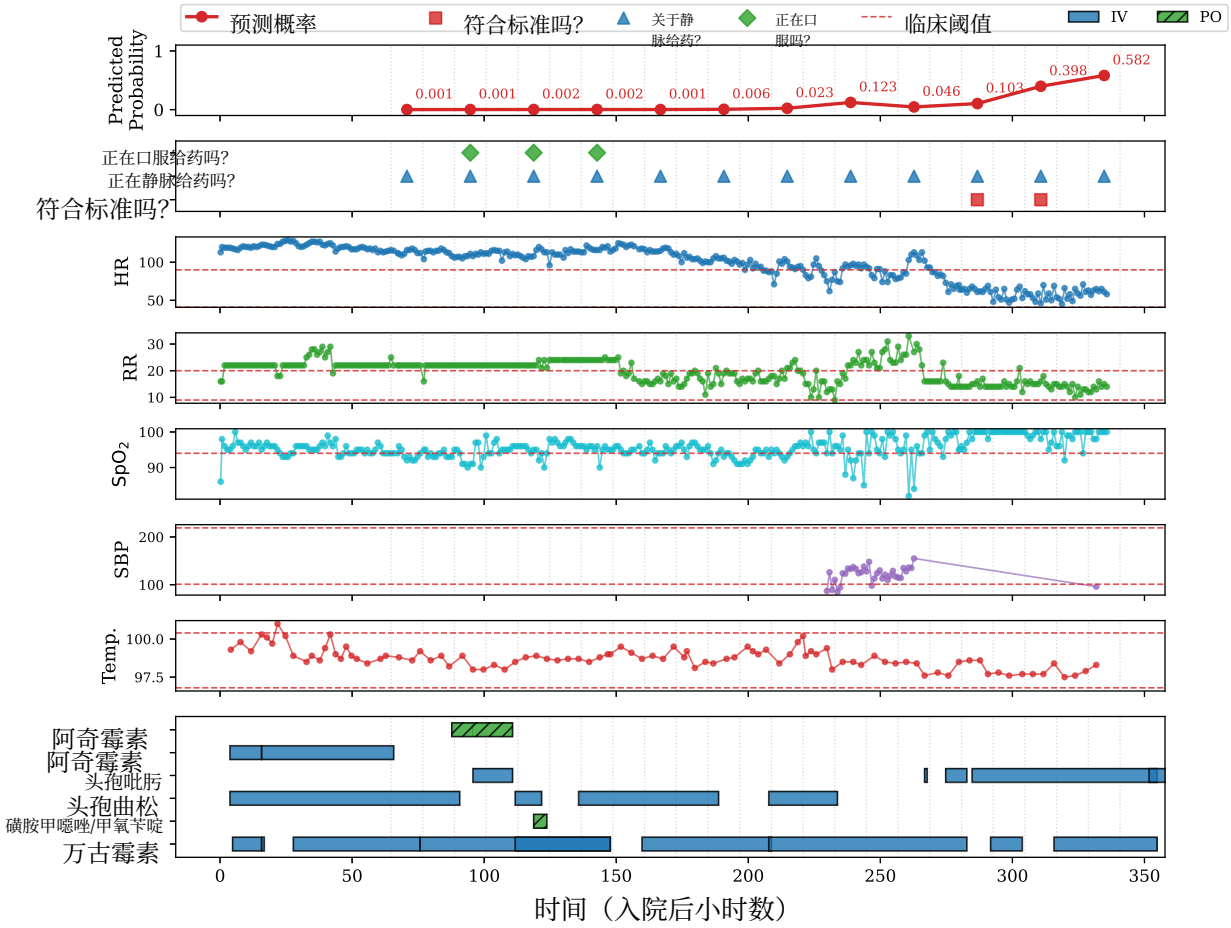


图5：MIMIC测试集中的一个就诊记录示例，展示了系统预期的运行方式。各面板从上到下依次显示：NP调优模型预测的转换准备度概率；标示口服（PO）和静脉给药（IV）抗生素处方，以及每个预测窗口内是否满足临床标准的标签；测得的生命体征，并叠加显示转换准备度的临床阈值；以及抗生素处方时间线，其中颜色和图案表示给药途径。灰色垂直线标示每日的12小时预测窗口。预测从入院48小时后开始，以确保有足够的历史数据用于预测。变量缩写：HR（心率）、RR（呼吸频率）、SpO_2（血氧饱和度）、SBP（收缩压）、Temp.（体温）。抗生素缩写：Sulfameth/Trimeth（磺胺甲噁唑-甲氧苄啶）。

数据集	模型	指标			
		F1 分数	准确率	精确率	召回率
UCLH	重复	0.704 [0.699-0.709]	0.700 [0.696-0.704]	0.587 [0.581-0.594]	**0.879 [0.874-0.884]**
	逻辑回归	0.719 [0.714-0.724]	0.736 [0.732-0.740]	0.634 [0.627-0.640]	0.832 [0.827-0.838]
	GBDT	**0.744 [0.740-0.749]**	**0.765 [0.761-0.769]**	**0.667 [0.661-0.673]**	0.843 [0.837-0.848]
	NP	0.741 [0.736-0.746]	0.754 [0.751-0.759]	0.649 [0.643-0.655]	0.863 [0.857-0.868]
	调优后的NP	**0.748 [0.744-0.754]**	**0.764 [0.760-0.768]**	0.660 [0.654-0.666]	0.864 [0.859-0.870]
MIMIC	重复	0.432 [0.419-0.446]	0.778 [0.773-0.784]	0.315 [0.303-0.327]	**0.691 [0.673-0.708]**
	逻辑回归	0.372 [0.359-0.385]	0.763 [0.758-0.769]	0.275 [0.264-0.287]	0.575 [0.557-0.594]
	GBDT	0.497 [0.482-0.511]	**0.861 [0.856-0.865]**	**0.444 [0.428-0.460]**	0.564 [0.545-0.581]
	NP	**0.519 [0.505-0.531]**	0.845 [0.840-0.850]	0.418 [0.404-0.432]	**0.683 [0.666-0.701]**
	调优后的NP	0.494 [0.479-0.508]	0.844 [0.839-0.848]	0.409 [0.393-0.423]	0.626 [0.608-0.644]

表4：二元分类模型性能比较，其中正类表示已适合转换的患者。UCLH 和 MIMIC 中正实例的比例分别为 40.2% 和 12.2%，这表明预测多数类在两种情况下的准确率分别可达 59.8% 和 87.8%。决策阈值的选择以验证集上的 F1 分数最大化为准。括号内为 95%置信区间。粗体数值表示性能最佳的模型（$p \geq 0.05$）。所有指标均为数值越高越好。模型缩写：GBDT（梯度提升决策树）、NP（神经过程）。

抗生素处方模式很少遵循简单的序列，即以一种口服对应药物替代静脉给药药物；在实际临床中，可能会同时开具多种药物用于不同适应证，因此仅凭处方数据很难识别转换。开具口服药物并不一定意味着患者病情正在好转。判断患者是否符合临床标准同样具有挑战性：某项测量值略微超过阈值，就可能导致患者被标记为不符合标准；例如，在最后一个预测窗口中，患者的心率曾短暂升高。还需注意的是，两名患者的生理测量值可能完全相同，但其中一名适合转换，另一名则不适合；潜在适应证或口服途径耐受性方面的问题，都可能使转换不恰当。

尽管存在这些复杂性，该模型仍能产生符合临床实际的预测，这些预测能够追踪患者的生理变化轨迹，并有助于及时识别可切换患者。这些挑战促使我们采用第4节中讨论的设计选择，尤其是用于处理不规则测量值的聚合策略，以及选择预测生命体征而非直接预测转换。

本研究的一项关键贡献在于其 formulation 本身，这一 formulation 不同于此前提出的将机器学习用于静脉转口服抗生素治疗决策支持的方法 [14]。神经过程模型和梯度提升决策树模型采用我们的 formulation 后，结果均显著优于随机选择，表明无论采用何种具体模型，预测生理指标都能有效识别需要进行静脉转口服抗生素转换审查的患者。神经过程模型和梯度提升决策树模型在排序性能方面表现相近，这就引出了一个问题：为什么还要使用概率预测？我们强调，我们的目标是创建一个实用的机器学习系统，而不仅仅是构建一个性能优异的模型。可解释性是任何基于机器学习的临床决策支持系统成功应用的关键因素[20]。临床医生必须信任模型提出的建议，并了解这些建议是如何生成的。虽然神经过程模型本身不可解释，但该系统通过其输出和架构提供了可解释性。系统会预测近期未来的生命体征，而这些指标是临床医生已经充分理解的；随后，系统将透明的

