arXiv:2602.08171v1 [cs.LG] 9 Feb 2026

用于临床试验治疗个性化的因果机器学习框架：在溃疡性结肠炎中的应用

Cristian Minoccheri[1*]，Sophia Tesic[1], KayvanNajarian[1,3,4], Ryan Stidham[1,2]1密歇根大学Gilbert S. Omenn计算医学与生物信息学系，美国密歇根州安娜堡市Geddes大道1109号，邮编48104。[2]密歇根大学胃肠病学系，美国密歇根州安娜堡市Geddes大道1109号，邮编48104。[3]密歇根大学急诊医学系，美国密歇根州安娜堡市Geddes大道1109号，邮编48104。[4]密歇根大学电气工程与计算机科学系，美国密歇根州安娜堡市Geddes大道1109号，邮编48104。

*通讯作者。电子邮箱：minoc@umich.edu;

摘要

随机对照试验可估计平均治疗效应，但治疗反应异质性促使人们采用个性化方法。一个关键问题是，统计上可检测的异质性是否能转化为更优的治疗决策——这两个问题彼此不同，所得结论可能相互矛盾。我们提出了一个模块化的因果机器学习框架，分别评估这两个问题：置换重要性用于识别能够预测异质性的特征，最佳线性预测器（BLP）检验用于评估统计显著性，双重稳健策略评估用于衡量基于这种异质性采取行动是否能改善患者结局。我们将该框架应用于UNIFI维持试验中溃疡性结肠炎患者的个体水平数据，比较安慰剂（n=144）、每12周一次的标准剂量乌司奴单抗（n=143）和每8周一次的剂量强化乌司奴单抗（n=274），并使用交叉拟合的X-learner模型，纳入基线人口统计学特征、用药史、第8周临床评分、实验室生物标志物和视频衍生的内镜特征。BLP检验发现，内镜特征与乌司奴单抗相对于安慰剂的治疗效应异质性之间存在很强的关联（p<0.001），但双重稳健

策略评估显示，纳入内镜特征并未改善预期缓解（95%置信区间： −1.6至 +6.6个百分点），而折外多臂评估显示表现更差（缓解率为30.5% vs 36.8%）。将预后贡献与策略价值进行诊断性比较后发现，内镜评分表现为疾病严重程度标志物——它们能够改善未接受治疗患者的结局预测，却会给治疗选择增加噪声；而临床变量（粪便钙卫蛋白、年龄、C反应蛋白）捕捉到了与决策相关的变异。这些结果表明，因果机器学习在临床试验中的应用应在异质性检验之外纳入策略层面评估：统计显著性与决策表现之间的差距本身就具有信息价值，有助于区分预后特征与真正的效应修饰因子。

1 引言

治疗反应的异质性是各治疗领域持续面临的挑战。随机对照试验（RCT）能够提供无偏的平均治疗效应估计，但患者间的结局差异显著 [1],因此，需要建立有原则的框架，以识别治疗获益存在差异的亚组。一个关键问题是，估计出的异质性是否能转化为更优的治疗决策，而非仅仅反映统计上的区分。

尽管统计学和计量经济学文献中估计异质性治疗效应的方法已取得长足进展，但这些方法在临床试验数据中的系统应用仍然有限。预后模型无法区分哪些患者即使不接受治疗也会改善，哪些患者是因治疗而改善；而异质性治疗效应则估计治疗对每位个体患者的因果效应，从而直接揭示哪些患者从特定治疗中获得的获益足以抵偿其成本与风险。异质性治疗效应可通过条件平均治疗效应（CATE）进行估计，即给定患者协变量时的预期治疗效应。关于因果森林 [2]和元学习器 [3] 的奠基性研究，已确立了估计CATE的灵活方法；双重稳健估计 [4–6] 则为评估个体化治疗规则提供了有原则的依据。与此并行的政策学习文献，已正式阐明了检测效应异质性与学习能够改善预期结局的治疗策略之间的区别 [7, 8]。最佳线性预测器检验 [9] 提供了一种框架，用于评估特定特征组是否在其他协变量之外，对治疗效应异质性有所贡献。尽管方法学取得了这些进展，鲜有研究将这些工具整合应用于临床试验数据，且仍有重要的实际问题有待解决：如何区分预后特征与真正的效应修饰因子；样本内异质性检测能否转化为样本外更优的治疗分配；以及如何处理剂量探索试验和维持治疗试验中自然出现的多臂比较。

近期研究 [1]将因果森林应用于随机对照试验数据，重点识别易于解释的亚组划分，并证明了基于树的效应估计的价值。本研究在侧重点和范围上有所不同：我们并不专注于亚组

识别，而是将 X-学习器估计与双重稳健策略评估相结合，直接评估候选特征集能否改进治疗分配决策，并通过嵌套折外评估将分析扩展至多臂设置。这种以策略为导向的视角——评估检测到的异质性是否能转化为更优决策——回应了临床试验文献中一个尚未得到充分探讨的问题。

我们提出了一种用于随机对照试验治疗个性化的因果机器学习框架，该框架结合了 X-learner 条件平均治疗效应估计、基于置换的特征重要性、最佳线性预测器（BLP）检验 [9], 、双重稳健策略价值评估和预后性能建模。该框架旨在区分统计关联与治疗选择的改进，并纳入折外评估，以量化所学治疗规则的泛化能力。该方法适用于关注治疗效应异质性的各类治疗领域。它有助于识别哪些特征能够预测异质性、异质性在统计学上是否显著，以及异质性是否能改善治疗决策。

为展示该框架，我们将其应用于 UNIFI 随机维持治疗试验中乌司奴单抗治疗溃疡性结肠炎（UC）的患者个体数据 [10]。炎症性肠病的治疗个性化正日益受到关注 [11, 12], ，但现有研究大多侧重于预后预测——识别哪些患者会对某种特定治疗产生应答——而非治疗效应修饰，后者关注的是哪些患者从一种治疗中获益多于另一种治疗。这一区别对临床决策有直接影响，但尚未在 UC 维持治疗试验中采用因果推断方法进行正式评估。

