arXiv:2603.03587v1 [stat.ME] 3 Mar 2026

用于因果推断的可控生成式沙盒

Qi Zhang埃默里大学
qi.zhang2@emory.edu

Harsh Parikh耶鲁大学
harsh.parikh@yale.edu

AshleyNaimi
大学 ashley.naimi@emory.edu

RaziehNabi埃默里大学 埃默里
razieh.nabi@emory.edu

Christopher Kim安进
chrkim@amgen.com

Timothy Lash埃默里大学
timothy.lee.lash@emory.edu

摘要

在因果推断中，方法验证与研究设计依赖于具有已知反事实的合成数据。现有模拟器需要在分布真实性与因果可控性之间权衡：前者指捕捉混合类型和多模态表格数据的能力，后者则包括对重叠、未测量混杂和治疗效应异质性的显式控制。我们提出 CAUSALMIX，这是一种变分生成框架，通过将高斯潜在先验的混合分布与针对连续型、二元和类别型变量的数据类型专用解码器相结合，弥合了这一差距。该模型纳入了显式的因果控制：利用重叠性正则项塑造倾向评分分布，并直接参数化混杂强度和效应异质性。这一统一目标函数在保持观测数据保真度的同时，支持对因果机制进行因子式操控，使重叠、混杂强度和治疗效应异质性能够在设计阶段独立变化。在各项基准测试中，CAUSALMIX 在混合类型表格数据上取得了最先进的分布指标，同时提供稳定、精细的因果控制。我们通过一项针对转移性去势抵抗性前列腺癌治疗的比较安全性研究，展示了其实际用途：使用 CAUSALMIX 在经过校准的数据生成过程中比较估计量、调整超参数，并在针对特定治疗效应异质性场景的条件下开展基于模拟的效能分析。

1 引言

对于每个个体，我们只能观测到一种潜在结果，因此个体因果效应从根本上说无法观测 [1]。因此，要评估因果方法，就需要在真实值已知的情境中进行。合成数据 因此成为因果方法学的得力工具：它提供真实反事实、标准化基准，以及可控的识别假设压力测试。在这个沙盒中，研究人员可以探究因果方法对混杂、模型设定错误和重叠不足的稳健性；刻画有限样本行为；并在贴近预期应用的条件下调整超参数（例如因果森林深度、超级学习器库和交叉拟合折），同时仍能掌握真实值 [2； 3]。受真实数据启发的模拟还支持前瞻性研究设计（针对特定异质性／混杂情形评估统计功效和样本量），并可通过发布逼真的不可识别替代数据来促进隐私保护协作 [4]。

在合成数据上的表现往往无法推广到复杂的真实场景，因为估计量行为取决于建模假设与数据特征的契合程度，包括依赖结构、非线性、混杂强度、重叠、结局变异性和效应量。因此，理想的基准必须既能提供对（因果）真实值的访问，又能忠实再现真实数据生成过程（DGP）的关键特征 [5； 6]。

研究空白。 尽管用于因果数据的深度生成架构已取得显著进展 [7; 8; 9], , 现有方法仍面临一个根本性挑战：如何同时实现对混合类型观察性数据的逼真建模，以及在设计阶段对效应异质性、混杂和重叠等关键因果属性进行显式控制。因此，经验真实性的提升往往以牺牲因果透明度为代价，限制了这些方法在方法评估和研究规划中的实用性。

贡献。 我们提出 CAUSALMIX，这是一种将分布真实性与细粒度因果控制相统一的变分生成框架。

• **混合类型保真度**。 一种采用贝叶斯高斯混合先验的混合潜变量架构，以及针对数据类型的解码器头，能够捕捉观察性数据集和临床数据集中常见的多峰依赖关系。

• **因果调节机制**。 我们提供用于异质性治疗效应、倾向评分重叠性和未测量混杂的设计阶段 控制机制。通过对数似然比正则项调节重叠程度，直接塑造治疗分配；混杂和效应异质性则在结局机制中进行参数化。

• **稳定的因果保真度**。 正则化和方差约束可确保预先指定的因果函数和对照在训练过程中得到忠实实现，尤其是在因果机制为低维或弱非线性时。

• **统一目标函数**。 该模型联合优化分布拟合与因果约束，确保在匹配观测数据的同时，不会削弱对指定因果结构的控制。

我们进一步提出一套用于因果合成数据的集成评估与应用流程，在合成数据发布的背景下，联合评估分布保真度、因果保真度和记录级披露风险方面的隐私。实践者需要经过校准的沙盒，既能呈现出与其数据相似的特征，又支持围绕重叠、混杂和异质性开展有针对性的“假设情景”研究。CAUSALMIX 支持：（i）在基于经验的 DGP 下进行公平的估计量比较；（ii）在同一校准条件下，以有原则的方式进行超参数选择；（iii）开展基于模拟的研究设计（在异质性和残余混杂可调的情况下进行统计功效／样本量计算）。我们以一个比较安全性场景为例说明这些优势（阿比特龙与恩扎卢胺用于治疗转移性去势抵抗性前列腺癌），其中观察性数据呈现出真实的数据结构和合理的偏倚。代码已公开发布于 https://github.com/zhangqiecho/causalmix。

2 背景

2.1 设置与符号

我们观测到独立同分布样本 $O = (X,\ T,\ Y)$，其中 $X \in \mathcal{X}$ 表示协变量，$T \in \{0, 1\}$ 表示二元处理，Y 表示观测结局。令 $Y(1)$ 和 $Y(0)$ 为潜在结果，根据一致性，$Y = T\,Y(1) + (1 - T)\,Y(0)$。我们将条件结局回归定义为 $\mu_t(x) \triangleq \mathbb{E}[Y(t) \mid X = x]$，将倾向评分定义为 $e(x) \triangleq \mathbb{P}(T = 1 \mid X = x)$ [10], , 并考虑平均处理效应 ATE $= \mathbb{E}[Y(1) - Y(0)]$ 以及条件平均处理效应 CATE$(x) = \mu_1(x) - \mu_0(x)$。

从观察性数据中识别平均处理效应和条件处理效应，通常依赖于以下假设：一致性、给定观测协变量 X 时的无混杂性，以及正值性。然而在实践中，因果估计量的有限样本行为很大程度上取决于混杂强度、模型复杂度和倾向得分重叠程度。未测量混杂因素导致的可忽略性违背，以及处理组之间重叠不足，都会削弱点识别、增大方差，并迫使我们在重叠不足的区域对反事实结果进行外推，从而使估计和方法比较都更加复杂。

在此类情境下，具有已知且可控因果结构的合成数据提供了一个有原则的测试平台，可用于研究这些挑战：它能够系统地控制未测量混杂，并

重叠，支持通过已知真实值进行诊断性评估，并有助于在偏离理想识别条件的现实情况下，对因果估计量进行公平的基准评测。

2.2 合成数据在因果推断中的背景

当反事实真值无法直接观测时，合成数据是算法开发、公平比较和研究设计的基础 [1]。我们按照与真实数据的紧密程度，对三类方法进行概述。

任意参数化模拟器。 经典设置为（X、 T、 Y）指定结构方程，从而能够完全控制效应大小、函数形式和噪声。它们易于调整和分析，但可能导致方法排名更多地反映设计者的选择，而非现实世界的复杂性，尤其是在处理混合类型、多模态表格数据时。

半合成基准。 这些方法重用 X 真实数据集中的协变量，同时根据固定机制模拟 $T|X$和 $Y|T$、X。这一范式是广泛使用的挑战赛和排行榜的基础 [11； 2； 12； 13],它支持可复现的比较，但对重叠和异质性等因果“旋钮”的控制有限且较为间接，在敏感领域还可能受到隐私约束的影响。