临床标准应用于这些预测，以确定患者是否具备转换条件。当患者被标记为具备转换条件时，临床医生可以查看底层生命体征预测，从而了解系统作出判断的原因。

NP方法的另一项结构性优势在于，单个模型即可同时用于预测和分类任务。由于NP会生成未来生命体征数值的完整预测分布，因此可将这些分布直接应用于临床标准，从而得出切换就绪概率，进而为需要复核的患者提供有依据的排序。相比之下，基线模型用于预测时只能给出点估计，因此无法生成评估患者生命体征保持在正常范围内的概率所需的分布输出。因此，需要两个完全独立的梯度提升决策树模型：一个用于预测，一个用于分类，且二者分别针对不同目标和特征表示进行训练。除了维护两个独立模型所带来的实际负担外，这种分离还意味着分类模型的预测并未以患者生理变化轨迹的明确预测为依据，从而削弱了预测生命体征与临床医生会查看的转换准备度评分之间可解释的联系。

将预测与标准的应用分离后，模型在部署时可以更加灵活地使用。以接受机械通气的患者为例，呼吸机可以控制呼吸频率，可能导致患者看起来已准备好转换，而实际上仍未康复。在这种情况下，测得的呼吸频率可能处于正常范围，并非因为患者有所好转，而是因为呼吸机在人工调节其呼吸。采用预测方法时，可以非常方便地在转换准备度计算中排除呼吸频率，或者根据具体情况调整阈值。同样，如果有新的证据导致建议的临床阈值发生变化，也可以调整所提出的系统，而无需重新训练或调整基础模型。这在考虑模型部署时非常有益，因为定期更新和重新训练模型会增加显著的复杂性 [21]。

据我们所知，目前只有一项研究讨论了将机器学习应用于静脉转口服抗生素转换问题，以评估患者是否适合进行转换 [14]。该研究的方法是训练一个模型，针对患者就诊记录中的某一天，预测该患者是否接受了静脉抗生素治疗。通过训练模型预测历史临床行动（即静脉抗生素的使用），所得系统必然会学习复制训练数据中现有的静脉转口服抗生素治疗决策模式及其频率。临床实践正在不断变化，并且越来越多的证据表明，对于许多疾病，长疗程静脉抗生素治疗并无必要 [2, 6, 22]。如果临床决策支持系统的目标是改进当前实践——在本例中，即鼓励更早进行转换，从而减少静脉治疗总天数——那么，训练一个用于模仿历史上次优行为的模型，其能力必然受到限制。对于复杂或非典型病例而言，这一局限性尤为突出，而这些病例正是最需要临床决策支持的情形。尽管此类模型在留出的回顾性数据上的表现可能看起来令人鼓舞，但这种方法不可避免地会产生一个脆弱的临床决策支持系统：它被定格在过去，无法为参与构建具有韧性的学习型卫生系统的临床医生提供持续支持 [23, 24]。我们通过对患者生理状况进行建模，将处方临床医生置于决策核心，从而构建出足够灵活的临床决策支持系统，使其能够适应临床证据和不断变化的指南。

从根本上说，获取转换准备度的真实标签颇具挑战。转换决策涉及多方面因素，除生命体征外，还包括感染类型和严重程度、抗生素的供应情况及禁忌证、患者合并症，以及影响出院计划的社会状况。电子健康记录中记录的历史切换决策反映的是有待改进而非照搬的欠佳实践，因此不适合作为真实标签。即便不考虑从既往实践中学习的问题，从电子健康记录中构建可靠的静脉转口服治疗标签也很困难，而标签质量会限制模型性能和临床效用。Dutey-

Magni 等人 [25] 讨论了仅凭处方数据重建抗生素疗程的复杂性。

Bolton 等人 [14] 采用的标签构建方法存在一些方法学问题，导致数据集噪声较大且存在系统性偏差。在单次住院期间，某一给药途径下所有抗生素处方的最早开始时间和最晚结束时间被用来界定静脉给药和口服治疗的时段，这可能导致转换被错误标记，并系统性地排除较复杂但常见的临床情形，使最终数据集偏向不切实际、线性的治疗轨迹。此外，所有抗生素都被视为可以互换，既未核实口服药物是否适合替代静脉药物，也未确认处方适应证是否针对同一病症。例如，使用广谱静脉抗生素可能导致艰难梭菌感染 [26], ，严重时会停止静脉疗程，改用口服抗生素（通常为万古霉素）控制艰难梭菌感染；若不考虑适应证，这种情况会被标记为一次转换，但实际上并未发生转换。

我们的系统并不依赖真实转换标签，因此必然不完整，仅捕捉决策中的生理状况改善部分。这一替代指标是否恰当，只能通过前瞻性部署加以验证；届时，临床医生会审查被标记的患者，并就建议的相关性与时机提供反馈。我们指出，仅对生理因素进行建模是有意设计，我们的系统有意将最终转换决策交由临床医生掌握。对临床医生行为和医院政策进行建模困难重重，并会导致模型强化次优实践。在考虑基于机器学习的临床决策支持系统时，这一问题至关重要。

我们这项工作的主要局限是缺乏前瞻性临床评估。如果不在实践中部署该系统，并衡量其对转换率、静脉治疗总天数、患者结局和临床医生决策的影响，我们就无法确切评估该模型的预测能否转化为更好的临床护理。我们的评估表明，该模型能够依据转换准备度的生理标准对患者进行准确排序，但回顾性性能指标并不能保证临床效用。这一局限在静脉转口服抗生素治疗临床决策支持系统文献中普遍存在：据我们所知，尚无用于静脉转口服抗生素转换的基于机器学习的系统经过前瞻性评估 [14], 尽管Bolton 等人 [27]开展了一项试验，以衡量基于机器学习的静脉转口服抗生素治疗建议对一小套精心整理的病例简介中临床医生决策的影响。虽然基于规则的系统已经过前瞻性评估 [10, 11, 12], 但这些研究通常将干预与教育和政策变更结合在一起，因此难以单独区分临床决策支持系统本身的贡献。

模型依赖五项生命体征，这意味着未纳入其他可能具有临床相关性的信息。考虑更广泛的输入变量或许能提升预测性能，例如炎症标志物（C反应蛋白、白细胞计数），此前的研究已考虑过这些指标 [28, 29]。之所以选择这组简单的五项生命体征，是因为在拟议的静脉转口服抗生素转换系统可能适用的临床环境中，这些指标的测量频率足够高。该预测模型还假设近期信息能够反映近未来的情况，但在临床状况迅速恶化或好转时，这一假设可能不成立，进而导致未能及时发现状态变化。最后，如补充材料S2所述，标准神经过程模型存在校准问题，这是高斯似然与电子健康记录数据中的异常值相互作用所致。虽然调优变体解决了这一局限，但仍需进行此类修正，凸显了对复杂电子健康记录数据建模的难度。

4 方法

本节详细介绍数据源和预处理流程、构建切换准备度标签的方法、用于预测的神经过程架构，以及作为比较对象的基线模型。

4.1 数据

为训练和验证模型，我们使用了两个由常规收集的电子健康记录构成的数据库。第一个是MIMIC-IV v3.1 [16], ，这是一个公开可用且已去标识化的数据库，收录了2008年至2022年间美国波士顿 Beth Israel Deaconess Medical Center 重症监护病房（ICU）的入院记录。第二个数据库来自伦敦大学学院医院（UCLH）NHS基金会信托的内部数据库，涵盖英国伦敦的11家医院，时间跨度为2019年至2025年。

这两个数据集代表了差异显著的人群：MIMIC仅包含重症监护病房入院病例，因此其中的患者通常病情严重得多；MIMIC数据中患者生命体征的测量频率也高得多，抗生素处方的特征在给药途径和药物种类上均有所不同。这些数据集来自不同国家，当地的医疗实践，尤其是抗生素开具方面的做法各异；表1展示了各数据集的特征。例如，两者的性别构成不同（MIMIC中女性占40.7%，而UCLH中占52.6%），抗生素开具模式也有明显差异：万古霉素是MIMIC中最常见的静脉抗生素（占入院记录的65.7%），但在UCLH很少使用，而阿莫西林-克拉维酸则占主导地位（46.5%）。此外，两地的监测强度也不同：与UCLH相比，MIMIC中的心率、呼吸频率和血氧饱和度测量频率约为其两倍；收缩压的差异则较小，而MIMIC中体温的记录频率实际上更低。

从每个数据库中，我们提取三类信息：生命体征、住院信息和抗生素处方。生命体征包括五项关键生理测量值：心率、呼吸频率、血氧饱和度（SpO_2）、收缩压和体温。这些数据在患者住院期间以不规则的时间间隔记录。住院信息包括入院和出院时间。抗生素处方包括抗菌药物医嘱记录，涵盖药物名称、给药途径（静脉给药或口服）以及每项医嘱的开始和结束时间。