内镜指标（如梅奥内镜评分（MES））广泛用于评估黏膜愈合 [13, 14], ，早期内镜反应也已被作为长期缓解的预后预测因子进行研究 [15]。然而，考虑到评估者间变异、侵入性和成本 [16]，内镜特征能否作为稳定的效应修饰因子——帮助识别需要剂量强化（每8周而非每12周治疗）的患者、从持续接受活性治疗而非停用安慰剂中获益的患者，或适合采用完全不同维持治疗策略的患者——目前仍不明确。

我们使用UNIFI维持治疗试验的患者个体数据展示该框架，评估内镜特征能否在二元比较和多臂比较中，提升治疗分配的个性化程度，且效果是否优于常规可获取的临床标志物。本研究使用的代码可在https://github.com/Minoch/UNIFI causal获取。

主要贡献如下：

1）提出一种模块化因果机器学习框架，用于随机对照试验中的治疗个性化。该框架将条件平均治疗效应估计、特征重要性、统计检验和双重稳健策略评估整合到单一流程中，并针对不同模块分别回答以下问题：哪些特征能够预测异质性（置换重要性）？异质性是否具有统计学显著性（BLP检验）？异质性是否有助于改善治疗决策（策略价值）？

2）实证证明，在溃疡性结肠炎的UNIFI维持治疗试验中，这些问题可能得出相互矛盾的答案：内镜特征显示

与治疗效应异质性存在强统计关联（最佳线性预测器 z = 8.28），却未改善治疗分配，并在多臂情境中降低了折外策略价值（缓解率30.5% vs. 36.8%）。这强调异质性检测与决策表现必须分别评估。

3) 一种用于区分预后特征与效应修饰因子的诊断策略：比较它们对结局预测的贡献（增量布里尔分数）与它们对不同治疗组中策略价值的贡献。将该方法应用于此处后发现，内镜评分是严重程度标志物——能改善未接受治疗患者的预后判断，却会给治疗选择增加噪声；而常规临床变量（钙卫蛋白、年龄、C反应蛋白）则反映了与决策相关的变异。

2 数据集

本分析使用了UNIFI维持期随机对照试验的数据。该试验研究了乌司奴单抗对中重度溃疡性结肠炎患者的疗效 [10]。队列由意向性治疗（ITT）人群组成：这些患者接受了诱导治疗，并被随机分配至为期44周的维持方案，包括安慰剂、每12周接受一次乌司奴单抗（Q12）或每8周接受一次乌司奴单抗（Q8） [10, 17]。

主要结局为第44周基于完全Mayo评分判定的临床缓解，未完成者被归类为无应答者（意向性治疗）。治疗分配由维持用药指示变量确定（Q12组为MaintMed UST 12，Q8组为MaintMed UST Q8），其余受试者归为安慰剂组。将诱导治疗期间是否接受皮下注射乌司奴单抗（InductionMed UST SC）作为协变量纳入分析。治疗分配变量用于定义比较组，但未纳入结局模型，以避免信息泄漏。

用于建模的特征集包括患者年龄和性别、二元用药史变量（既往未暴露于生物制剂、基线免疫调节剂使用情况、治疗不耐受或难治史），以及意向治疗纳入指标。第8周的实验室测量值提供了炎症的客观标志物，包括粪便钙卫蛋白（第8周粪便钙卫蛋白）和C反应蛋白（第8周C反应蛋白）。

临床活动度变量来自第8周评估：绝对排便次数（第8周排便次数）、直肠出血评分、大便频率评分、医师总体评估和部分Mayo评分（第8周部分Mayo评分）。根据从基线至第8周的变化，计算了上述各项基于症状的测量指标的变化评分（例如，pMayo abs 0 8、RBscore abs 0 8、SFscore abs 0 8、PGscore abs 0 8、ABSstoolabs 0 8）。

内镜和视频衍生特征通过对乙状结肠进行详细的节段评分，评估黏膜炎症及其愈合情况。在第8周的每次结肠镜检查中，量化显示各Mayo内镜亚评分（0–3）的组织百分比（第8周MES 0百分比、第8周MES 1百分比、第8周MES 2百分比、第8周MES 3百分比），并纳入近期提出的累积疾病评分（CDS） [18]。同时纳入这些分布从基线至第8周的变化，以捕捉早期反应动态（MES 0 perc Abs 0 8、MES 1 perc Abs 0 8、MES 2 perc Abs 0 8、MES 3 perc Abs 0 8、CDS abs 0 8）。