数据拟合生成器。一种较新的方向是从真实数据中学习合成数据生成过程（DGP），使边缘分布和条件分布与经验结构相匹配。*Plasmode* 模拟保留观测到的 X 和（有时） T，同时根据已知模型生成结局，以注入真实值 [14； 15； 16； 17]；这类方法透明，但可能缺乏灵活性，且对设计选择较为敏感[18； 19； 20； 21]。神经网络方法利用灵活的生成模型拟合 $T|X$（或 $X|T$）和 $Y|T,\ X$，其中包括 REALCAUSE 中的归一化流 [7], 、Wasserstein生成对抗网络（WGAN） [8], 以及条件变分自编码器（VAE）框架 [9； 22]。这些模型提高了分布保真度，但可能面临训练不稳定（对于 GAN） [23], 、对混合类型表格数据的支持有限 [9], ，或只能对重叠、未测量混杂和效应异质性进行粗略且间接的控制 [7]。
强调经验真实性的方法（如 REALCAUSE 和基于 WGAN 的生成器）能够紧密模拟观测数据分布，但对因果参数的显式控制有限。例如，REALCAUSE 通过在拟合的极端值之间插值来调整效应尺度、异质性和重叠，同时假设不存在未测量混杂 [7], ；而基于 WGAN 的方法缺乏直接操控治疗效应的机制 [8]。Frengression [24] 通过干预建模实现对边际因果量的直接控制，但无法显式设定条件效应、重叠或未测量混杂，而这些因素对于评估观察性场景中估计量的稳健性至关重要。CREDENCE [9] 允许用户预先指定治疗效应和未测量混杂偏倚的函数形式，但在模拟逼真的混合类型、多峰表格数据方面能力有限。此外，它缺乏系统的实证验证，无法确认学习得到的生成器是否确实满足预先指定的因果函数和混杂控制。

基准的期望准则。在各种应用中，以下三项性质尤为重要：**(i) 混合类型真实性**——忠实地模拟具有多峰性的连续型、二元和类别型变量；**(ii) 因果可控性**——将效应量和异质性、未测量混杂以及重叠视为明确的设计参数；以及 **(iii)联合优化**——在不顾此失彼的前提下，同时实现分布保真度和因果控制。这些期望准则推动了本文所提出框架的开发。该框架在基于条件变分自编码器的生成器上扩展了混合先验、数据类型特定解码器，以及对处理效应、重叠和混杂的显式控制。

3 方法

3.1 问题表述

本工作的主要目标是学习一种灵活的数据生成机制 G_θ，使生成样本的联合分布 $\mathcal{O}' = (X',\ T',\ Y') \sim G_\theta$能够紧密近似观测数据的经验分布 $p_{\text{real}}(X,\ T,\ Y)$，同时还能明确地由用户控制关键因果属性。为实现这一目标，我们采用非参数结构因果模型（SCM），其中所有结构方程均为不受限制的映射。

设观测样本 $\mathcal{O} = \{(X_i,\ T_i,\ Y_i)\}_{i=1}^{n}$ 由潜在因果过程 $G^{\dagger} = [\phi_X, \phi_T, \phi_{Y(1)}\ \phi_{Y(0)}]$ 生成，该过程由以下内容指定：

$$\begin{aligned} X &= \phi_X(U_X), \\ T &= \phi_T(X, U_T), \\ Y(t) &= \phi_{Y(t)}(X, U_Y), \quad t \in \{0,1\}, \\ Y &= T\,Y(1) + (1-T)\,Y(0), \end{aligned}$$

其中，U_X、U_T和U_Y是相互独立的外生变量，且每个 $\phi_\cdot$ 都是任意的非线性函数。该结构因果模型诱导出真实的联合分布 $p_{\text{real}}(X,\ T,\ Y)$，我们的生成器旨在学习并近似该分布。

因果控制函数。除了学习与观测数据相匹配的分布外，我们还要求能够指定生成的合成数据中的关键因果参数。我们通过引入一组用户指定的、依赖协变量的控制函数，将因果控制形式化为三个维度：

$$\Psi = \{\alpha(X),\, \tau(X),\, \kappa(X,T)\}.$$

每个函数都控制着结构因果模型中的一种不同因果属性：

• **重叠**指处理组 $P(X \mid T = 1)$与未处理组 $P(X \mid T = 0)$的协变量分布有多相似。为了指定所需的正值性/重叠结构，我们引入一个依赖协变量的条件密度比

$$\alpha(x) = \frac{P(X = x \mid T = 0)}{P(X = x \mid T = 1)}.$$

$\alpha(X)$的值接近1表示处理组之间具有较强的重叠；较大或较小的值则对应重叠较弱的区域，或近似正值性违反的区域。在这些区域中，协变量空间分别由未处理组或处理组主导。

• **效应函数** $\tau(X)$ 定义了条件平均处理效应：

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x],$$

决定协变量间因果效应的大小 和异质性。

• **未测量混杂** $\kappa(X,T)$通过共享潜在因素，使 T与潜在结局 $Y(t)$之间产生依赖关系，从而刻画超出 X范围的隐藏混杂程度及其形式。其定义如下：

$$\kappa(x,t) = \mathbb{E}[Y(t) \mid X = x, T = 1] - \mathbb{E}[Y(t) \mid X = x, T = 0].$$

$\kappa(x,t)$通过其在不同处理组的潜在结果中引发的选择偏倚，对结构性未测量混杂进行编码，进而决定其偏离条件可忽略性的程度。

半参数因果推断的最新进展已将识别与估计理论扩展至含潜变量的复杂有向无环图，使得能够在经典后门和前门设定之外一致地估计因果泛函 [25；26；27]。这些方法采用灵活的机器学习估计量，结合一步校正或目标最小损失程序，在较弱的建模假设下获得双重稳健且渐近有效的估计。然而，其实际可靠性在很大程度上取决于底层数据生成过程的性质，包括重叠、干扰参数复杂度，以及处理分配与潜在混杂之间的相互作用。由于无法在观察性数据集中系统地改变这些因素，因此，严谨的评估需要一种能够独立控制这些因果机制的生成器。上文介绍的控制函数正是为支持此类估计量层面的压力测试而设计的。

联合优化问题。给定因果控制函数 $\Psi = \{\alpha, \tau, \kappa\}$，我们希望学习一个数据生成器 G_θ，使其生成样本 $\mathcal{O}' = (X',\ T',\ Y') \sim G_\theta$所形成的分布尽可能接近经验联合分布 $p_{\text{real}}(X,\ T,\ Y)$，同时忠实于 Ψ 中编码的因果结构，其严格程度由用户控制。

对于任意生成器 G_θ，我们使用生成样本上的期望来定义其诱导的因果量：

$$\alpha_\theta(x) := \frac{p_\theta(X = x \mid T' = 0)}{p_\theta(X = x \mid T' = 1)},$$

$$\tau_\theta(x) := \mathbb{E}[Y'(1) - Y'(0) \mid X = x],$$

$$\kappa_\theta(x,t) := \mathbb{E}[Y'(t) \mid X = x, T = 1] - \mathbb{E}[Y'(t) \mid X = x, T = 0].$$

严格满足因果规格的生成器子集为

$$\mathcal{G}^*_\Theta(\Psi) := \left\{ G_\theta : \begin{array}{c} \alpha_\theta(X) = \alpha(X), \\ \tau_\theta(X) = \tau(X), \\ \kappa_\theta(X,T) = \kappa(X,T) \end{array} \right\}.$$

同样，令

$$\mathcal{G}^\dagger_\Theta := \{G_\theta : p_\theta(X', T', Y') \approx p_{\text{real}}(X, T, Y)\}$$

为能够重现经验分布的生成器集合。若 $\mathcal{G}^\dagger_\Theta \cap \mathcal{G}^*_\Theta(\Psi) \neq \emptyset$，则理想解位于该交集中；否则，我们会在 $\mathcal{G}^*_\Theta(\Psi)$ 中寻找一个生成器，使其在选定的差异度下与 $\mathcal{G}^\dagger_\Theta$ 的距离最小。

由此得到以下优化原则：

$$\begin{aligned} \min_\theta \mathbb{E}\big[d\big((X,T,Y),(X',T',Y')\big)\big] \\ + \lambda_\alpha \, \mathbb{E}_X[\mathcal{R}_\alpha(\alpha_\theta(X), \alpha(X))] \\ + \lambda_\tau \, \mathbb{E}_X[\mathcal{R}_\tau(\tau_\theta(X), \tau(X))] \\ + \lambda_\kappa \, \mathbb{E}_{X,T}[\mathcal{R}_\kappa(\kappa_\theta(X,T), \kappa(X,T))]\,. \end{aligned}$$