现在，我们将数据结构和任务定义形式化。设 $\mathcal{X} = \{X_1, X_2, \ldots, X_M\}$ 为包含 M 次相互独立的医院就诊记录的数据集。对于每次就诊记录 X_m，我们有一组 N_m测量值，这些测量值是在患者住院期间采集的。每项测量值对应C 个变量之一，在本例中即上述五项生命体征。我们采用三元组表示法，因此每次就诊记录可表示为集合 $X_m = \{(t_{m,\,i},\ c_{m,\,i},\ x_{m,\,i})\}_{i=1}^{N_m}$，其中 $t_{m,\,i} \in \mathbb{R}^+$ 是变量 $c_{m,\,i} \in \{1, \ldots, C\}$ 的测量时间，而 $x_{m,\,i} \in \mathbb{R}$ 是测量值。设 $X_m^\tau = \{(t',\ c',\ x') \in X_m | t' < \tau\}$ 表示时间 τ之前的就诊记录历史。预测任务是根据 X_m^τ预测患者在未来某个时间 $t > \tau$ 的状态。分类任务是根据历史记录 X_m^τ，预测一个二元标签 $y_m^\tau \in \{0, 1\}$ ，该标签表示患者在预测窗口 $[\tau, \tau + \Delta]$内是否符合转换准备度的临床标准。第 4.2 节将讨论确定这些标签的方法。

我们采用以下预处理步骤：剔除超出预定义合理范围的测量值，具体范围见补充材料S2。将每条就诊记录的时长截断为最多2周。MIMIC 数据集已排除儿科患者；对于 UCLH 数据，排除未满18岁的患者。作为提取 UCLH 数据所需假名化的条件，移除了年龄超过90岁的患者。全文中，我们使用相对于

索引就诊入院时间的相对时间。除非另有说明，模型输入采用48小时的病史。为匹配模型的预期部署方式，我们为每条就诊记录每天构建一个任务，并将预测时间设为上午9点。只有在预测时间患者被认为正在接受静脉处方、此前48小时的回溯窗口内至少有10个测量值，且12小时预测窗口（上午9点至晚上9点）内至少有一个测量值时，才会在当天创建任务。如果患者在预测时间有有效的静脉抗生素医嘱，或静脉医嘱在此前36小时内结束，则认为患者正在接受静脉用药。确实存在更复杂的抗生素疗程重建方法[25], ，但本研究并未采用。

为模拟真实的部署场景并防止前视偏差，我们按时间划分数据以构建测试集。对于 UCLH 数据，2024年1月1日之后入院的就诊记录构成测试集。对于 MIMIC 数据，我们以 2019年1月1日为截止日期；不过，MIMIC-IV 的去标识化流程会为每位患者的所有日期加上随机偏移，从而打乱不同患者入院日期的时间顺序。因此，我们按照补充材料 S3.1 所述，重建近似的入院日期，以便进行时间划分。由于重建的入院日期只是近似值，因此与 UCLH 数据相比，解读 MIMIC 数据的前5项精确率指标时应更加谨慎。我们从非测试数据中选取 10% 创建验证集，其余数据构成训练集。表 S2 列出了各数据集训练集、验证集和测试集中任务数量的完整细分。

基线分类模型使用如上所述构建的任务进行训练和评估。然而，为了训练和评估预测模型，我们采用不同的方式构建任务，以更好地体现各模型的一般预测能力，而非特定的转换准备度预测场景。对于每次就诊记录，我们通过均匀抽样预测起始时间 τ来创建任务。对于每个 τ，我们使用 48 小时的回溯窗口作为上下文，预测 τ之后 12 小时预测时域内的所有测量值。我们按就诊记录的时长成比例抽样任务，每 24 小时的数据生成一个任务，因此，7 天的就诊记录会生成 7 个预测时间随机分布的任务。这遵循了训练神经过程时构建任务的标准做法 [30]。

4.2 转换标注

我们根据患者的生命体征在预测窗口 $[\tau, \tau + \Delta]$内是否处于临床正常范围，定义转换就绪度标签。对于每项生命体征 c，我们设定代表正常范围的上下限阈值；这些阈值见补充材料 S3.3。在静脉转口服抗生素转换的背景下，临床正常范围存在多种不同定义 [8]。我们咨询了本机构的临床医生，以确定他们认为患者适合接受审查的水平，并在 MIMIC 和 UCLH 数据集上统一采用这些水平。在不同情境下，适用的水平可能有所不同；在我们的框架中调整这些水平很简单，调整后符合标准的患者比例也会随之变化。为考察结果对临床阈值变化的敏感性，我们使用较宽松的阈值重新运行了 MIMIC 排名实验，使阳性（具备转换条件）天数的比例从 12.2% 增至 50.0%。我们发现，在各项排序指标（AUROC、平均精确率、布里尔分数和前5项精确率）上表现最佳的模型几乎完全相同。完整结果见 S7.1。

构建这些标签时，一个关键考量是，测量值的采集时间无法预先得知。在模型部署期间，必须在固定的评估时间点进行预测，而不能预先知道未来何时会进行测量。若在实际测量时间点评估标准，就会引入前视偏差，因为这些时间点只有在事后才能获知

。为解决这一问题，我们将预测窗口内的测量值汇总到固定网格上。我们将12小时预测窗口划分为四个3小时区间，并计算每个区间内各项生命体征的中位数值。如果患者（$y_m^\tau = 1$）的所有汇总值均处于各自的正常范围内，则标记为具备转换条件。如果某个区间内没有某项生命体征的测量值，则在检查该变量是否符合标准时排除该区间。将缺失值视为符合标准而非不符合标准，是一项建模决策，必须向终端用户明确说明；另一种做法会导致类别分布明显更加失衡，因为缺失数据很常见。

4.3 神经过程

神经过程（NP）是一种概率元学习模型，它利用神经网络对预测映射进行参数化，将可能的数据集空间映射到随机过程空间；这些随机过程表示模型在任意一组目标点上的概率预测[17, 30]。虽然上述定义在技术上准确无误，但对于理解神经过程如何用于生命体征预测而言，其作用或许有限。用元学习的术语来说，我们可以将第 4.1 节所述的每条就诊记录视为一个独立任务或数据集。所有就诊记录构成的数据集称为元数据集。我们的目标是构建一个模型，使其在获得来自未见过的就诊记录的数据后，能够预测患者状态的值。神经过程通常通过获取以往任务的数据集，并将每个任务划分为上下文集和目标集来解决这一问题。随后，模型经过训练，利用上下文集中的信息，对一组目标输入生成校准良好的概率预测。在本例中，上下文集是回溯窗口内所有生命体征测量值，而目标点则是预测窗口内的测量值。

卷积条件神经过程（ConvCNP）是一类将平移等变性编码到预测映射 [31]中的神经过程。平移等变性意味着，如果数据的时间值平移一个常数时间，模型的预测也会平移相同的时间。平移等变的神经过程对应于平稳随机过程，因此对于以平稳性作为合理归纳偏置的时间或时空问题，它们是一类有用的模型 [30]。ConvCNP采用基于输入数据核平滑的函数编码，随后由卷积神经网络进行处理。模型架构、训练目标以及从预测分布推导切换就绪概率的说明见补充材料S4。

我们的ConvCNP采用UNet架构 [32], , 在瓶颈层前后各有4个卷积层，核平滑的长度尺度为1小时。在预测窗口内的每个目标时间 t 以及每项生命体征 c 上，模型输出高斯预测分布的参数；在给定上下文数据 X_m^τ的条件下，这些预测是条件独立的。模型通过最小化目标观测值在该预测分布下的负对数似然进行训练。我们最多训练500轮，并采用早停（耐心值为250轮），批量大小为512。我们对学习率和通道数进行网格搜索，选择验证损失最低的模型。完整超参数设置见补充材料S4。所有模型均在单个NVIDIA A10或NVIDIA A40 GPU上训练。

为计算切换就绪概率，我们考虑两种方法。在标准神经过程模型中，对于每种生命体征，我们都会在预测时间生成预测分布，这些预测时间与第 4.2 节所述 3 小时聚合区间的中心对齐。随后，我们使用高斯分布的累积分布函数，计算每个时间点的测量指标落在临床阈值范围内的概率。在条件

独立假设下，所有生命体征和时间点对应的这些概率的乘积即为总体切换就绪概率。在调优变体（调优后的NP）中，我们在预测模型中添加一个分类头。该分类头通过最大池化操作，沿时间维度聚合最后一个卷积层的编码，然后使用多层感知机输出单一的切换准备度概率。训练分类头时，预测模型的权重保持冻结。分类头的架构详情见补充材料 S5。

4.4 基线方法

我们针对预测和分类任务分别开发了基线模型。这些基线方法采用表格模型，其输入是由回溯窗口 X_m^τ 得到的固定长度特征向量。对于每种生命体征 $c \in \{1, \ldots, C\}$，我们从X_m^τ中的测量值提取16个特征：统计摘要（均值、标准差、最小值、最大值、中位数、计数、第25和第75百分位数）、时间特征（距最近一次测量的时间、观测值的时间跨度和测量频率），以及趋势特征（连续斜率的均值和标准差、从首次观测值到末次观测值的总变化量，以及正向变化和负向变化的计数）。如果某种测量类型对于给定任务完全缺失，则使用默认值填补其对应特征。我们还计算了两个描述缺失情况的全局特征：缺失测量类型的数量和总体完整率。分类任务总共有82个特征。