所有特征定义均列于下方表格1 。缺失数据采用中位数插补处理连续变量，采用众数插补处理二元指标。连续特征经过z分数标准化，二元变量则保留其原始尺度。

表 1 数据集特征描述

特征	描述
Endoscopic Features	
第8周MES 0百分比	第8周乙状结肠Mayo评分为0的百分比
第8周MES 1百分比	第8周乙状结肠Mayo评分为梅奥1分的百分比
第8周MES 2百分比	第8周乙状结肠Mayo评分为梅奥2分的百分比
MES 3 perc wk8	第8周乙状结肠Mayo评分为梅奥3分的百分比
MES 0 perc Abs 0 8	第0周至第8周乙状结肠梅奥0分百分比的变化
MES 1级百分比绝对变化（第0周至第8周）	第0周至第8周乙状结肠梅奥1分百分比的变化
MES 2级百分比绝对变化（第0周至第8周）	第0周至第8周乙状结肠梅奥2分百分比的变化
MES 3级百分比绝对变化（第0周至第8周）	第0周至第8周乙状结肠梅奥3分百分比的变化
第8周 CDS	第8周乙状结肠 CDS评分 （连续型黏膜活动性标志物）
CDS绝对值 0 8	第0周至第8周 CDS评分的绝对变化
Non-Endoscopic Features	
AGE	入组时年龄（岁）
男性	男性二元指标
生物制剂初治	无既往生物制剂暴露（用药史）
BIMM	基线时使用免疫调节剂（用药史）
对所有相关药物不耐受或难治	不耐受或难治史
	免疫调节剂/类固醇
对皮质类固醇不耐受或难治	不耐受或难治史 仅限于皮质类固醇
第8周粪便钙卫蛋白	第8周粪便钙卫蛋白（实验室检测，肠道炎症）
第8周C反应蛋白	第8周C反应蛋白（实验室检测，全身性炎症）
第8周排便次数	第8周绝对排便次数
第8周医师总体评估	医师总体评估（0-3），第8周
第8周部分Mayo评分	第8周部分Mayo评分（症状综合评分）
第8周直肠出血评分	直肠出血评分（0-3），第8周
第8周粪便频率评分	大便频率评分（0-3），第8周
部分Mayo评分绝对变化 0 8	部分Mayo评分变化（第0周至第8周）
直肠出血评分绝对变化 0 8	直肠出血评分变化（第0周至第8周）
大便频率评分绝对值 0 8	大便频率评分变化（第0周至第8周）
医生总体评估评分绝对值 0 8	医师总体评估变化（第0周至第8周）
绝对大便次数绝对值 0 8	绝对大便次数变化（第0周至第8周）
诱导治疗药物 UST 皮下注射	接受皮下注射UST诱导治疗（二元变量）
Treatment Variables	
维持用药 UST 每12周	维持治疗分配：UST 每12周（二元变量）
维持用药 UST 每8周一次	维持治疗分配：UST 每8周（二元变量）
Primary Outcome	
第44周完全缓解-CALC 意向性治疗	第44周临床缓解 （完全Mayo评分，未完成者视为无应答者）

3 方法

我们采用了 X-learner 方法 [3] 来估计条件平均治疗效应（CATEs），并识别患者层面预测因素，以刻画异质性治疗反应。

我们开展了三项复杂程度递增的分析：（1）乌司奴单抗联合维持治疗（每8周一次和每12周一次）与安慰剂撤药的比较；（2）强化剂量（每8周一次）与标准剂量（每12周一次）维持治疗的比较；以及（3）多臂策略分析，将每位患者分配至三个选项中的最优策略。多臂分析采用嵌套交叉验证，以进行无偏的折外策略价值估计。

X-learner 分三个阶段进行 [3]。在第一阶段，我们使用梯度提升回归树（XGBoost）为每个治疗组分别拟合结局模型，根据基线和第8周特征预测第44周缓解。对于每项二元比较，分别为对照患者和接受治疗的患者拟合模型 μ_0（X）和 μ_1（X）（800 棵树，最大深度为 4，子采样率为 0.8，列采样率为 0.8，学习率为 0.05，L2 正则化为 1.0）。所有模型均采用五折交叉验证进行交叉拟合。

在第二阶段，我们通过生成伪结局来计算插补治疗效应。对于接受治疗的患者，$D_i^{(1)} = Y_i - \mu_0(X_i)$；对于对照组患者，$D_i^{(0)} = \mu_1(X_i) - Y_i$。我们分别拟合条件平均治疗效应模型，将这些伪结局对基线特征和第8周特征进行回归，使用梯度提升树（400个估计器，最大深度为4，子采样率为0.8，列采样率为0.8）。

在第三阶段，我们使用倾向评分加权合并两个条件平均治疗效应估计值。由于治疗采用随机分配，我们使用等于经验治疗概率的常数倾向评分。最终的条件平均治疗效应估计值为：

$$\tau(X_i) = e \cdot \tau^{(0)}(X_i) + (1 - e) \cdot \tau^{(1)}(X_i), \tag{1}$$

其中，$e = P(T = 1)$为常数倾向评分，$\tau_{(0)}(X_i)$和 $\tau^{(1)}(X_i)$分别为对照组和治疗组的条件平均治疗效应估计值。

对于多臂比较，我们使用相同的XGBoost架构拟合三个结局模型 $\mu_0(X)$、$\mu_{12}(X)$和$\mu_8(X)$。多臂策略定义为：

$$\pi(X_i) = \arg\max_{t \in \{0,12,8\}} \mu_t(X_i). \tag{2}$$

我们使用基于经验随机分配概率的双重稳健估计来评估该策略：$P(T = 0) = 0.257$、$P(T = \text{Q12}) = 0.255$，以及 $P(T = \text{Q8}) = 0.488$。为避免过拟合，我们采用嵌套式五折分层交叉验证：在训练折上拟合结局模型，并在留出折上评估策略，再汇总所有折中的双重稳健贡献。

为量化内镜特征与临床特征的相对贡献，我们在学习得到的条件平均治疗效应模型上进行了基于置换的特征重要性分析 [19, 20] 。对于二元比较，我们评估了在接受治疗的患者中拟合的第二阶段模型 $\tau^{(1)}(X)$。对于每个特征，在患者之间随机置换其取值 10 次，并计算均方误差相对于未置换基线的增加量

测量 [20]。整个过程重复 10 次，以计算均值和标准误。

对于多臂分析，我们计算了代表最优性差距的伪结局，即采用最佳替代治疗方案可实现的改善：

$$D_i^{\text{gap}} = \max_{t \in \{0,12,8\}} \mu_t(X_i) - \mu_{T_i}(X_i). \tag{3}$$

我们拟合了一个梯度提升模型（400棵树，最大深度为4，子采样率为0.8，列采样率为0.8）来预测这一差距，并在该模型上评估置换重要性。

组级重要性通过对以下两个类别中各项特征的重要性求和得到：（1）内镜特征（MES百分比、MES百分比变化和CDS评分）；（2）非内镜特征（人口统计学特征、用药史、第8周临床评分和实验室生物标志物）。

为正式检验内镜特征是否在临床变量之外具有额外贡献，我们采用了最佳线性预测器框架 [9]。对于每个特征组（临床和内镜），我们计算了每位患者的两项汇总统计量（该组内各特征的均值和标准差），并使用普通最小二乘法，将估计的条件平均治疗效应 $\tau(X_i)$ 对拼接后的汇总统计量进行回归。统计推断采用乘子自助法重抽样，迭代1000次。我们通过对两个内镜系数的z分数平方求和，构造Wald型检验统计量。