其中，$d(\cdot,\ \cdot)$ 是一种分布距离，$\mathcal{R}_\alpha$，$\mathcal{R}_\tau$，$\mathcal{R}_\kappa$ 是刚性惩罚项，用于约束对重叠、处理效应和混杂规格的遵循程度；λ_α，λ_τ，λ_κ 是正则化参数。

第一项确保与经验分布保持一致，而惩罚项则引导 G_θ 朝着用户指定的因果结构优化。该框架学习一种数据生成机制，使其近似真实的观测过程，同时能够沿三个可解释的因果维度进行受控变化。

问题陈述：我们的目标是开发一种模型架构、训练算法和数据生成程序，使其能够：

- 有效优化统一训练目标，以找到能够平衡所有目标的参数值
- 以高保真度满足因果约束（约束在可接受的容差范围内得到满足）
- 处理混合类型表格数据且不产生预处理伪影
- 通过易于调节的超参数，对因果参数提供可解释的控制
- 高效生成高质量合成样本

以下各节将介绍我们针对这一问题表述提出的解决方案，并说明模型架构、先验分布、因果正则化机制和训练流程的具体设计选择。

3.2 模型概述与数据生成过程

CAUSALMIX 实现了一种基于条件 VA G_θ 构建的模块化数据生成机制
Ł 观测分布分解为

$$p(X,T,Y) = p(T)\, p(X \mid T)\, p(Y \mid X, T),$$

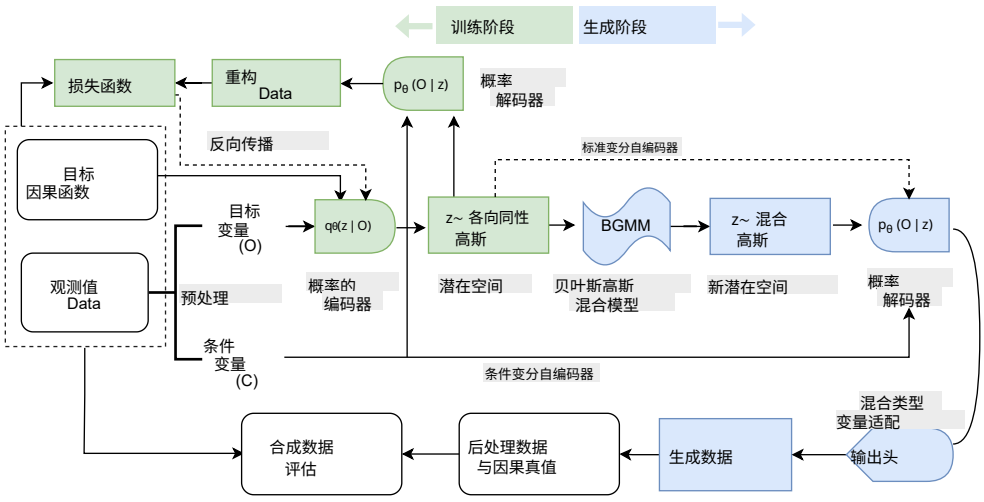


图1：CAUSALMIX 生成框架概览。

and 由三个组成部分建模：用于 T 的处理模型、用于 $X \mid T$ 的处理前生成器 ，
and 一个用于潜在结果的处理后生成器（$Y(0)$ $Y(1)$）$\mid X,\ T$。

模块化生成器。 给定变量元数据，CAUSALMIX封装器负责预处理并实例化：(i) 用于治疗分配的伯努利模型，(ii) 条件变分自编码器 $G_{X,\theta}$ 用于$X \mid T$，以及 (iii) 条件变分自编码器 $G_{Y,\theta}$用于 $(Y(0)\ Y(1)) \mid X, T$。合成数据通过依次采样 T'、然后 $X' \mid T'$、最后 $(Y'(0)\ Y'(1)) \mid X', T'$ 顺序生成，观测结局构造为 $Y' = T'Y'(1) + (1 - T')Y'(0)$。生成样本利用预处理元数据被确定性地变换回原始数据尺度与表示（例如逆标准化、重新整数化与类别解码）。

条件变分自编码器骨干网络。 两个生成器共享同一条件变分自编码器结构，采用多层感知机（MLP）编码器与解码器。编码器将目标变量映射为潜在表示，并输出对角高斯后验的均值与对数方差。解码器将拼接后的潜在编码与条件变量（T或 (X, T)）映射到共享隐藏表示，再经由面向二元、类别型与连续变量的变量特定输出头进行路由，从而实现对混合类型表格数据的真实建模。

潜在结果与因果约束。 处理后生成器 $G_{Y,\theta}$ 对两个潜在结果进行联合建模，从而能够直接获取个体层面处理效应。训练期间，重构以观测到的治疗组为条件，同时通过额外的正则化，确保模型与用户指定的处理效应函数和混杂函数相一致。我们引入稳定机制，以确保这些因果目标得到忠实实现；相应的目标函数将在第3.3节中介绍。

重叠控制。 协变量生成器 $G_{X,\theta}$ 采用了显式的重叠控制机制，通过调整条件分布 $p(X \mid T)$，使其符合用户指定的对数密度比。这让用户能够在设计阶段直接控制倾向得分的重叠；具体实现细节将在第3.3节中介绍。

潜在先验。 CAUSALMIX的一个关键设计选择是使用灵活的基于混合的潜在先验 来捕捉观察性数据中的多峰结构。与其将潜在空间限制为标准高斯分布，每个条件变分自编码器都可以在训练后于潜在空间中拟合贝叶斯高斯混合模型（BGMM）。从该混合先验中抽取的潜变量样本随后被条件解码，以生成 X' 和（$Y'(0)\ Y'(1)$），从而提升模型表示复杂、异质分布的能力。

总体而言，这种模块化设计使CAUSALMIX能够生成保留经验结构的合成数据集，同时对重叠、未测量混杂和处理效应异质性提供显式控制。图1概述了CAUSALMIX框架。

3.3 目标函数

通过优化统一目标函数来学习生成机制 G_θ 的参数 θ。该目标函数将生成建模与用于约束用户指定因果结构的惩罚项相结合，包含三个部分：(i) 用于保证分布保真度的变分自编码器损失；(ii) 使诱导出的因果量与目标因果函数相一致的惩罚项；以及 (iii) 用于控制协变量分布 $p_\theta(X \mid T)$ 的重叠惩罚。下面依次介绍各个部分。

生成建模目标。CAUSALMIX 中的每个模块都实现为一个条件变分自编码器。对于一批观测数据 $(X,\ T,\ Y)$ 和潜变量 $Z \sim q_\theta(z \mid \cdot)$，根据各变量的类型选用相应的似然函数进行重构：连续变量使用高斯似然（高斯负对数似然，而非 CREDENCE 中使用的均方误差），二元变量使用伯努利似然（逻辑交叉熵），离散变量使用类别似然（Softmax交叉熵）。

训练过程通过最大化证据下界（ELBO）来进行。对于观测变量 $\mathcal{O}$（即 X 或 Y），ELBO 为

$$\mathrm{ELBO}(\theta) = \mathbb{E}_{q_\theta(z)}\big[\log p_\theta(\mathcal{O} \mid z)\big] - \mathrm{KL}\big(q_\theta(z) \,\|\, p(z)\big),$$

其中 p（z）表示潜在先验。等价地，最小化负证据下界可得

$$-\mathrm{ELBO}(\theta) = -\mathbb{E}_{q_\theta(z)}\big[\log p_\theta(\mathcal{O} \mid z)\big] \ + \ \mathrm{KL}\big(q_\theta(z) \,\|\, p(z)\big).$$

The 编码器将对角高斯后验 q_θ（z）参数化为 $= \mathcal{N}$（μ_θ，diag(σ_θ^2)），对于该后验，e KL 散度有闭式解

$$\mathrm{KL}\big(q_\theta(z) \,\|\, p(z)\big) = \frac{1}{2}\sum_{j=1}^{d}\big(\mu_{\theta,j}^2 + \sigma_{\theta,j}^2 - \log\sigma_{\theta,j}^2 - 1\big).$$