分类基线方法直接根据特征向量预测切换准备度标签 y_m^τ。我们采用逻辑回归（使用 scikit-learn 实现） [33], 和梯度提升决策树（GBDT，使用 LightGBM 实现） [18]。我们对超参数进行网格搜索，并选择在留出的验证集上平均精度最高的模型；超参数搜索空间见补充材料 S6。与基于 NP 的系统相比，这些模型构成了更具挑战性的基线，因为它们直接优化分类目标，而不是先预测生命体征再应用标准。

对于预测任务，我们使用两种基线方法。重复基线将预测窗口之前每项生命体征最近一次的测量值，作为所有未来时间点的预测值。GBDT 预测基线更为复杂，必须处理目标时间的不规则性以及多个输出变量。预测窗口中的每个时间点都被视为一个独立的预测实例。表示回溯窗口的共享特征向量会附加独热编码，用以指明要预测的生命体征 c，以及自预测时间 τ 以来经过的时间（即预测时域）。这种表述使单个 GBDT 模型能够预测预测窗口内任意未来时间点的所有生命体征。两种基线都无法生成概率预测，因此难以量化预测的不确定性。

贡献声明

M.R.构思并设计了这项研究。M.R.开发了模型架构、代码和可视化内容。S.H.、A.L.、E.M和N.S.提供了临床指导并进行了验证。V.L.、S.H.和I.J.C.提供了建模专业知识，并就实验提供指导。S.H.和A.L管理了数据治理和访问审批。M.R.撰写了论文初稿。M.R.、V.L.、S.H.、I.J.C和A.L.对论文提出了反馈并进行了编辑。V.L、S.H.和I.J.C.提供了监督。所有作者均批准了论文的最终版本。

致谢

M.R.、I.J.C.和V.L.获得EPSRC抗微生物药物耐药数字健康中心（EP/X031276/1）的资助。S.H.和A.L.获得国家健康与护理研究所伦敦大学学院医院生物医学研究中心的资助。

代码可用性

本研究中使用的计算机代码可在
https://github.com/SAFEHR-data/ivos-model-public 获取。

数据可用性

MIMIC-IV数据集可在 https://physionet.org/content/mimiciv/3.1/ 获取。UCLH数据集包含个人身份信息，尚未获准对外发布。希望使用UCLH数据集验证或复现研究结果的研究人员，需获准与伦敦大学学院医院国民保健署基金会信托开展研究合作，并取得UCLH/UCL联合研究办公室的适当许可。符合这些要求的研究人员可联系通讯作者，以进一步了解如何获取这些数据集。

参考文献

[1] Jissa Maria Cyriac 和 Emmanuel James。从静脉治疗转换为口服治疗：简明概述。《药理学与药物治疗学杂志》，5(2):83–87，2014年。doi: 10.4103/0976-500X.130042.[2] Noah Wald-Dickler、Paul D. Holtom、Matthew C. Phillips、Robert M. Centor、Rachael A. Lee、Rachel Baden 和 Brad Spellberg。口服疗法就是新型静脉疗法。挑战血液和骨感染领域数十年来的教条：系统综述。《美国医学杂志》，135(3): 369–379.e1，2022年。doi: 10.1016/j.amjmed.2021.10.007.[3] Marie-C´eline Zanella、Gaud Catho、Holly Jackson、Nasim Lotfinejad、Val´erie Sauvan、Marie- No¨elle Chraiti、Walter Zingg、Stephan Harbarth 和 Niccol`o Buetti。外周静脉导管的留置时间与血流感染风险。《*JAMA Network Open*》，8(4): e257202，2025年。doi: 10.1001/jamanetworkopen.2025.7202.[4] Abi Jenkins。静脉给药转为口服：一种新颖的观点。《抗菌化疗杂志》，78(10):2603–2604，2023年。doi: 10.1093/jac/dkad239.[5] 健康改善与差异办公室。公共卫生概况：在符合静脉给药转口服标准后仍继续接受静脉注射抗生素的患者百分比。
https://fingertips.phe.org.uk/profile/amr-local-indicators，2025年。 [访问日期：08 September 2025]。[6] Ho-Kwong Li、Ines Rombach、Rhea Zambellas、A. Sarah Walker、Martin A. McNally、Brid- get L. Atkins、Benjamin A. Lipsky、Harriet C. Hughes、Deepa Bose、Michelle K¨umin、Claire

Scarborough、Philippa C. Matthews、Andrew J. Brent、Jose Lomas 等。骨与关节感染的口服抗生素与静脉抗生素治疗。《新英格兰医学杂志》，380(5): 425–436，2019年。doi: 10.1056/NEJMoa1710926。

[7]英国卫生安全局。成人尽早转换的全国抗菌药物静脉转口服标准。 https://www.gov.uk/government/publications/ antimicrobial-intravenous-to-oral-switch-criteria-for-early-switch/ national-antimicrobial-intravenous-to-oral-switch-ivos-criteria-for-early-swi tch，2025年。 [访问日期：12月10日， 2025]。[8] Eleanor J Harvey、Monsey McLeod、Caroline De Br´un 和 Diane Ashiru-Oredope。《住院成人群体中实现抗菌药物安全静脉转口服切换的标准：系统性快速综述》。《英国医学杂志开放版》，13(7):e068299，2023年。doi: 10.1136/bmjopen-2022-068299。[9] P F Dutey-Magni、M Brown、S Harris、C Curtis、R Dobson、H Chowdhury、A Cawthorn、S De、N Stone、J Cooper 和 L Shallcross。通过监督式机器学习实时识别需要调整抗生素治疗的住院患者：临床前诊断评估与可行性研究。《开放论坛：传染病》，13(1):ofaf721，2025年。ISSN 2328-8957。doi: 10.1093/ofid/ofaf721。[10] Patrick E. Beeler、Stefan P. Kuster、Emmanuel Eschmann、Rainer Weber 和 J¨urg Blaser。借助电子提醒，提前将抗生素从静脉给药切换为口服。《Interna- tional Journal of Antimicrobial Agents》，46(4):428–433，2015年。doi: 10.1016/j.ijantimicag.2015. 06.013。[11] Marvin A. H. Berrevoets、Johannes (Hans) L. W. Pot、Anne E. Houterman、Anton (Ton) S. M. Dofferhoff、Marrigje H. Nabuurs-Franssen、Hanneke W. H. A. Fleuren、Bart-Jan Kullberg、Jeroen A. Schouten 和 Tom Sprong。用于优化抗生素静脉转口服切换的电子触发工具：一项对照中断时间序列研究。《抗菌药物耐药与感染控制》，6(1):81，2017年。doi: 10.1186/s13756-017-0239-3。[12] Tiffany Kan、Derrick Kwan、Thomas Chan、Pavani Das 和 Sumit Raybardhan。在社区教学医院实施临床决策支持工具，以提高抗生素静脉转口服转换率。《加拿大医院药学杂志》，72(6): 455–461，2019年。[13] Charlotte Quintens、Marie Coenen、Peter Declercq、Minne Casteels、Willy E Peetermans 和 Isabel Spriet。从基础到高级：对乙酰氨基酚和抗生素静脉转口服切换的计算机化方案：一项中断时间序列分析。《英国医学杂志开放版》，12(4):e053010，2022年。doi: 10.1136/bmjopen-2021-053010。[14] William J. Bolton、Richard Wilson、Mark Gilchrist、Pantelis Georgiou、Alison Holmes 和 Timothy M. Rawson。通过公平且可解释的机器学习实现个性化静脉转口服抗生素切换决策。《自然·通讯》，15(1)，2024年。doi: 10. 1038/s41467-024-44740-2。[15] Magnus Ross、Nel Swanepoel、Akish Luintel、Emma McGuire、Ingemar J Cox、Steve Harris 和 Vasileios Lampos。利用神经过程加快抗菌药物从静脉给药向口服治疗的切换。载于《2025年健康领域机器学习会议》，2025年。URL https://openreview. net/forum?id=j8wsqyowid。