为评估内镜特征在治疗选择中的实际效用，我们使用双重稳健估计 [4–6]评估了假设治疗分配策略。对于每个特征集（全部特征或仅临床），我们定义策略$\pi(X_i) = \text{argmax}_t\, \mu_t(X_i)$，将治疗分配给预测结果最高的治疗方案。策略价值估计如下：

$$V(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\mu_{\pi(X_i)}(X_i) + \frac{\mathbb{1}(T_i = \pi(X_i))}{e_{\pi(X_i)}} \left(Y_i - \mu_{\pi(X_i)}(X_i) \right) \right], \tag{4}$$

其中，$\mathbb{1}(T_i = \pi(X_i))$是指示变量，表示观察到的治疗是否与策略分配一致；$e_{\pi(X_i)}$ 是通过随机化接受治疗 $\pi(X_i)$的概率； $\mu_t(X_i)$是X-learner第一阶段的结局模型。该估计量在结局模型或倾向评分至少有一个正确设定时是一致的 [4, 5]。

我们比较了内镜信息辅助策略 π_{all} 与仅临床策略 π_{clinical}的策略价值，计算 $\Delta = V(\pi_{\text{all}}) - V(\pi_{\text{clinical}})$。采用5000次配对自助法重抽样构建置信区间，以反映比较中的配对性质。对于多臂策略，嵌套交叉验证的折外估计提供了对泛化能力的保守评估。

作为补充分析，我们采用监督分类方法，在不考虑治疗分配的情况下，评估了基线特征和第8周特征对第44周缓解的预测准确度。经过校准的XGBoost分类器（600棵树，最大深度为4，子采样率0.85，列采样率0.85，学习率0.05，L2正则化

1.0）通过五折分层交叉验证结合保序回归校准进行评估。我们采用AUROC和Brier评分评估性能。我们比较了使用全部特征与仅使用临床特征的模型，比较范围包括全队列以及按治疗组分层的模型；采用配对自助法重抽样（5000次迭代）构建指标差异的95%置信区间。全队列包括所有治疗组中的患者（n=561）；按治疗组分层的分析用于评估各组内结果的一致性。

我们还计算了增量 Brier评分 $\text{Brier}_{\text{incremental}}$ 统计量，以量化在临床特征之外加入内镜特征后，额外解释的零模型误差比例：

$$\text{Brier}_{\text{incremental}} = \frac{\text{Brier}_{\text{clinical}} - \text{Brier}_{\text{all}}}{\text{Brier}_{\text{null}}}, \tag{5}$$

其中 $\text{Brier}_{\text{null}}$ 是预测边际结局概率（即所有患者的$\hat{p}_i = Y$）的零模型的 Brier评分。该指标等同于两个模型之间的预测准确性指数（IPA）差值： [21]两者之间：

$$\text{Brier}_{\text{incremental}} = \text{IPA}_{\text{all}} - \text{IPA}_{\text{clinical}}, \tag{6}$$

其中 $\text{IPA} = 1 - \text{Brier}/\text{Brier}_{\text{null}}$。正值表示内镜特征捕捉到了额外的预测信号，而负值则表明内镜特征引入了噪声或导致过拟合。

为使用观察到的治疗效应验证 CATE重要性发现，我们在预先指定的患者亚组内计算了亚组平均治疗效应（ATE）。由于治疗经过随机分配，任何亚组内的 ATE 均可通过结局均值的简单差值来识别：

$$\widehat{\text{ATE}}_S = \frac{1}{n_{1,S}} \sum_{i \in S: T_i = 1} Y_i - \frac{1}{n_{0,S}} \sum_{i \in S: T_i = 0} Y_i, \tag{7}$$

其中 S 表示该亚组，$n_{1,\ S}$ 和 $n_{0,\ S}$ 分别为亚组内接受治疗和对照的患者人数，Y_i 为二元缓解结局。亚组依据置换重要性分析得出的排名靠前特征定义；在有临床公认阈值的情况下，采用相应阈值：粪便钙卫蛋白 $\geq$150 μg/g [22]，C反应蛋白 $\geq$5 mg/L [23]，以及部分Mayo评分 $\geq$3 [24]。年龄采用中位数二分法，因为目前没有公认的临床临界值。置信区间通过进行5000次自助法重抽样构建。此分析提供了无模型验证：如果 CATE重要性排名反映了真实的异质性，那么由高重要性特征定义的亚组应呈现出有实际意义的观察到的平均治疗效应差异。

为方便读者，我们在表中简要总结了方法 2。

4 结果

基于置换的特征重要性分析显示，在识别差异化治疗反应方面，临床和人口统计变量的预测价值始终高于内镜特征（表3，图1）。

表2 方法通俗易懂的摘要

方法	摘要
X-learner 条件平均治疗效应估计 tion	分别为接受治疗和未接受治疗的患者构建预测模型 然后估计每位患者的个体治疗效应 通过提出以下问题来评估获益：“这位患者的结局会是怎样？ 如果接受另一种治疗？”
置换重要性	通过识别最能影响治疗获益的患者特征， 打乱每个特征的取值，并测量模型在预测哪些患者会获益方面 的表现会变差多少。
BLP检验	检验内镜特征是否能解释 临床特征 已捕捉到的估计治疗效果变异。结果显著表明存在关联， 但不一定能改善决策。
双重 稳健 策略评估	模拟若遵循模型的治疗 rec- 将会发生什么 ommendations 得到遵循，并估计由此产生的 remis- sion 率。回答检测异质性是否确实 转化为更优的治疗分配。
增量布里尔分数	衡量加入内镜特征是否能改善对谁将实现 缓解的预测。
亚组平均处理效应	一种无模型检验；患者被划分为临床定义的 比较各组之间的观察缓解率 以及各组内治疗组与对照组的缓解率，不使用任何模型。