将重构项与 KL 项结合，可得标准 VAE 损失

$$\mathcal{L}_{\mathrm{VAE}}(\theta) = -\mathbb{E}_{q_\theta(z)}\big[\log p_\theta(\mathcal{O} \mid z)\big] \ + \ \lambda_{\mathrm{KL}}\,\mathrm{KL}\big(q_\theta(z) \,\|\, p(z)\big),$$

其中 λ_{KL} 控制潜变量正则化的强度。在处理后生成器中，重构以观测到的治疗组为条件，并且仅使用相应的结局 Y（T）。

因果参数控制。为确保学习到的生成器能够实现所需的因果结构，我们针对由用户指定的控制函数定义的诱导因果量引入惩罚项：重叠度的 $\alpha(X)$、处理效应幅度和异质性的 $\tau(X)$，以及未测量混杂的 $\kappa(X,\ T)$。每项约束的强度由刚性参数 $\lambda_\alpha,\ \lambda_\tau,\ \lambda_\kappa \geq 0$ 控制。

处理效应控制。为使学习到的处理效应函数与其目标相一致，$\tau_\theta(X) \approx \tau(X)$，我们采用加权复合惩罚（Huber损失）来塑造条件平均处理效应残差 $\Delta\tau_\theta(X) = \tau_\theta(X) - \tau(X)$ 的分布。二次项锚定这些残差的均值对齐，从而隐式稳定诱导的平均处理效应；平滑L1项则提高了对离群值和弱识别区域的稳健性。当目标函数为低维或弱非线性时，这会在精确对齐与稳定性之间权衡。单独的方差正则项可抑制围绕目标因果结构出现虚假的个体层面离散，从而稳定训练，同时避免预期的异质性坍缩。

$$\mathcal{L}_\tau^{\mathrm{mean}}(\theta) = \mathbb{E}_X\Big[\lambda_\tau^{\mathrm{mse}}(\Delta\tau_\theta)^2 + \lambda_\tau^{\mathrm{sl1}}\,\mathrm{SmoothL1}(\Delta\tau_\theta)\Big], \quad \text{and} \quad \mathcal{L}_\tau^{\mathrm{var}}(\theta) = \mathrm{Var}\big[\Delta\tau_\theta\big].$$

未测量混杂控制。令 $\kappa_\theta(X,\ T)$ 表示由结果解码器引起的混杂偏倚。我们对混杂残差$\Delta\kappa_\theta$ $(X,\ T) = \kappa_\theta(X,\ T) - \kappa(X,\ T)$施加类似的复合惩罚，使用二次项和平滑L1项实现均值对齐和稳健性，同时通过方差正则项控制过大的个体层面离散度：

$$\mathcal{L}_{\kappa}^{\text{mean}}(\theta) = \mathbb{E}_{X,T}\Big[\lambda_{\kappa}^{\text{mse}}(\Delta\kappa_\theta)^2 + \lambda_{\kappa}^{\text{sl1}}\,\text{SmoothL1}(\Delta\kappa_\theta)\Big], \quad \text{and} \quad \mathcal{L}_{\kappa}^{\text{var}}(\theta) = \text{Var}\big[\Delta\kappa_\theta\big].$$

重叠控制。 CAUSALMIX 的一项关键特性是通过对数密度比直接控制处理组之间的协变量重叠

$$\log \alpha_\theta(X) = \log p_\theta(X \mid T' = 0) - \log p_\theta(X \mid T' = 1),$$

用于量化协变量特征在处理相对于对照下的相对似然。通过对 $\log \alpha_\theta(X)$ 偏离用户指定目标 $\log \alpha(X)$ 的程度施加惩罚，生成器可以诱导从强重叠到近似正值性违反的各种情形。

条件对数密度 $\log\ p_\theta(X \mid T' = t)$ 由逐特征解码器似然计算得到。二元、类别型和连续变量分别贡献伯努利、Softmax 和高斯对数密度，详见附录A。重叠惩罚的形式为

$$\mathcal{L}_{\alpha}(\theta) = \mathbb{E}_X\left[\big(\log \alpha_\theta(X) - \log \alpha(X)\big)^2\right],$$

其中 λ_α 控制约束的严格程度。

统一目标函数。 完整的学习目标将分布保真度与显式因果控制相结合：

$$\begin{aligned}\mathcal{L}(\theta) = \mathcal{L}_{\text{VAE}}(\theta) + \lambda_\alpha\,\mathcal{L}_\alpha(\theta) + \lambda_\tau\,\mathcal{L}_\tau^{\text{mean}}(\theta) + \lambda_\tau^{\text{var}}\,\mathcal{L}_\tau^{\text{var}}(\theta) \\ + \lambda_\kappa\,\mathcal{L}_\kappa^{\text{mean}}(\theta) + \lambda_\kappa^{\text{var}}\,\mathcal{L}_\kappa^{\text{var}}(\theta).\end{aligned}$$

最小化 $\mathcal{L}(\theta)$ 得到一个生成器，它在逼近观测分布的同时，忠实地实现用户指定的重叠、处理效应和混杂结构。

3.4 混合变量类型

真实世界的表格数据通常包含异质类型的变量，包括连续变量、二元变量、分类变量和整数值特征。为在统一的生成框架中适应这种异质性，每个条件变分自编码器模块都采用多头解码器，并为每个变量分配与其数据类型相匹配的似然族。

二元变量使用由解码器对数几率参数化的伯努利似然建模。分类变量使用 Softmax 对数几率建模，其维度与类别数量相匹配。连续变量使用高斯似然建模，均值由解码器生成，对数方差则受到限制，从而使模型能够捕捉异方差性和变量特定的不确定性。在我们的实验中，对连续变量使用高斯负对数似然，而非 CREDENCE 中采用的均方误差，显著提升了模型性能，尤其是对于方差存在异质性或取值范围受限的变量。这一提升源于负对数似然目标函数允许解码器在保持梯度尺度适当的同时，学习位置和离散程度。

整数值变量（例如，以浮点数存储的计数）在标准化后按连续变量处理，后处理时可选择进行舍入。在潜在结果模式下，每个变量头都会为 $Y(0)$ 和 $Y(1)$ 输出单独的参数集；为确保数值稳定性，此模式不允许使用多类别分类结果。

训练和重叠计算中使用的各变量精确对数似然表达式（伯努利、Softmax 和高斯）见附录A，其中总对数密度由各个特征独立贡献的对数密度求和得到。总体而言，这种多头解码器设计能够在变分自编码器框架内，准确且灵活地对混合类型的观察性数据进行建模。

3.5 高斯混合模型先验

标准变分自编码器通常假设各向同性高斯潜变量先验，$p(z) = \mathcal{N}(0,\ I)$。然而，近期关于表格数据的变分自编码器与高斯混合模型集成的研究 [28] 表明，由于复杂的非线性

特征依赖，真实数据集通常会产生显著偏离单峰高斯分布的潜在表示。因此，从标准高斯先验中采样可能会扭曲所学习潜在空间的几何结构，并降低生成样本的真实性；在具有多峰性和聚类特征的异构表格领域中，这一局限尤为明显。

为解决这一问题，CAUSALMIX 将固定的高斯先验替换为在 VAE 训练后对编码器的潜在均值进行事后拟合的贝叶斯高斯混合模型 (BGMM)。对于每个条件变分自编码器模块，我们收集编码器输出 $\{\mu_\theta(\mathcal{O}_i)\}_{i=1}^{N}$，并使用狄利克雷过程先验拟合 BGMM，使其能够自适应地推断表示潜在分布所需的混合分量数量。由此得到的先验形式为

$$p_{\mathrm{BGMM}}(z) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(z \mid \mu_k, \Sigma_k),$$

其中，π_k 表示分量 k的混合权重，μ_k 和 Σ_k 分别表示其均值和协方差，且$\sum_k \pi_k = 1$。狄利克雷过程先验鼓励混合权重保持稀疏，因此只有一部分分量会获得显著的概率质量。实践中，K 使用截断折棍变分推断从数据中学习，其上界由潜在维度设定。正如 [28], 所强调的，这种事后建模保留了原始的 VAE 训练目标，同时更忠实地近似了所学习的潜在几何。

生成时，从 $p_{\mathrm{BGMM}}(z)$中抽取潜变量样本，可选择添加与平均编码器方差相符的噪声，再与条件特征（例如 T' 或$(X',\ T')$）拼接，并传入解码器主干，以生成共享隐藏表示；随后，各变量专属的输出头基于该表示生成模型参数。