[16] Alistair E. W. Johnson、Lucas Bulgarelli、Lu Shen、Alvin Gayles、Ayad Shammout、Steven Horng、Tom J. Pollard、Sicheng Hao、Benjamin Moody、Brian Gow、Li-wei H. Lehman、Leo A. Celi 和 Roger G. Mark。MIMIC-IV：一个免费开放的电子健康记录数据集。*Scientific Data*，10(1):1，2023年。doi: 10.1038/s41597-022-01899-x。[17] Marta Garnelo、Dan Rosenbaum、Christopher Maddison、Tiago Ramalho、David Saxton、Mur- ray Shanahan、Yee Whye Teh、Danilo Rezende 和 S. M. Ali Eslami。条件神经过程。载于第*35*届国际机器学习会议论文集，第1704–1713页。PMLR，2018年。[18] Guolin Ke、Qi Meng、Thomas Finley、Taifeng Wang、Wei Chen、Weidong Ma、Qiwei Ye 和 Tie-Yan Liu。LightGBM：一种高效的梯度提升决策树。载于《神经信息处理系统进展》，第30卷，2017年。[19] Glenn W. Brier。概率表达的预报验证。《Monthly Weather Review》，78(1):1，1950年。ISSN 0027-0644。doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0. CO;2。[20] Ian A. Scott、Anton van der Vegt、Paul Lane、Steven McPhail 和 Farah Magrabi。实现人工智能赋能的决策支持在临床医生中的大规模应用。《BMJ Health & Care Informatics》，31(1)，2024年。ISSN 2632-1009。doi: 10.1136/bmjhci-2023-100971。[21] Steve Harris、Tim Bonnici、Thomas Keen、Watjana Lilaonitkul、Mark J. White 和 Nel Swanepoel。临床部署环境：健康领域转化机器学习的五大支柱。《Frontiers in Digital Health》，4，2022年。ISSN 2673-253X。doi: 10.3389/fdgth.2022. 939292。[22] Kasper Iversen、Nikolaj Ihlemann、Sabine U. Gill、Trine Madsen、Hanne Elming、Kaare T. Jensen、Niels E. Bruun、Dan E. Høfsten、Kurt Fursted、Jens J. Christensen、Martin Schultz、Christine F. Klein、Emil L. Fosbøll、Flemming Rosenvinge、Henrik C. Schønheyder、Lars Køber、Christian Torp-Pedersen、Jannik Helweg-Larsen、Niels Tønder、Claus Moser 和 Henning Bundgaard。心内膜炎的部分口服治疗与静脉抗生素治疗。《New Eng- land Journal of Medicine》，380(5):415–424，2019年。doi: 10.1056/NEJMoa1808312。[23] Charles P. Friedman。学习型卫生系统有何独特之处？《Learning Health Systems》，6(3)，2022年。ISSN 2379-6146。doi: 10.1002/lrh2.10328。[24]Amy M. Kilbourne、Amanda E. Borsky、Robert W. O'Brien、Melissa Z. Braganza 和 Melissa M. Garrido。学习型卫生系统的基础科学。《Health Services Research》，59(6)，2024年。ISSN 1475-6773。doi: 10.1111/1475-6773.14374。[25] P. F. Dutey-Magni、M. J. Gill、D. McNulty、G. Sohal、A. Hayward、L. Shallcross、Niall Ander- son、Elise Crayton、Gillian Forbes、Arnoupe Jhass、Emma Richardson、Michelle Richardson、Patrick Rockenschaub、Catherine Smith、Elizabeth Sutton、Rosanna Traina、Lou Atkins、Anne Conolly、Spiros Denaxas、Ellen Fragaszy、Rob Horne、Patty Kostkova、Fabiana Lorencatto、Su- san Michie、Jennifer Mindell、John Robson、Claire Royston、Carolyn Tarrant、James Thomas、Jonathan West、Haydn Williams、Nadia Elsay 和 Chris Fuller。利用电子健康记录开展医院抗菌药物管理分析的可行性研究。《JAC-antimicrobial resistance》，3(1):dlab018，2021年。doi: 10.1093/jacamr/dlab018。

[26]Michael Cymbal、Arjun Chatterjee、Brian Baggott和Moises Auron。艰难梭菌感染的管理：诊断、治疗与未来展望。《美国医学杂志》，137(7):571–576，2024年。doi: 10.1016/j.amjmed.2024.03.024.[27] William J Bolton、Richard Wilson、Mark Gilchrist、Pantelis Georgiou、Alison Holmes和Timothy M Rawson。人工智能驱动的决策支持对不确定的抗菌药物处方的影响：一项随机多方法研究。《柳叶刀·数字健康》，7 (11)，2025年。[28] Jingge Xiao、Leonie Basso、Wolfgang Nejdl、Niloy Ganguly和Sandipan Sikdar。Ivp-vae：使用初值问题求解器对电子健康记录时间序列进行建模。AAAI'24/IAAI'24/EAAI'24。AAAI Press，2024年。ISBN 978-1-57735-887-9。doi: 10.1609/aaai.v38i14.29534.[29] Edward De Brouwer、Jaak Simm、Adam Arany和Yves Moreau。Gru-ode-bayes：间歇观测时间序列的连续建模。《神经信息处理系统进展》，32，2019年。[30] Wessel P. Bruinsma。卷积条件神经过程。博士论文，2024年。[31] Anna Vaughan、Will Tebbutt、J. Scott Hosking和Richard E. Turner。用于局地气候降尺度的卷积条件神经过程。《地球科学模型开发》，第251–268页，2022年。doi: 10.5194/gmd-15-251-2022.[32] Olaf Ronneberger、Philipp Fischer和Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。收录于医学图像计算与计算机辅助干预 *MICCAI 2015*年，第234–241页。Springer International Publishing，2015年。[33] Fabian Pedregosa、Ga¨el Varoquaux、Alexandre Gramfort、Vincent Michel、Bertrand Thirion、Olivier Grisel、Mathieu Blondel、Peter Prettenhofer、Ron Weiss、Vincent Dubourg、Jake Vanderplas、Alexandre Passos、David Cournapeau、Matthieu Brucher、Matthieu Perrot和´Edouard Duchesnay。Scikit-learn：Python中的机器学习。《机器学习研究杂志》，12:2825–2830，2011年。[34] Saurav Jha、Dong Gong、Xuesong Wang、Richard E. Turner和Lina Yao。神经过程家族：综述、应用与展望，2023年。

补充信息

S1 基于规则的临床决策支持系统的扩展讨论

已有多项研究使用简单的电子提醒，在经过固定时间后提醒临床医生审查静脉转口服抗生素转换的适用性。Beeler等人 [10] 实施了一套系统，在患者接受静脉给药疗程超过60小时后向临床医生发出警报。要符合触发警报的条件，患者的处方剩余时间必须至少为24小时，中性粒细胞正常且无发热，并且能够吞咽。他们在一项前瞻性评估中发现，该系统使静脉治疗天数下降了17%。Berrevoets等人 [11]构建了类似工具：若患者的炎症指标显示出改善迹象，且患者能够耐受口服途径、感染类型未被归类为重症，则在初始处方后72小时触发警报。他们将该系统与面向临床医生的每周教育培训相结合，并在前瞻性评估中发现，72小时后的静脉抗生素使用下降了13%。Kan等人 [12] 实施了一套系统，针对静脉抗生素医嘱已持续至少48小时、无固定疗程、无发热且有经口饮食医嘱的患者生成警报。他们发现，该系统配合静脉转口服治疗政策的修订和员工培训，使干预后时期的转换率显著提高。Quintens等人[13] 讨论了一套基于13项标准的更复杂规则的实施，旨在提高临床医生对转换提醒的接受度。该规则关注口服治疗的适宜性，并不直接针对患者的临床改善状态。

如上文所述，临床决策支持系统已表明，实施简单的临床规则并结合对临床医生的教育，能够显著提高静脉转口服抗生素转换率。这些研究中讨论的规则主要针对患者是否适合口服治疗，并未（或几乎未）以复杂的方式评估患者的临床改善情况。在本研究中，我们旨在利用机器学习更好地理解静脉转口服抗生素治疗决策的这一方面。需要注意的是，由于我们无法获取可靠的真实转换标签，因此无法将这些基于规则的方法与所提出的系统进行基准比较。

S2 模型校准

如正文结果所示，尽管标准神经过程模型的排序性能很强，但在转换准备度分类任务中的校准效果较差。较高的布里尔分数表明，模型的预测概率未能准确反映患者达到转换准备度的真实可能性。具体而言，模型持续低估转换准备度的概率。

这种误校准源于训练目标与电子健康记录数据中异常值之间的相互作用。模型通过最小化高斯预测分布下观测测量值的负对数似然进行训练。由于高斯分布的尾部较轻，对于日常采集的生命体征中常见的测量误差和瞬时尖峰，它们赋予的概率非常低。模型在训练中遇到此类异常值时，对于不确定性较小的预测，负对数似然会变得非常大，从而使模型学会扩大其预测分布。图 S1 通过两个互补视角展示了这一现象：左图显示校准曲线位于对角线上方，表明不确定性估计过高；右图显示分箱平均误差低于对角线，意味着模型预测的实际误差始终低于预测不确定性。