对于成功完成诱导治疗后，乌司奴单抗联合维持治疗与安慰剂撤药的比较，非内镜特征的总体重要性达到0.79，约为内镜视频分析特征的1.5倍（重要性为 $= 0.54$）。在单个预测因子中，信息量最大的临床变量是第8周粪便钙卫蛋白（重要性为 $= 0.202$）、年龄（0.144）、第8周C反应蛋白（0.101），以及第8周乙状结肠中MES 1级所占百分比（0.093）。在内镜特征中，CDS相较基线的变化重要性最高（0.078），其次是MES 0级所占百分比相较基线的变化（0.058）。

对接受活性治疗的患者进行剂量强化维持治疗（每8周使用乌司奴单抗）与标准剂量（每12周）比较的分析显示了相似的模式。非内镜特征的重要性汇总值为0.71，约为内镜特征（重要性 =0.50）的1.4倍。对剂量决策预测能力最强的临床变量是粪便钙卫蛋白（重要性 $= 0.133$）、年龄（0.129）、C反应蛋白（0.096）以及内镜Mayo评分1级占比相较于基线的变化（0.093）。

对于在安慰剂、Q12 或每8周一次中为每位患者分配最佳维持治疗策略的多臂策略，重要性比率降至 1.2 倍（非内镜：0.92 ± 0.02，内镜：0.75 ± 0.01）。虽然这表明在复杂的多臂治疗决策中，内镜特征发挥的作用相对更大，但内镜特征的重要性绝对值仍低于临床特征。

表3 比较内镜与临床特征的汇总特征重要性

比较	非内镜	内镜	比率
UST 与安慰剂	0.79 ± 0.02	0.54 ± 0.01	1.46×
Q8 与 Q12	0.71 ± 0.03	0.50 ± 0.02	1.42×
多臂策略	0.92 ± 0.02	0.75 ± 0.01	1.22×

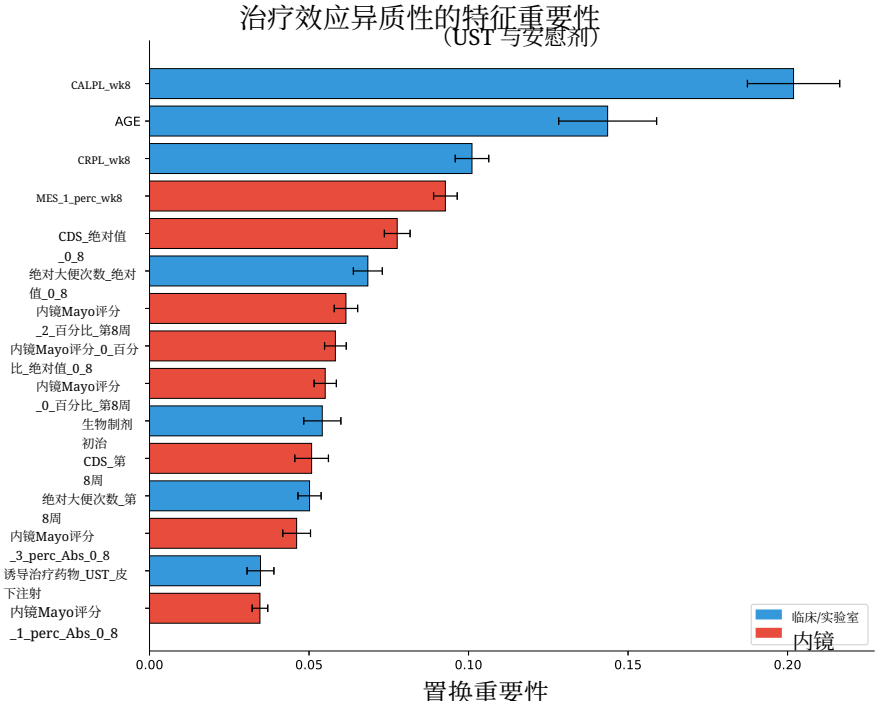

图 1 预测异质性治疗效应（UST 对比安慰剂）的置换重要性。特征按类别以不同颜色标示：蓝色表示临床/实验室特征，红色表示内镜特征。误差线表示 95%置信区间。临床特征，尤其是炎症生物标志物（粪便钙卫蛋白、C反应蛋白）和患者年龄，在识别治疗效应异质性方面的重要性最高。

为正式评估内镜特征是否在临床变量之外提供额外信息，我们进行了 BLP检验 [9]。对于乌司奴单抗与安慰剂的比较，BLP 检验结果具有高度统计学显著性（$z = 8.28$，$p < 0.001$），表明内镜汇总指标能够解释临床汇总指标之外估计治疗效果的额外变异（表4） [14, 15]。

然而，对于剂量强化（Q8 与 Q12），检验结果不显著（$z = 1.01$，$p = 0.311$），多臂策略也仅呈边缘显著性（$z = 1.82$，$p = 0.069$）。因此，尽管内镜特征对于继续使用生物制剂还是停药具有统计显著性，但对剂量优化或多臂分配的贡献微乎其微。

表 4 超出临床特征的内镜检查贡献的最佳线性预测器检验结果

比较	Z统计量	P值
UST 与安慰剂	8.28	<0.001
Q8 与 Q12	1.01	0.311
多臂策略	1.82	0.069

尽管乌司奴单抗与安慰剂比较的最佳线性预测器结果具有统计显著性，但双重稳健策略评估显示，将内镜检查纳入治疗分配规则并无实际益处（见图 2）。

对于乌司奴单抗与安慰剂，使用全部特征的策略估计缓解率为33.0%（95%置信区间：27.7%–38.4%），而仅临床策略为30.5%（95%置信区间：24.9%–36.1%）——相差2.5个百分点（95%置信区间：-1.6 至 $+6.6$ 个百分点），与零无显著差异。对于剂量强化，差异可以忽略不计（0.6 个百分点，95%置信区间：-4.6 至 $+6.0$ 个百分点）。

在多臂情境中，内镜特征降低了治疗分配质量。折外评估显示，使用内镜检查的策略缓解率较低（30.5%，95%置信区间：23.1%–37.9%），而仅临床策略的缓解率为36.8%（95%置信区间：29.7%–44.0%），相差 -6.3 个百分点（95%置信区间：-13.3 至 $+0.4$ 个百分点）。样本内估计（58.8%和60.5%）与折外估计（30.5%和36.8%）之间的差异表明，内镜特征过度拟合了噪声，而非捕捉可重复的治疗反应模式。