BGMM先验有两项关键优势。首先，通过将潜在空间建模为混合分布而非单一高斯分布，它能够捕捉异构表格数据中固有的多峰结构，从而生成更逼真的合成样本。其次，它能与我们的因果控制框架无缝结合：重叠、处理效应和未测量混杂在解码器层面得到约束，而混合先验则保留了潜变量的灵活性和多样性，且不会干扰这些约束。

这些特性共同使经BGMM增强的VAE成为CAUSALMIX的核心组件，能够在经验数据和用户指定的因果结构下生成高质量合成数据。

3.6 优化过程

训练通过最小化第3.3节所述的统一目标函数来进行，该目标函数结合了负证据下界以及用于约束用户指定因果结构的惩罚项。每个条件变分自编码器模块均使用随机梯度下降和Adam优化器（学习率 1×10^{-3}）独立优化。模型使用从80/20训练-验证划分中抽取的小批量进行训练。我们使用PyTorch Lightning管理批处理、设备分配、日志记录和早停。

前向传播与潜变量采样。 对于每个小批量，编码器生成潜变量参数（μ，$\log\ \sigma^2$），并使用标准重参数化技巧抽取潜变量样本 $z = \mu + \sigma \odot \epsilon$，$\epsilon \sim \mathcal{N}(0,\ I)$。 条件特征（对于$G_X$为 $T,\ \theta$，或对于G_Y为（$X,\ T$）$,\ \theta$）在解码前与潜在编码拼接。针对特征的输出头随后为二元变量和分类变量生成对数几率，并为连续变量生成高斯参数。

重构与KL项。 重构损失通过对各变量在相应似然族下的负对数似然求和来计算，而KL散度使用闭式高斯-高斯表达式。对于处理后生成器，重构以观测到的治疗组为条件，并且仅包含相应的潜在结果。

因果惩罚项。 在处理后生成器的训练过程中，解码器还会评估两个潜在结果 $Y'(0)$ 和 $Y'(1)$，从而计算诱导出的处理效应函数 $\tau_\theta(X)$ 和混杂函数 $\kappa_\theta(X,\ T)$。按照第3.3节所述应用均值对齐惩罚和方差惩罚，以满足相应的因果约束，并通过刚性参数控制它们的相对强度。同样，处理前生成器会计算每个样本的对数密度比 $\log\ p_\theta(X \mid T' = 1)\ - \log\ p_\theta(X \mid T' = 0)$，并对其施加惩罚，使其接近目标重叠函数 $\log \alpha(X)$。

早停与监控。 训练期间会监控验证损失，并在连续若干训练轮次未观察到改进时触发早停。PyTorch Lightning 会分别记录 VAE 损失、KL散度和各项因果惩罚项，从而能够诊断分布拟合和因果约束满足度。

训练后 BGMM 拟合。 收敛后，使用训练集的潜在均值 $\mu_\theta(\mathcal{O}_i)$ 拟合带有狄利克雷过程先验的贝叶斯高斯混合模型。最大分量数设置为等于潜在维度。该 BGMM 定义了生成过程中使用的灵活潜在先验，且不参与基于梯度的优化。

这些步骤共同产生的生成器能够逼近经验数据分布，同时忠实地实现指定的重叠、处理效应和混杂结构。

3.7 评估指标

我们从三个互补维度评估所提出的生成器：(i) 真实数据与合成数据之间的分布保真度，(ii) 因果结构保真度——包括处理效应、混杂和重叠行为，以及 (iii) 隐私保护。所有指标都将生成的合成数据集与观测数据进行比较。我们结合使用基于标准库的指标和自定义评估指标；指标定义与实现的摘要见 表 B.1。

Ma缘保真度。 我们分别针对连续型和离散评估一维边缘分布 e
vari变量。对于每个连续型特征 c，我们计算归一化瓦瑟斯坦距离

$$W_{\text{norm}}(c) = \frac{W_1(F^{(c)}_{\text{real}}, F^{(c)}_{\text{synth}})}{\text{sd}^{(c)}_{\text{real}} + \varepsilon},$$

其中 $F^{(c)}$ 真实数据 和 $F_{(c)}$ 合成数据 分别表示特征 c 在真实数据和合成数据中的经验分布， W_1 为 1-瓦瑟斯坦距离，$\text{sd}^{(c)}$ 真实数据 为特征 c 在真实数据中的标准差， $\varepsilon > 0$为用于保证数值稳定性的小常数。我们还报告 KS 补数，其定义为$1 - D_{\text{KS}}$，其中 D_{KS} 为柯尔莫哥洛夫–斯米尔诺夫距离。对于二元和类别型变量，我们报告基于总变差距离的 TVComplement，

$$\text{TVComplement}(c) = 1 - \frac{1}{2}\sum_{x} \left|p_{\text{real}}(x) - p_{\text{synth}}(x)\right|,$$

其中 $p_{\text{real}}(x)$ 和 $p_{\text{synth}}(x)$ 分别表示类别 x在真实数据和合成数据中的频数。聚合分数通过对各变量取平均值获得。

成对依赖性。 为评估低阶依赖结构，我们计算三种互补指标。对于连续型–连续型变量对，我们报告基于皮尔逊相关系数的平均相关性相似度分数。对于离散型–离散型变量对，我们报告基于总变差距离的平均列联相似度分数。此外，我们计算对称不确定性（SU），这是一种归一化互信息度量 [29]:

$$\text{SU}(i, j) = \frac{2\,\text{MI}(i, j)}{H(i) + H(j)},$$

其中 $\text{MI}(i, j)$ 表示变量 i 和 j之间的互信息， $H(i)$ 和 $H(j)$ 是它们的边缘熵。互信息和熵的估计方法是：对于连续变量 [30] ，使用 k 近邻法；对于离散变量，则使用经拉普拉斯平滑处理的列联表。我们报告 SU 相似度

$$1 - \left|\text{SU}_{\text{real}}(i, j) - \text{SU}_{\text{synth}}(i, j)\right|$$

对所有变量对取平均。

条件保真度。 我们通过评估$p(Z \mid C)$中的差异度来衡量条件结构的保留程度，其中 Z 表示除条件变量 C之外的所有特征。对于选定的条件变量（例如治疗分配），当 C 为连续型时，我们使用分位数分箱将真实数据和合成数据分别划分为由各层定义的层；当 C 为离散型时，则根据观测类别进行划分。在每个层内，我们使用混合核 [31], 计算真实样本和合成样本之间的最大均值差异（MMD²），其中连续变量使用RBF核，离散变量使用汉明核。最终的条件保真度分数是各层MMD²值按层大小加权后的

平均值。为便于理解差异度的大小，我们使用基于真实数据集分层拆分计算的真实数据内部参考值对该数值进行归一化；接近一的值表示其差异度与真实数据本身的内在变异程度相当。

联合保真度。 全局保真度通过以下方式评估：(i) 对编码表示 [32] 计算归一化能量距离；(ii) 进行分类器双样本检验（C2ST） [33]。C2ST 使用逻辑回归训练分类器，以区分真实数据与合成样本，并报告 ROC 曲线下面积以及重新缩放后的分数

$$1 - 2\left|\mathrm{AUC} - 0.5\right|,$$

当两个分布无法区分时，该分数达到最大值。

因果结构诊断。 由于生成器会显式建模潜在结果和用户指定的因果控制函数，我们从三个维度评估保真度：处理效应、未测量混杂和重叠。

处理效应诊断。 对于每个生成样本，处理后生成器会生成潜在结果 $Y'(0)$ 和 $Y'(1)$，从而得到相应的个体化效应

$$\tau_\theta(X') = Y'(1) - Y'(0).$$

给定目标函数 $\tau(X')$，我们计算平均绝对误差

$$\mathrm{MAE}_\tau = \mathbb{E}\big[\left|\tau_\theta(X') - \tau(X')\right|\big],$$