当使用欠自信预测模型，将临床标准应用于预测分布以预测转换准备度时，这种误校准会传递到分类任务中。

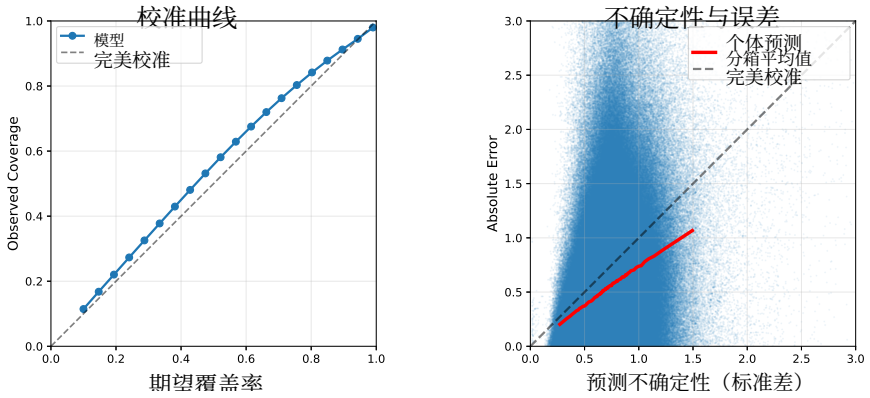


图 S1：左图展示了预测模型预测不确定性的校准曲线。对于每个置信水平（例如 80%），我们使用正态预测分布生成预测区间，该区间应以相应概率包含真实值。横轴表示所选置信水平，纵轴表示落在区间内的测试观测值的实际比例。完美校准呈对角线分布：80% 的区间应包含 80% 的观测值。对角线以上的点表示欠自信预测。右图展示了预测不确定性与绝对预测误差之间的关系。散点表示个体预测，按第1至第99百分位数分箱后的平均值（红线）总结了变化趋势。对角线表示理想校准，即预测不确定性的大小与实际误差的大小相匹配。在此情况下，欠自信模型的分箱平均误差低于对角线；也就是说，预测不确定性高于实际误差。

较宽的预测分布会增加预测值落在定义转换准备度的正常范围之外的概率。因此，预测的转换准备度概率会系统性地偏低，从而导致结果中观察到的校准不佳。

NP调优变体通过添加一个直接基于切换准备度标签训练的分类头来解决这一问题。该分类头学习校正预测模型的置信度不足，从而生成与观测频率相匹配的校准概率。调优后的模型在保持较强排序性能的同时，实现了与基线方法相当的布里尔分数。这种方法使系统能够保留预测框架在可解释性方面的优势，同时提供适用于临床决策支持的校准概率输出。

变量	范围	单位
心率	10 – 400	bpm
收缩压	0 – 400	mmHg
血氧饱和度	0 – 100	%
呼吸频率	0 – 120	bpm
体温	50 – 120	°F

表 S1：数据清理所用的合理值范围。

S3 数据设置

S3.1 时间划分

我们实验设计中的一个关键步骤是对数据进行严格的时间划分，这是模拟真实部署场景并防止前视偏差所必需的。MIMIC-IV 中的去标识化流程会为同一患者的所有日期加上一个随机偏移量，从而保留患者内部的时间线，但打乱不同患者之间入院记录的时间顺序。在这种被打乱的数据上进行随机划分，会错误地混合过去和未来的数据，导致过于乐观的评估结果。为解决这一问题，我们为每条就诊记录重建一个近似且可按时间排序的入院日期。按照 MIMIC-IV 创建者所述的方法，我们将入院表中以锚定年份形式提供的实际入院年份，与经过偏移的入院时间中的月份和日期结合起来。虽然这无法还原确切日期，但能够使我们按年份正确排列所有就诊记录。-

S3.2 任务数量

	MIMIC-IV		UCLH	
划分	**预测**	**转换准备度**	**预测**	**转换准备度**
训练集	167,923	79,148	344,133	76,586
Val.	18,804	8,902	38,864	8,885
Test	52,537	22,615	186,759	41,053

表 S2：预测任务和转换准备度预测任务在各数据集分区中的任务数量。

S3.3 切换标准

测量指标	**切换标准**
体温 (°F)	96.8 – 100.4
心率 (每分钟心跳次数)	41 – 90
呼吸频率（次/分钟）	9 – 20
收缩压（mmHg）	101 – 219
血氧饱和度（%）	> 94

表 S3：切换准备就绪的生理标准。

S4 神经过程的进一步细节

在本节中，我们给出神经过程（NP）模型的形式化描述。我们采用卷积条件神经过程（ConvCNP）[31, 30], ，该模型在给定已观测测量值的上下文集的条件下，定义目标测量值上的条件分布。

S4.1 模型定义

沿用正文中确立的记号，令一次患者就诊由一组观测值表示： $X_m = \{(t_{m,i}, c_{m,i}, x_{m,i})\}_{i=1}^{N_m}$，其中$t_{m,i} \in \mathbb{R}^+$ 表示观测时间，$c_{m,i} \in \{1, \ldots, C\}$ 索引生命体征变量（例如心率、体温），$x_{m,i} \in \mathbb{R}$ 为测量值。令 X_m^{τ} 表示就诊记录 m 在时间 τ之前的病史，令 $X_m^{[\tau,\tau+\Delta]}$表示预测窗口内的目标观测值。ConvCNP通过三个组成部分对预测分布$p(X_m^{[\tau,\tau+\Delta]} \mid X_m^{\tau})$进行建模：

编码器。 上下文集 X_m^{τ} 通过径向基函数（RBF）核平滑映射为函数表示，将观测值置于离散时间网格上。这种称为 ConvDeepSet [30, 34], 的编码方式，为每个生命体征变量分别保留数据密度和信号值的通道。

处理器。 UNet 卷积架构 [32]对网格表示进行处理。我们的实现采用瓶颈层前后各四个卷积块，特征通道数 D 作为超参数。令 Φ 表示编码器与处理器的组合。所得表示为

$$\mathbf{R} = \Phi(X_m^{\tau}) \in \mathbb{R}^{D \times G}, \tag{S1}$$

其中， G 是离散时间域中的网格点数。

解码器。 核大小为 1 的卷积层将 D 个特征通道投影到 $2C$ 个输出通道，从而得到每个生命体征 $c \in \{1, \ldots, C\}$的均值 $\mu_{t,c}$和方差 $\sigma_{t,c}^2$。通过对投影后的网格进行插值，可恢复任意连续目标时间 t 处的参数。

S4.2 训练目标

条件神经过程框架假设，给定上下文集后，目标观测值在条件上相互独立。联合预测似然可分解为

$$p(X_m^{[\tau,\tau+\Delta]} \mid X_m^{\tau}) = \prod_{(t_j, c_j, x_j) \in X_m^{[\tau,\tau+\Delta]}} \mathcal{N}\big(x_j \mid \mu_{t_j,c_j}, \sigma_{t_j,c_j}^2\big), \tag{S2}$$

其中$\mu_{t_j,\ c_j}$ 和 $\sigma_{t_j,\ c_j}^2$ 是时间 t_j 和变量 c_j对应的解码器输出。

模型通过最小化 目标观测值的负对数似然（NLL）进行训练 ：

$$\mathcal{L}_{\text{forecast}} = -\frac{1}{|X_m^{[\tau,\tau+\Delta]}|} \sum_{(t_j, c_j, x_j) \in X_m^{[\tau,\tau+\Delta]}} \log \mathcal{N}\big(x_j \mid \mu_{t_j,c_j}, \sigma_{t_j,c_j}^2\big). \tag{S3}$$

该目标通过均值实现准确的点预测，同时通过方差学习不确定性估计，以捕捉生命体征数据中固有的噪声。

S4.3 转换准备度概率

转换准备度取决于在整个预测窗口 $[\tau, \tau + \Delta]$ 内，所有生命体征是否始终处于临床定义的正常范围内。对于每项生命体征 c，令 $[\ell_c, u_c]$ 表示其可接受范围。令 $\mathcal{T} = \{t_1, \ldots, t_K\}$ 表示预测网格上的评估时间点集合（以~3小时区间为间隔），并令 $X_{t,c}$ 表示生命体征 c 在时间 t 的预测测量值这一随机变量，其分布为 $\tilde{X}_{t,c} \sim \mathcal{N}(\mu_{t,c}, \sigma^2_{t,c})$。

患者具备转换条件的概率，记为 $P(y^\tau_m = 1 \mid X^\tau_m)$，即所有预测的测量值均符合临床标准的概率：

$$P(y^\tau_m = 1 \mid X^\tau_m) = P\left(\bigcap_{t\in\mathcal{T}} \bigcap_{c=1}^{C} \ell_c \le \tilde{X}_{t,c} \le u_c\right). \tag{S4}$$

在条件独立假设下，该联合概率可分解为 s:

$$P(y^\tau_m = 1 \mid X^\tau_m) = \prod_{t\in\mathcal{T}} \prod_{c=1}^{C} P\big(\ell_c \le \tilde{X}_{t,c} \le u_c\big). \tag{S5}$$

每个边缘概率均通过高斯分布的累积分布函数进行解析计算。

S4.4 分类头

对于NP调优变体，我们在模型中添加一个分类头，以直接预测二元切换准备度标签 $y^\tau_m \in \{0, 1\}$。该分类头作用于处理器生成的特征表示$\mathbf{R} \in \mathbb{R}^{D\times G}$。沿时间维度进行全局最大池化，可得到一个固定大小的嵌入：

$$h_d = \max_{g\in\{1,\ldots,G\}} R_{d,g}, \quad \mathbf{h} = [h_1, \ldots, h_D]^\top \in \mathbb{R}^D. \tag{S6}$$