交叉验证预后模型提供了独立佐证。在全队列（所有治疗组中的患者）中，包含全部特征的模型AUROC为0.663，Brier评分为0.204；而仅使用临床特征的模型表现更优（AUROC为0.684，Brier评分为0.196；$\Delta\mathrm{AUROC} = -0.021$，95%置信区间：$-0.054$ 至 $+0.011$；表5）。增量Brier评分分析证实，内镜特征增加了预测误差（$\mathrm{Brier}_{\mathrm{incremental}} = -3.4\%$），与过拟合的情况一致。

分层分析显示出不同的模式：内镜特征略微改善了安慰剂组的表现（AUROC：0.626 vs 0.572，$\Delta = +0.055$），但降低了每8周一次组的表现（AUROC为0.634 vs 0.658，$\Delta = -0.025$）。这种不一致进一步表明，内镜特征无法稳定地改善结局预测。

亚组ATE分析验证了CATE重要性发现（表 6）。乌司奴单抗相对于安慰剂的总体平均治疗效应为10.9%（95%置信区间：2.4%至19.0%）。按排名靠前的特征划分的亚组呈现出预期的差异：粪便钙卫蛋白 ≥ 150 μg/g的患者平均治疗效应为14.6%（95%置信区间：5.8%至23.0%），低于阈值的患者则为5.4%；生物制剂初治患者的平均治疗效应为13.9%（95%置信区间：0.9%至27.3%），而生物制剂经治患者为8.0%。亚组平均治疗效应的范围为10.9个百分点（从3.7%到14.6%），这证实异质性虽在统计学上可检测到，但其绝对程度较小。

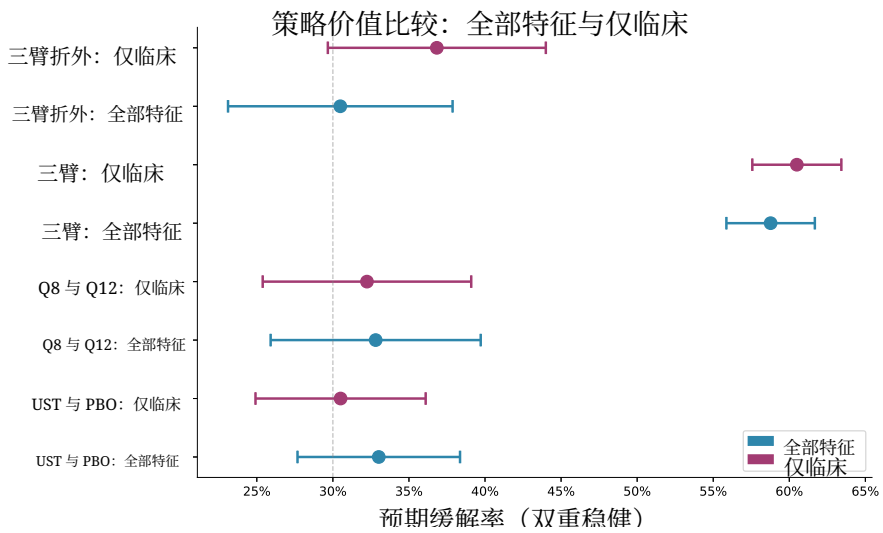

图2 森林图比较各治疗比较中的双稳健策略价值。每个点表示学习得到的策略下的预期缓解率，并附有95%置信区间。蓝色点表示使用全部特征（包括内镜检查）的策略；紫色点表示仅临床策略。多臂折外比较显示，全部特征策略的性能明显较低，提示纳入内镜特征时存在过拟合。

表5 内镜特征的预后性能与边际贡献

队列	模型	AUROC	ΔAUROC	布里尔$_{\text{incr}}$
全队列	全部特征	0.663	−0.021	−3.4%
	仅临床	0.684		
安慰剂	全部特征	0.626	+0.055	+7.5%
	仅临床	0.572		
Q12	全部特征	0.701	+0.001	+2.0%
	仅临床	0.700		
Q8	全部特征	0.634	−0.025	−2.9%
	仅临床	0.658		

注：全队列表示合并全部三个组别（安慰剂 + Q12 + 每8周一次，样本量=561）。 ΔAUROC = AUROC（全部） − AUROC（临床）；正值表示内镜检查提高了区分度。布里尔$_{\text{incr}}$ = IPA（全部） − IPA（临床）；正值表示内镜检查提高了校准度。

5 讨论

这一案例研究阐明了因果机器学习在治疗个性化中的一个关键区别：哪些特征能够预测异质性、异质性在统计学上是否显著，以及异质性是否能改善治疗决策，这些问题可能得出相互矛盾的答案。一个模块化框架分别评估每个问题——通过置换重要性、BLP检验和双重稳健策略

表6 亚组平均治疗效应：UST 对比安慰剂（PBO）

亚组	N	N_{UST}	N_{PBO}	ATE	95%置信区间
总体	561	417	144	10.9%	（2.4%，19.0%）
粪便钙卫蛋白 <150 μ克/克	137	98	39	5.4%	(-13.1%，23.8%)
粪便钙卫蛋白 >150 μ克/克	396	298	98	14.6%	(5.8%，23.0%)
年龄 <中位数	269	204	65	12.1%	(-0.6%, 24.5%)
年龄 >中位数	285	208	77	10.1%	(-1.2%, 21.0%)
C反应蛋白 <5mg/L	404	295	109	14.0%	(3.7%，23.7%)
C反应蛋白 ≥5mg/L	130	102	28	3.7%	(-13.5%，19.3%)
生物制剂初治	266	199	67	13.9%	(0.9%，27.3%)
生物制剂经治	295	218	77	8.0%	(-2.6%，18.1%)
第8 <3周部分Mayo评分	349	256	93	14.4%	(3.2%，25.4%)
部分Mayo评分 wk	185	141	44	6.4%	(-5.1%，16.9%)
8 ≥3					