皮尔逊相关系数 $\rho(\tau_\theta(X'), \tau(X'))$ 用于评估异质性对齐程度；计算诱导的平均处理效应 $\mathrm{ATE}_\theta = \mathbb{E}[\tau_\theta(X')]$ 及其与目标 ATE 的偏差；并计算 $\tau_\theta(X')$ 和 $\tau(X')$ 经验分布之间的一维 Wasserstein 距离。 $\tau_\theta(X')$ 与 $\tau(X')$ 的散点图及叠加直方图可提供定性诊断。

未测量混杂诊断。 结合观测到的和翻转后的处理分配，生成器会诱导出混杂

$$\kappa_\theta(X', T') = \mathbb{E}[Y'(T') \mid X', T'] - \mathbb{E}[Y'(1 - T') \mid X', T'].$$

给定目标 $\kappa(X', T')$，我们计算

$$\mathrm{MAE}_\kappa = \mathbb{E}[|\kappa_\theta(X', T') - \kappa(X', T')|],$$

以及治疗组和对照组个体的组别特异性平均绝对误差，并计算 $\kappa_\theta(X', T')$ 与 $\kappa(X', T')$ 的分布之间的 Wasserstein 距离。散点图和分布叠加图直观地展示了混杂对齐情况。

重叠诊断。 我们采用两种互补方法评估重叠（正值性）。基于解码器的诊断会计算诱导的对数密度比

$$\Delta_\theta(X') = \log p_\theta(X' \mid T' = 0) - \log p_\theta(X' \mid T' = 1),$$

该比值与重叠控制函数仅符号约定不同。我们报告相对于目标对数 $\alpha(X')$的均方误差、 $\Delta_\theta(X')$的汇总统计量，以及落在用户指定容差带内的样本比例。此外，模型无关的基于倾向得分的诊断会拟合逻辑回归 $e(X') = \mathbb{P}(T' = 1 \mid X')$，通过高斯核密度估计来估计平滑的倾向评分密度，并计算重叠系数 [34]

$$\mathrm{OverlapCoeff} = \int_0^1 \min\left(p_{T'=0}(e),\, p_{T'=1}(e)\right) de.$$

评估还会生成一张信息丰富的图，将两条密度曲线叠加显示，并对它们的交集区域进行着色。

这些诊断指标共同量化生成数据对用户指定的因果约束的满足程度，这些约束编码于 $\tau(X)$、 $\kappa(X, T)$ 和 $\alpha(X)$ 中。

隐私。 隐私通过记录级披露风险进行操作性定义，并使用基于最近记录距离（DCR）的诊断指标进行量化，以评估合成样本是否与特定真实个体异常接近。 [35； 36]。将真实数据集和合成数据集编码到共同特征空间后，我们计算每条真实记录与最近合成近邻之间的距离

以及与最近其他真实近邻之间的距离。根据这些距离，我们报告三个密切相关的汇总指标。*DCR*保护比例是最近合成近邻比最近真实近邻更远的真实记录所占比例，数值越高表示保护越强。此外，我们还报告*DCR*距离比（合成数据与真实数据之间的距离除以真实数据之间的距离）的汇总统计量，用以量化记录级的相对接近程度。最后，我们报告标准化的DCRBaselineProtection评分，该指标将合成数据与真实数据之间的DCR中位数除以随机生成数据所对应的DCR中位数进行归一化；数值接近1表示隐私保护水平与随机噪声相当。详细信息见附录B。

3.8 条件平均处理效应估计方法

对条件平均处理效应的准确估计是个体化决策的核心，然而来自不同方法族的估计量依赖于各自不同的建模假设，并且在混杂、重叠不足和模型设定错误条件下的表现可能截然不同。因此，有意义的比较需要具备已知真实效应且因果违背可控的数据生成过程——而这些条件在观察性研究中很少同时满足。CAUSALMIX 提供了一个可控的生成式沙盒，其中治疗分配、结果机制、重叠和未测量混杂均可被系统性地改变，从而能够在多样的因果情形下对 CATE估计器进行有原则的压力测试。

我们评估来自四大方法族的代表性直接 CATE估计器。元学习器 [37],以 X 学习器为代表，使用针对治疗组和对照组个体的独立结果模型，通过治疗效应的加权平均来填补个体治疗效应，从而构建条件平均处理效应估计。双重稳健（*DR*）学习器 [38； 39]将结果回归与倾向评分建模相结合，形成伪结果；只要任一干扰模型设定正确，这些伪结果仍保持一致。双重机器学习（*DML*）方法 [40] 使用正交化估计方程和交叉拟合，以降低高维干扰项估计带来的偏倚。最后，基于树的方法包括因果森林 [41], ，其借助诚实样本拆分下的自适应局部平均来估计异质性效应，以及贝叶斯因果森林（BCF） [42], ，其分别对处理和预后效应进行建模，并通过贝叶斯收缩对异质性加以正则化。

这些模型族共同涵盖了广泛的建模策略和稳健性特性，因此非常适合在CAUSALMIX框架内进行系统评估。

4 案例研究：晚期前列腺癌中的异质性治疗效应

4.1 临床背景

转移性去势抵抗性前列腺癌（mCRPC）是前列腺癌的晚期阶段，其特征为对雄激素剥夺疗法产生耐药，并与疾病快速进展、预后不良以及总生存期有限相关 [43]。mCRPC最常用的一线治疗是新型雄激素受体靶向疗法阿比特龙和恩扎卢胺；由于缺乏头对头的III期随机试验，这两种疗法在常规临床实践中常被互换使用。两种疗法均与具有临床意义的不良事件相关，且新出现的观察性证据表明，其安全性特征可能因患者亚组而异。尤其是，已有报告指出，在患有既往心血管疾病或糖尿病的患者、老年人以及合并症负担较高的患者中，风险存在差异 [44]。

尽管已有这些观察结果，但比较治疗安全性中异质性的程度及其驱动因素仍未得到充分理解。更准确地刻画这种异质性，对于为转移性去势抵抗性前列腺癌患者的个体化治疗决策提供依据至关重要。因此，本案例研究考察了阿比特龙与恩扎卢胺作为转移性去势抵抗性前列腺癌一线治疗时比较安全性中的异质性治疗效应，重点关注不良事件风险。

4.2 数据描述

数据来源与队列整理。经验数据来自 OptumClinformatics Data Mart，这是一个纵向行政理赔数据库，包含商业保险和 Medicare Advantage 计划受益人的信息

。该数据库包含会员参保信息和人口统计信息、住院及门诊医疗索赔、药房索赔、实验室检查结果以及医疗服务提供者信息。本分析使用了 2014年1月1日至2024年3月31日期间的理赔数据。

采用 Freedland 等人提出并经验证的基于理赔数据的算法识别转移性去势抵抗性前列腺癌患者，该算法综合了诊断代码、前列腺特异性抗原检测结果、药物暴露，以及提示转移性疾病和去势抵抗的手术操作代码 [45]。该算法经过更新，纳入了近期获批的治疗方案。分析队列包括 4,098 名在转移性去势抵抗性前列腺癌索引日期后开始接受恩扎卢胺或阿比特龙一线治疗的患者。

变量。主要结局为治疗相关不良事件的发生，定义为治疗期间因任何临床相关不良事件而导致的住院或急诊就诊。治疗暴露根据转移性去势抵抗性前列腺癌索引日期后的首次处方确定：开始使用阿比特龙的患者归为处理组，而开始使用恩扎卢胺的患者则作为对照组。

基线协变量包括 18 个变量，涵盖二元、类别型和连续型变量。这些变量包括人口统计学特征（年龄和种族）、转移部位、既往前列腺癌治疗（包括紫杉烷类及其他全身治疗）、既往使用恩扎卢胺或阿比特龙的情况、以 Charlson 合并症指数衡量的合并症负担、心血管疾病和糖尿病病史、其他临床相关疾病，以及可能存在药物相互作用的合并用药。

4.3 模型训练与超参数选择

CAUSALMIX生成器采用上述变分架构进行训练，所选超参数旨在促进稳定优化、忠实再现经验分布，并在合成重复样本中稳健地控制因果目标。潜在维度设为与目标变量数量相等，而贝叶斯高斯混合先验中的混合分量数量则由模型自动推断。编码器和解码器均使用一个包含 64 个单元的隐藏层，训练时的批量大小设为 10。