多层感知机将该嵌入映射为预测概率 $\hat{p} = \mathrm{MLP}(\mathbf{h}) \in [0, 1]$。

分类头在第二阶段进行训练，此时预测模型的权重被冻结，并使用二元交叉熵损失：

$$\mathcal{L}_{\text{class}} = -\big[y^\tau_m \log \hat{p} + (1 - y^\tau_m) \log(1 - \hat{p})\big]. \tag{S7}$$

这种两阶段方法将概率校准为二元任务标签，同时保留从预测目标中学到的特征表示。

S5 神经过程超参数

参数	值
架构	
层数	4
通道数*	{64, 128, 256, 512}
长度尺度	1.0 小时
训练	
学习率*	$\{10^{-5}, 5\times10^{-5}, 10^{-4}, 5\times10^{-4}, 10^{-3}\}$
批量大小	512
每轮大小	16384 个任务
验证轮次大小	4096 个任务
最大轮数	500
早停耐心值	250 轮
学习率调度器	余弦，线性预热 50 轮
Data	
回看窗口	48小时
预测窗口	12小时
预测网格大小	3小时
最小任务测量	10
每日训练任务数	1.0

表 S4：神经过程模型超参数。标记为 * 的参数通过网格搜索选出，搜索空间显示在括号中。

参数	值
架构	
隐藏层 1 大小	256
隐藏层 2 大小	128
丢弃率	{0.0, 0.1, 0.5}
训练	
学习率	$\{5\times10^{-5}, 10^{-4}, 5\times10^{-4}, 10^{-3}\}$
其他设置	与基础模型相同

表 S5：NP 调优的分类头超参数。训练分类头时，基础预测模型的权重保持冻结。

参数	值
GBDT	
估计器数量	{5, 10, 50, 100, 200}
最大深度	{3, 4, 5, 6}
学习率	{0.001、0.005、0.05、0.1、0.5}
逻辑回归	
正则化强度倒数 (C)	{0.1，1.0，10.0，100.0，1000.0}

表 S6：基线模型超参数。我们对各组超参数进行了网格搜索。选择在验证集上平均精度最高的模型，在测试集上进行评估。

S6 评估指标

我们将测试集记为 $\{(\hat{p}_i, y_i)\}_{i=1}^{N}$，其中 $\hat{p}_i \in [0,1]$ 是预测的切换准备度概率，$y_i \in \{0,1\}$ 是实例 i 的真实标签。对于决策阈值 $\theta \in [0,1]$，令 TP(θ)、FP(θ)、TN(θ) 和 FN(θ) 分别表示真正例、假正例、真负例和假负例的数量。

ROC曲线下面积（AUROC）AUROC评估模型区分可切换患者与不可切换患者的能力。它表示随机选取的正实例获得的预测概率高于随机选取的负实例的概率。AUROC计算为受试者工作特征曲线的积分，该曲线绘制不同阈值下真正率相对于假阳性率的变化：

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}\big(\mathrm{FPR}^{-1}(t)\big)\, dt, \tag{S8}$$

其中，阈值 θ 处的真正率和假阳性率定义为

$$\mathrm{TPR}(\theta) = \frac{\mathrm{TP}(\theta)}{\mathrm{TP}(\theta) + \mathrm{FN}(\theta)}, \qquad \mathrm{FPR}(\theta) = \frac{\mathrm{FP}(\theta)}{\mathrm{FP}(\theta) + \mathrm{TN}(\theta)}. \tag{S9}$$

在实践中，该积 分采用梯形法则对经验ROC曲线进行近似 .

平均精度（平均精确率）平均精度概括精确率-召回率曲线，对类别不平衡分类任务尤其具有参考价值。与对称处理正类和负类的AUROC不同，平均精确率关注相对于正类的性能。

平均精确率计算为精确率-召回率曲线下的面积，近似为每个阈值处精确率的加权和：

$$\mathrm{AP} = \sum_{k=1}^{K} \big[R(\theta_k) - R(\theta_{k-1})\big] \cdot P(\theta_k), \tag{S10}$$

其中，$\theta_1 < \theta_2 < \cdots < \theta_K$为各不相同的阈值，精确率和召回率定义如下：

$$P(\theta) = \frac{\mathrm{TP}(\theta)}{\mathrm{TP}(\theta) + \mathrm{FP}(\theta)}, \qquad R(\theta) = \frac{\mathrm{TP}(\theta)}{\mathrm{TP}(\theta) + \mathrm{FN}(\theta)}. \tag{S11}$$

A h高平均精确率表示模型同时具有较高的召回率和较高的精确率。

布里尔分数 布里尔分数用于评估概率预测的准确率和校准度。它定义为预测概率与观测结果之间的均方误差:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2. \tag{S12}$$

布里尔分数越低，表示概率估计的校准效果越好、准确度越高。该分数范围为 0（预测完美）到 1（错误程度最大）。始终预测类别先验概率 p 的模型，其布里尔分数为 $p(1-p)$。

前5项精确率 前5项精确率通过评估模型为患者安排审查优先级的能力来衡量临床效用。对于测试集中至少有 10 条活跃就诊记录的每一天 d，按切换准备度预测概率从高到低排列患者。令 $\mathcal{D}$ 表示有效日期的集合，令 $\pi_d(j)$ 表示第 j 高预测概率对应的患者在第 d 天的索引。前5项精确率是所有有效日期中排名前 5 的患者里实际具备转换条件的患者比例的平均值:

$$\text{Precision@5} = \frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\frac{1}{5}\sum_{j=1}^{5} y_{\pi_d(j)}. \tag{S13}$$

S7 其他结果

S7.1 标准阈值敏感性

测量指标	切换标准
体温（°F）	96.8 – 100.58
心率（每分钟心跳次数）	40 – 131
呼吸频率（次/分钟）	8 – 24
收缩压（mmHg）	90 – 229
血氧饱和度（%）	> 91

表 S7：用于判定切换准备就绪的替代宽松标准。

由于不同医疗机构很可能有不同的要求，我们测试了所提出系统对临床标准变更的敏感性。S7.1 节展示了一组替代的宽松标准。将这些标准应用于 MIMIC 数据后，阳性（具备转换条件）天数的比例从 12.2% 增至 50.0%。采用较宽松的标准进行测量时，我们的模型仍保持了相对性能。排序指标的结果见表 S8，二元指标的结果见表 S9。

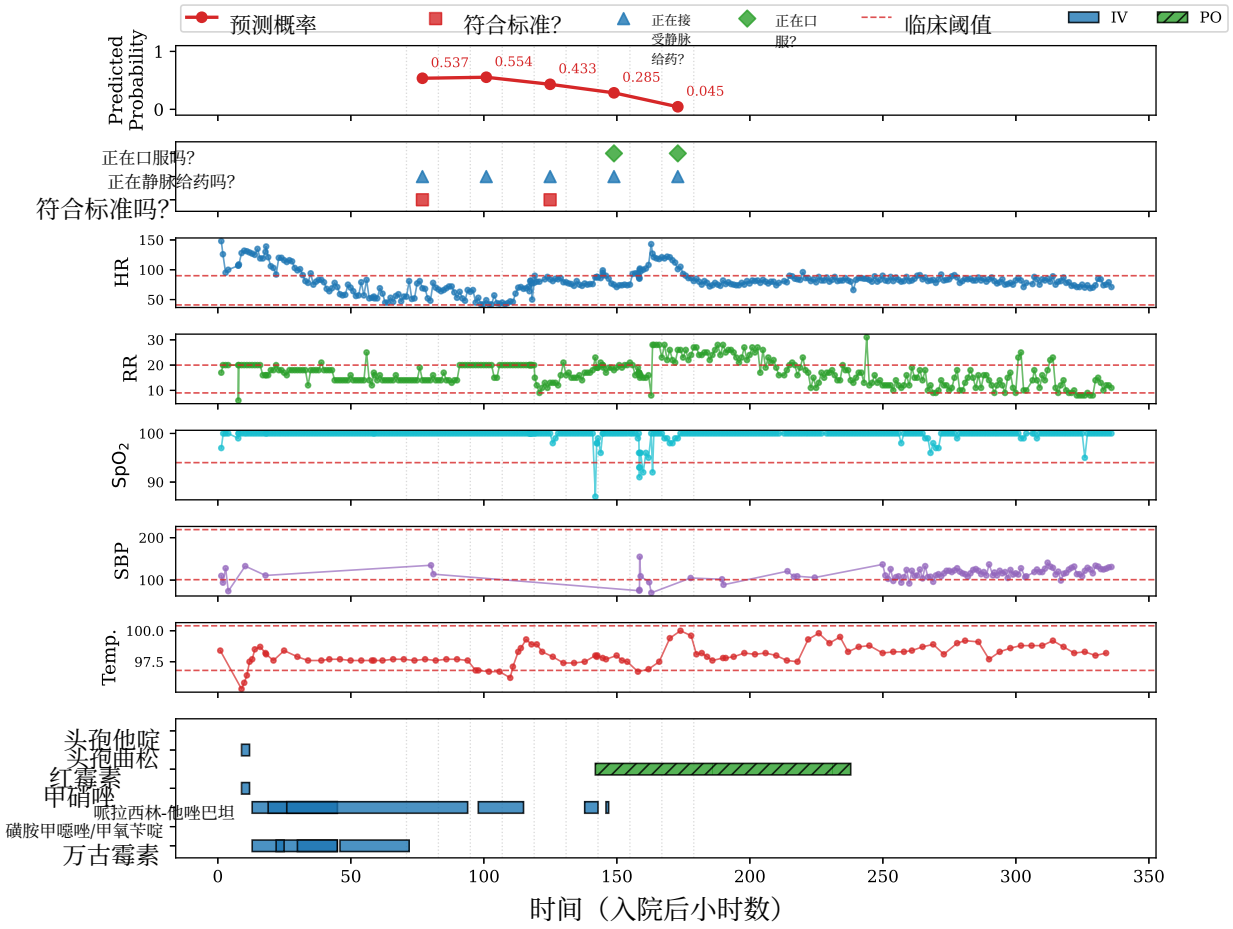
数据集	模型	指标			
		AUROC	AP	布里尔分数	前5项精确率
MIMIC-loose	逻辑回归	0.676 [0.669-0.682]	0.650 [0.640-0.660]	0.229 [0.228-0.231]	0.680 [0.663-0.696]
	GBDT	0.829 [0.824-0.834]	**0.809 [0.801-0.817]**	**0.168 [0.165-0.170]**	**0.835 [0.821-0.849]**
	NP	**0.836 [0.831-0.841]**	**0.810 [0.802-0.817]**	0.239 [0.235-0.243]	**0.841 [0.827-0.855]**
	调优后的NP	0.814 [0.809-0.820]	0.792 [0.783-0.800]	0.175 [0.172-0.177]	0.822 [0.808-0.838]