注：ATEs按各亚组内缓解率之差（UST − 安慰剂）计算。95%置信区间通过自助法重抽样（5000次迭代）获得。亚组按临床阈值定义：粪便钙卫蛋白 ≥150 μg/g（STRIDE-II [22]）、C反应蛋白 ≥5 mg/L（标准炎症阈值 [23]）、部分Mayo评分 ≥3（临床活动度 [24]）。年龄采用中位数二分法（尚无公认的临床界值）。

评估——这对于避免将统计关联过度解读为与决策相关的信号是必要的。在 UNIFI 维持试验中，BLP检验发现内镜特征与估计治疗效果之间存在显著关联（乌司奴单抗相较安慰剂：$z = 8.28$，$p < 0.001$），但双重稳健策略评估显示预期缓解并未相应改善，折外多治疗组评估则表明，纳入内镜特征后表现更差（缓解率为30.5%，而非36.8%） [25]。这种异质性检测与决策表现之间的脱节可能并非该试验独有——只要候选特征包含的预后信息被误认为是效应修饰，就可能出现这种情况。将特征集对结局预测的贡献（增量Brier评分）与其在各治疗组中对策略价值的贡献进行比较，这一诊断策略有助于区分这些作用。内镜评分改善了安慰剂患者的结局预测（增量Brier评分：$+7.5\%$），符合其预后价值；但在治疗队列中，内镜评分降低了预后性能（Q8组降低-2.9%，合并后降低 -3.4%），纳入该特征进行治疗选择也降低了折外策略价值。这种模式可能是特征反映疾病严重程度而非差异性治疗反应的标志。年龄、既往治疗暴露以及第8周炎症生物标志物（C反应蛋白、粪便钙卫蛋白）等临床变量具有较稳定的特征，测量误差也低于内镜评分。在本数据集中，这些常规可获得的指标足以支持所考察的维持治疗决策。亚组ATE分析提供了无模型验证：被确定为异质性主要预测因子的特征，对应的亚组具有明显不同的观察治疗效应（例如，粪便钙卫蛋白 ≥150 μg/g：ATE为14.6%，而低于该阈值时为5.4%）；不过，总体范围为10.9个百分点，证实异质性程度有限。

各项分析结果一致——在二元比较和多臂比较中纳入内镜特征后，策略价值增益仍然有限，折外表现也有所下降——这表明内镜带来的增量价值有限，并非

单一建模选择造成的假象。针对多臂策略进行嵌套交叉验证，明确检验了泛化能力 [4——6]。更广泛地说，这些结果表明，因果机器学习在临床试验中的应用应常规地将策略层面评估与异质性检验结合起来。BLP检验和特征重要性分析有助于描述治疗效应异质性的结构，但无法回答临床实践中最重要的问题：依据估计出的异质性采取行动，能否改善患者结局？折外双稳健策略评估能够回答这一问题，而两者之间的差距本身也可能具有参考价值。尽管内镜评估仍是确认黏膜愈合和指导长期管理的核心手段 [14, 15], 我们的结果表明，对于本文研究的决策，第8周视频衍生内镜评分无法在常规临床标志物的基础上进一步改善对第44周缓解的预测。这些结论并不意味着内镜评估缺乏临床价值，而是说明鉴于现实中的样本量和测量变异，使用基于第8周Mayo评分的汇总指标来选择个体化维持治疗可能具有挑战性。未来研究可考虑减少测量误差（例如采用标准化阅片或自动化分析）、纳入互补模态，并检验内镜信息能否更可靠地应用于其他终点或决策节点。

6 结论

我们提出了一种用于随机对照试验中治疗个性化的因果机器学习框架，旨在回答三个不同的问题：哪些特征能够预测治疗效应异质性、这种异质性在统计学上是否显著，以及它是否能改善治疗决策。该框架将 X-学习器估计、置换重要性、BLP检验和双重稳健策略评估整合到一个模块化流程中，可应用于不同治疗领域。将该框架应用于UNIFI维持试验后，我们发现异质性检测与决策表现明显脱节。内镜特征虽与治疗效应存在显著的统计学关联，但将其纳入治疗分配后，折外策略价值反而下降。通过诊断性比较预后贡献与策略价值，我们发现内镜评分是疾病严重程度标志物，而非稳定的效应修饰因子；相比之下，常规可获得的临床变量捕捉到了与决策相关的变异。这些发现具有普遍意义：将因果机器学习应用于临床试验时，评估候选特征不应只看它们与治疗效应的统计学关联，还应考察它们对样本外治疗决策的影响。需要考虑若干局限性。内镜输入仅限于 Mayo子评分分布和累积疾病评分；其他指数或自动提取的特征可能表现不同。研究未获得组织学数据，主要终点是第44周缓解，而非至事件时间结局。最后，这些结果反映的是乌司奴单抗用于溃疡性结肠炎维持治疗的情况，未必适用于其他生物制剂或克罗恩病。

缩略语列表

ATE - 平均治疗效应 AUROC - 受试者工作特征曲线下面积 BLP - 最佳线性预测器 CATE - 条件平均治疗效应 CDS - 累积疾病评分 IPA - 预测准确性指数 ITT - 意向性治疗 MES - 内镜 Mayo 评分 RCT - 随机对照试验 UC - 溃疡性结肠炎 XGBoost - 极端梯度提升

声明

资金支持

本出版物所报告的研究得到了数据驱动药物开发与治疗评估中心（DATA）的支持。该中心是一个产学合作研究中心，部分经费由美国国家科学基金会根据资助编号2209546提供，另有部分经费来自DATA的行业合作伙伴。

利益冲突/竞争性利益

RS曾担任艾伯维、百时美施贵宝、CorEvitas、礼来、Exact Sciences、吉利德、杨森、默克、辉瑞和武田的顾问或顾问委员会成员。RS和KN拥有医学影像及内镜分析技术方面的知识产权和股权，这些技术由密歇根大学授权给PreNovo, LLC、AMI, LLC和PathwaysGI, Inc.。其余作者声明不存在利益冲突。