控制处理效应与未测量混杂对齐的刚性参数固定为 10^3, ，而控制重叠的参数则设在 10^1 至 10^2, 之间，以反映它们对模型设定错误的敏感度差异。处理效应损失和混杂损失结合了均方误差与 SmoothL1惩罚项。当控制函数为低维或弱非线性时，降低均方误差权重（0.2–0.4），加大 Smooth L1 惩罚项的权重，并提高方差正则化，以避免合成效应塌缩或过于同质化。变量界限依据观测数据的经验范围设定。

我们通过验证损失稳定性、分布相似性的保持情况，以及受控因果量在独立生成的合成数据集之间的一致性，评估训练收敛性和稳健性。这些诊断指标为下文介绍的生成器验证设计提供了依据。

4.4 CAUSALMIX生成器验证与合成数据评估

我们从三个互补维度验证了CAUSALMIX生成器：分布保真度、因果结构保真度和记录级隐私。我们主要关注分布保真度和因果保真度，同时将隐私作为次要诊断指标，以确保提升真实感和因果控制不会导致过度记忆化。对于下文介绍的每种因果情景，我们还比较了从所提出的BGMM潜变量先验和标准高斯先验中采样的结果，以分离潜在多模态性对合成数据质量的影响。

因果场景设计。 我们考虑了三个难度逐步提升的因果情景，每种情景均由明确指定的处理效应函数 $\tau(X)$、未测量混杂函数 $\kappa(X,\ T)$ 和重叠函数 $\alpha(X)$ 定义。

情景*1*（同质效应、无混杂、完全重叠）。我们指定了恒定处理效应 $\tau(X) \equiv 0.1$、无未测量混杂 $\kappa(X,\ T) \equiv 0$，以及完全重叠 log $\alpha(X) \equiv$0（等价于 $\alpha(X) \equiv 1$）。此基线设置用于合理性检查，以验证 CAUSALMIX 在保留经验数据分布的同时，能够恢复简单的因果结构。

情景 *2*（线性异质效应、轻度混杂、恒定中度重叠）。我们引入了由具有临床相关性的协变量（既往心血管疾病、年龄和 Charlson 合并症指数）驱动的线性治疗效应异质性，

$$\tau(X) = 0.05 + 0.015\,\text{cvd} + 0.01\,\text{age} + 0.01\,\text{Charlson},$$

具有恒定的未测量混杂 $\kappa(X,\ T) \equiv 0.02$ 以及中等恒定重叠 $\log\ \alpha(X) \equiv 1$。该场景用于评估在轻微隐藏偏倚下恢复低维异质性的能力。

场景 *3*（非线性异质效应、协变量依赖型混杂和重叠）。这一最具挑战性的设置引入了非线性处理效应异质性、结构化未测量混杂以及协变量依赖型重叠。处理效应被设定为有界非线性函数，

$$\tau(X) = 0.02 + 0.05\ \tanh\big(0.4\,\text{cvd} + 0.2\,\text{age} + 0.2\,\text{Charlson} + 0.4\,\text{dementia}\big),$$

从而产生具有饱和效应的异质性效应。未测量混杂同时取决于协变量和处理分配，

$$\kappa(X,T) = 0.05 - 0.01\,(2T-1)\ \tanh\big(0.5\,\text{Charlson} + 0.6\,\text{cvd} + 0.2\,\text{age}\big),$$

导致处理组之间出现不对称的隐藏偏倚。重叠明确依赖于处理，并由既往阿比特龙暴露所驱动，

$$\log \alpha(X) = 2\big(2\,\text{Abiraterone_prev} - 1\big),$$

从而形成与具有临床意义的协变量相对应的强重叠和弱重叠区域。该场景考验 CAUSALMIX 联合控制非线性处理效应异质性、选择偏倚和治疗依赖型重叠的能力。

因果保真度的可视化。除定量因果保真度指标外，还辅以诊断性可视化。通过比较目标 CATE 分布与学习得到的 CATE 分布，并结合散点图和联合图来揭示各个体之间的一致性和离散程度，从而评估处理效应保真度。未测量混杂则通过类似的比较进行评估。通过倾向评分分布展示重叠情况，以说明共同支撑的保留情况。

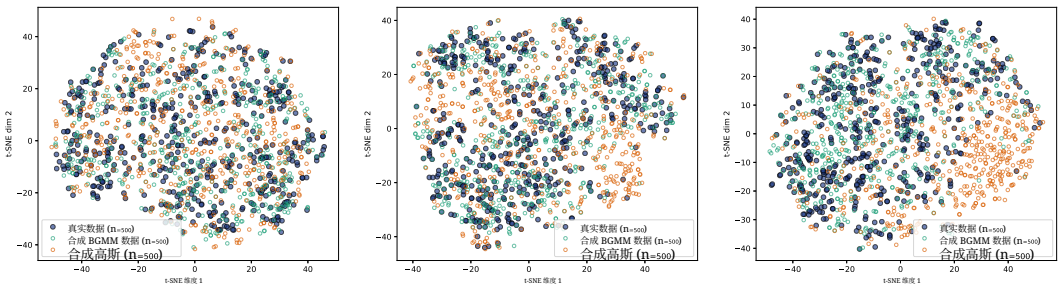
分布保真度的可视化。我们通过多个依赖尺度上的可视化诊断评估分布真实性。对于连续变量，通过叠加密度图考察边际保真度；对于离散变量，则使用条形图。对于连续型变量对，使用散点图或密度图评估成对保真度；对于离散型变量对，使用列联表热图；对于离散型与连续型变量对，则使用离散化的条件分布。最后，通过真实样本和合成样本的二维联合嵌入来可视化联合保真度，从而全面呈现整体对齐情况，并揭示边际或成对汇总未能捕捉到的差异。

结果。在所有场景中，我们从三个维度评估合成数据质量：分布保真度（表 D.1–D.3）、因果保真度（表 D.4–D.6）和隐私（表 D.7–D.9）。诊断图（图 3 和图 D.1 至 D.6）表明，在两种潜在先验下，CAUSALMIX 在所有维度上均实现了较高的分布保真度和因果保真度；随着因果复杂性增加，性能差异才开始显现。

如图 2 的二维联合嵌入所示，随着因果复杂度增加，基于 BGMM 的采样与真实数据保持了更紧密的对齐。在结合了多模态协变量分布、非线性异质性、协变量依赖型混杂和重叠的场景 3 中，BGMM 先验在边缘分布、成对、条件和联合分布指标上的表现显著优于高斯先验（表 D.3）。具体而言，BGMM 对连续边际分布的归一化 Wasserstein 距离更低，对离散边际分布和成对关系的相似性更高，条件对齐效果更好，并且在归一化能量距离和分类器两样本检验（C2ST）的衡量下，联合保真度显著更高。

两种先验都以很高的准确率恢复了设定的因果结构。然而，BGMM 在具有挑战性的情形下表现出更强的稳健性：在场景 3 中，它取得了更高的条件平均处理效应相关性，并且解码器级重叠重建的准确性显著更高，同时保持了相当的平均处理效应误差和混杂恢复能力（表 D.6）。

隐私诊断指标显示，真实性与隐私之间的权衡得到了有效控制。高斯先验始终能提供更强的记录级保护，其保护比例更高，距离比也更大。



(a) 情景1　　(b) 场景 2　　(c) 场景 3

图 2：在不同因果情景下，采用 BGMM 先验和高斯先验生成的合成数据的二维联合嵌入，并与真实 mCRPC 数据集进行比较。情景1：同质效应、无混杂、完全重叠；场景 2：线性异质效应、轻度混杂、恒定重叠；场景 3：非线性异质性，同时存在协变量依赖型混杂和重叠