表 S8：采用较宽松临床标准（切换率 50.0%）时，MIMIC 数据集上切换准备度预测的模型性能。粗体值表示表现最佳的模型（$p \geq 0.05$）。模型缩写：GBDT（梯度提升决策树）、NP（神经过程）。

数据集	模型	指标			
		F1 分数	准确率	精确率	召回率
MIMIC-loose	重复	0.739 [0.733-0.745]	0.700 [0.695-0.706]	0.654 [0.647-0.662]	0.849 [0.842-0.856]
	逻辑回归	0.684 [0.678-0.690]	0.560 [0.554-0.567]	0.534 [0.527-0.541]	**0.950 [0.946-0.954]**
	GBDT	0.772 [0.766-0.777]	0.741 [0.736-0.747]	0.690 [0.683-0.697]	0.875 [0.869-0.881]
	NP	**0.779 [0.773-0.784]**	**0.753 [0.748-0.759]**	**0.707 [0.699-0.714]**	0.867 [0.861-0.872]
	调优后的NP	0.762 [0.756-0.768]	0.730 [0.724-0.736]	0.681 [0.674-0.689]	0.865 [0.859-0.872]

表 S9：采用较宽松临床标准（切换率 50.0%）时，MIMIC 数据集上二元切换准备度预测的模型性能。粗体值表示性能最佳的模型（$p \geq 0.05$）。模型缩写：GBDT（梯度提升决策树）、NP（神经过程）。

S7.2 就诊记录示例



图S2：来自MIMIC测试集的一则示例就诊记录，展示系统预测获益并不那么明确的情形。系统建议更早进行转换，但患者随后在转换后病情加重。在该就诊记录中，由于静脉处方在150小时后停止，测试日仅覆盖整个住院时长的大约一半。各面板所展示的信息与正文图中的面板相同。变量缩写：心率（心率）、呼吸频率（呼吸频率）、SpO_2（血氧饱和度）、收缩压（收缩压）、体温（体温）。抗生素缩写：哌拉西林-他唑巴坦，哌拉西林-他唑巴坦；磺胺甲噁唑/甲氧苄啶，磺胺甲噁唑-甲氧苄啶

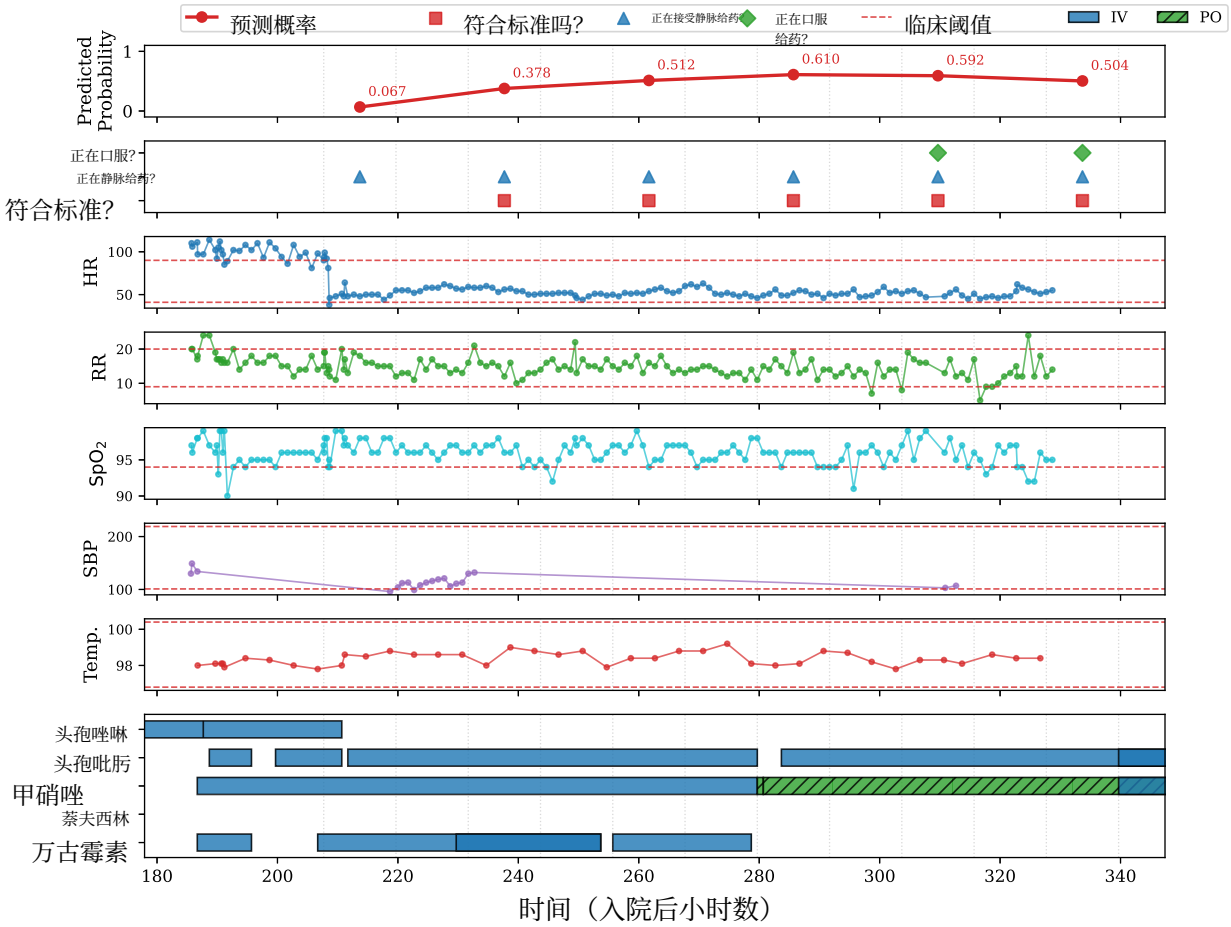


图 S3：MIMIC测试集中的一个示例就诊记录。在此示例中，患者在就诊记录开始时即处于不适状态，但除少数几次峰值外，其生命体征在剩余的大部分时间内均保持稳定。随着就诊记录的推进，模型越来越确信患者已做好转换准备，尽管实际并未发生转换。各面板展示的信息与正文图中的相同。变量缩写：HR（心率）、RR（呼吸频率）、SpO_2（血氧饱和度）、SBP（收缩压）、Temp.（体温）。

S8 统计显著性检验

我们使用配对自助法重采样，结合基于自助法的假设检验和邦费罗尼校正来比较模型性能。在 1,000 次自助法迭代中，每次都从测试集中有放回地抽样生成一组索引。每次迭代中，这组相同的索引均用于选取每个模型的预测，确保性能指标是在相同的数据子集上计算得出的。

我们使用自助法差异检验，将各模型与性能最佳的模型进行成对比较。对于每次比较，我们计算了最佳模型优于对比模型的自助法迭代所占比例。如果该比例超过 $1-\alpha/2$ （这表明在自助样本中，最佳模型始终优于对比模型），则认为模型之间存在显著差异。我们针对多重检验应用邦费罗尼校正：对于 k 个模型，显著性阈值调整为 $\alpha/(k-1)$，其中 $\alpha = 0.05$。在此阈值下，与表现最佳模型之间差异不显著的模型被标记为“显著最佳”，并在结果表中以粗体显示。