伦理审批与参与知情同意

UNIFI研究（NCT02407236）按照《赫尔辛基宣言》原则和药物临床试验质量管理规范开展。各参与中心的机构审查委员会或独立伦理委员会均批准了原始方案及所有修订内容，且所有参与者在入组前均提供了书面知情同意。本分析使用的是去标识化数据。由于这项二次分析仅使用匿名化数据，因此无需额外获得机构审查委员会批准或知情同意。

数据和材料的可用性

本分析使用了 UNIFI 乌司奴单抗治疗溃疡性结肠炎 III 期维持治疗试验（NCT02407236）的患者个体数据，该数据依据与试验申办方签订的数据共享协议获得。该数据集未公开，但符合条件的研究人员可根据申办方的数据共享策略申请访问。

作者贡献

数据获取：RS；资金获取：CM；项目构思：CM、RS、KN；数据预处理：RS、CM、ST；软件：CM；正式分析：CM；初稿：CM；最终修订：CM、RS、KN。

参考文献

[1] Van Vogt E, Gordon AC, Diaz-Ordaz K, Cro S. 将因果森林应用于随机对照试验数据，以识别异质性治疗效应：一项案例研究。*BMC Med Res Methodol*. 2025;25:50.[2] Athey S, Imbens GW. 使用随机森林估计和推断异质性治疗效应。*J Am Stat Assoc*. 2018;113(523):1228-1242.[3] K¨unzel SR, Sekhon JS, Bickel PJ, Yu B. 使用机器学习估计异质性治疗效应的元学习器。*Proc Natl Acad Sci USA*. 2019;116(10):4156-4165.[4] Robins JM, Rotnitzky A, Zhao LP. 当部分回归变量并非总是可观测时，回归系数的估计。*J Am Stat Assoc*. 1994;89(427):846-866.[5] Bang H, Robins JM. 缺失数据和因果推断模型中的双重稳健估计。*Biometrics*. 2005;61(4):962-973.[6] Kennedy EH. 迈向异质性因果效应的最优双重稳健估计。*Electron J Stat*. 2023;17(2):3008-3049.[7]Athey S, Wager S. 基于观察性数据的策略学习。*Econometrica*. 2021;89(1):133-161.[8] Kitagawa T, Tetenov A. 谁应该接受治疗？治疗选择的经验福利最大化方法。*Econometrica*. 2018;86(2):591-616.[9] Chernozhukov V, Demirer M, Duflo E, Fern´andez-Val I. 随机实验中异质性治疗效应的通用机器学习推断。NBER 工作论文第24678号，2018.[10] Sands BE, Sandborn WJ, Panaccione R, et al. 乌司奴单抗用于溃疡性结肠炎的诱导和维持治疗。*N Engl J Med*. 2019;381(13):1201- 1214.[11] Ceccato HD,Silva TAOE, Genaro LM, et al. 人工智能在炎症性肠病精准医学中的应用：系统综述。*Am J Transl Res*. 2025;17(1):28-46.[12]Gisbert JP, Chaparro M. 炎症性肠病患者生物制剂治疗的初始应答预测因素：[抗肿瘤坏死因子、维得利珠单抗和乌司奴单抗] 在炎症性

肠道疾病：从基础科学到临床实践。*J Crohns Colitis*. 2020;14(5):694-709.[13] SchroederKW, Tremaine WJ,Ilstrup DM. 口服包衣5-氨基水杨酸治疗轻至中度活动性溃疡性结肠炎：一项随机研究。*N EnglJ Med*. 1987;317(26):1625-1629.[14] Travis SP, Schnell D, Krzeski P, et al. 溃疡性结肠炎内镜严重程度指数的可靠性和初步验证。*Gastroenterology*. 2013;145(5):987- 995.[15] Shah SC, Colombel JF, Sands BE, Narula N. 黏膜愈合与溃疡性结肠炎患者长期结局改善相关：系统综述与荟萃分析。*Clin Gastroenterol Hepatol*. 2016;14(9):1245-1255.[16] Harbord M, Eliakim R, Bettenworth D, et al. 第三版欧洲溃疡性结肠炎诊断与管理循证共识。第二部分：当前管理。*J Crohns Colitis*. 2017;11(7):769-784.[17]Afif W, Arasaradnam RP, Abreu MT, et al. 乌司奴单抗治疗溃疡性结肠炎4年的疗效和安全性：UNIFI长期维持研究的最终结果。*Am J Gastroenterol*. 2024;119(5):910-921. [18] Stidham RW, Cai L, Cheng S, et al. 利用计算机视觉改进溃疡性结肠炎治疗性临床试验中的内镜疾病定量评估。*Gastroenterology*. 2024;166(1):155-167.e2.[19] Breiman L. 随机森林。*Mach Learn*. 2001;45:5-32.[20]Fisher A, Rudin C, Dominici F. 所有模型都有缺陷，但许多模型很有用：通过同时研究整类预测模型来学习变量的重要性。*J Mach Learn Res*. 2019;20(177):1-81.[21] Kattan MW, Gerds TA. 预测准确性指数：一种直观且有助于评估风险预测模型的指标。*Diagn Progn Res*. 2018;2:7.[22] Turner D, Ricciuto A, Lewis A, et al. STRIDE-II：国际炎症性肠病研究组织（IOIBD）的炎症性肠病治疗目标选择（STRIDE）倡议更新：确定炎症性肠病达标治疗策略的治疗目标。*Gastroenterology*. 2021;160(5):1570-1583.[23] Vermeire S, Van Assche G, Rutgeerts P. 炎症性肠病的实验室标志物：有用、神奇，还是多余的摆设？*Gut*. 2006;55(3):426-431.[24] Lewis JD, Chuai S, Nessel L, et al. 使用Mayo评分的非侵入性组成部分评估溃疡性结肠炎的临床应答。*Inflamm Bowel Dis*.

2008;14(12):1660-1666。

[25] Steyerberg EW。《临床预测模型：开发、验证与更新的实用方法》。第二版。施普林格。2019。