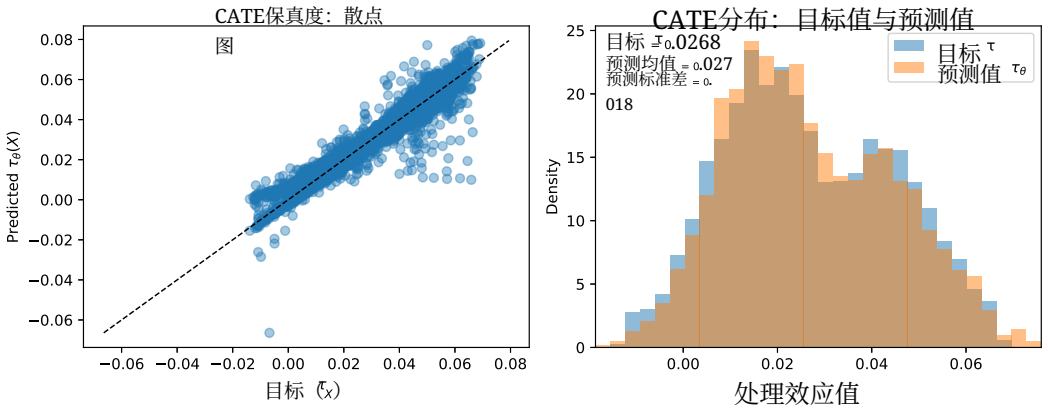


图 3：在场景 3 中，比较 CausalMix 使用 BGMM 先验生成的 mCRPC 数据集治疗效应与目标治疗效应。

BGMM先验的保护效果略低，尤其是在场景 3 中；但其保护比例仍远高于 0.5，中位距离比也超过 1，表明不存在系统性的记忆化。BGMM先验导致的保护效果下降主要集中在距离比分布的低分位数，反映的是局部邻近，而非普遍的披露风险（表 D.7–D.9）。

总体而言，这些结果表明，CAUSALMIX 使用 BGMM潜变量先验，能够提升复杂临床场景下的分布保真度和因果保真度，同时保持稳定的记录级隐私，展现了分布表达能力与披露风险之间可控的权衡。

4.5 应用1：CATE估计器的基准评测

从观察性数据中估计条件平均处理效应，对底层数据生成过程高度敏感，包括混杂程度、倾向得分重叠、结局变异性、函数形式（线性与非线性），以及处理效应的大小和异质性 [3]。我们将 CAUSALMIX 用作受控沙盒，在经过校准、能够匹配前列腺癌队列关键经验特征的数据生成过程中，对 CATE估计器进行基准评测。我们的目标是找出既准确又具备计算可行性的估计器，以检测比较治疗安全性中的异质性。

设计与目标估计量。我们聚焦于 CAUSALMIX 中第二个经过验证的因果情景，该情景具有中度重叠（$\log \alpha(X) = 1$）和线性处理效应异质性，

$$\tau(X) = 0.05 + 0.015 \cdot \texttt{cvd} + 0.01 \cdot \texttt{age} + 0.01 \cdot \texttt{Charlson}.$$

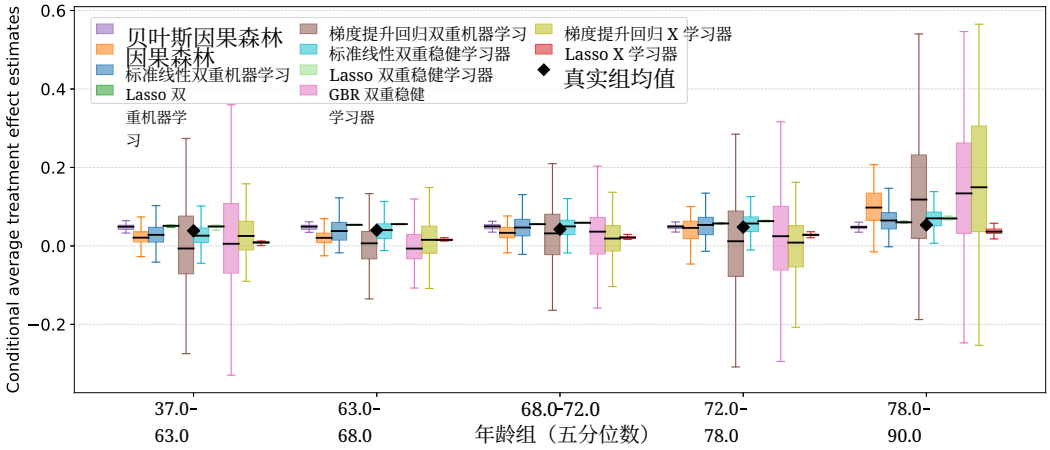
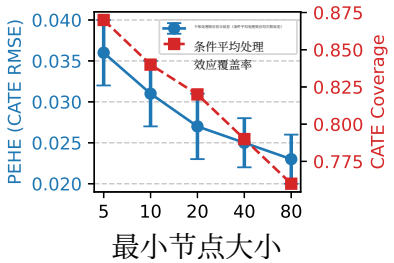


图 4：按年龄五分位组划分的分组条件平均处理效应估计值。对于每个年龄组，箱线图汇总了在校准的合成数据生成过程中（不存在未测量混杂，$\kappa = 0$）且查尔森 = 0、无既往心血管疾病的亚组中，各估计量得出的个体层面条件平均处理效应估计值。黑色菱形表示组层面的真实平均条件处理效应。



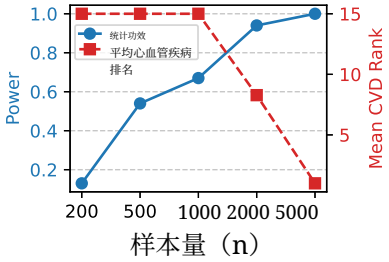


图 5：PEHE 与最小叶节点大小的关系。图 6：基于包含 2,000 棵树的因果森林计算的统计功效和平均心血管疾病排名。误差线表示跨重复试验的 ±1 标准差。因果森林计算的变量重要性随样本量增加的变化。

主要估计量为 ATE $= \mathbb{E}[\tau(X)]$ 和 CATE $\tau(X)$。我们考虑了两种设置：（i）无未测量混杂（$\kappa(X,\ T) = 0$）；（ii）少量未测量混杂（$\kappa(X,\ T) = 0.02$）。对于每种设置，我们生成了 50 个独立合成数据集，并保留真实的平均处理效应和个体处理效应用于评估。

估计量。我们比较了十种条件平均处理效应估计量，涵盖元学习器、基于正交化的方法和贝叶斯正则化方法。具体而言，我们考察了：

- 采用梯度提升回归（GBR）和 Lasso 模型进行最终阶段条件平均处理效应估计的 X 学习器；
- 最终阶段条件平均处理效应模型采用 {GBR、Lasso、线性}的双重机器学习估计量；
- 采用最终阶段条件平均处理效应模型的双重稳健学习器，包括 {梯度提升回归、Lasso、线性模型}；
- 基于树的方法，包括因果森林和贝叶斯因果森林。

对于所有非树估计量，均使用梯度提升树估计干扰函数（倾向得分和结果回归），以确保不同方法之间具有可比性。所有分析均采用交叉拟合。

预处理。由于线性条件平均处理效应模型在采用完整独热编码时可能出现多重共线性，进而导致协方差估计不稳定、渐近推断不可靠，因此我们为每个数据集构建了两种特征表示：X 学习器、DML 和 DR 变体采用**虚拟变量编码**，基于树的估计量（因果森林和贝叶斯因果森林）采用**完整独热编码**。连续协变量经过标准化，二元协变量则保持不变。每个数据集只进行一次预处理，并在各估计量之间重复使用。

不确定性量化与重复试验。 线性最终阶段 CATE 模型采用渐近（解析）推断；Lasso 和非参数（梯度提升回归）最终阶段模型则采用包含 100 次重抽样的自助法推断。每个场景均进行了 50 次独立重复实验。为比较计算效率，各估计量均限制为单线程，并记录实际运行时间。

评估指标。 令 $\hat{\tau}(X_i)$ 表示估计的条件平均处理效应，$\tau(X_i)$ 表示真实个体处理效应。性能首先在每个合成数据集内计算，然后在各重复试验间汇总。

对于平均处理效应估计，我们报告了各重复试验间平均处理效应偏差的均方根误差（RMSE）、平均处理效应偏差的均值和标准差、平均估计标准误，以及平均处理效应置信区间的经验覆盖概率。

对于条件平均处理效应估计，我们报告了均方根误差——亦称异质性效应估计精度（个体处理效应估计误差） [46],

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) - \tau(X_i)\right)^2},$$

在每次重复试验内计算并在各重复试验间取平均，同时给出个体处理效应估计误差的标准差以及逐点条件平均处理效应区间的平均覆盖概率。
最后，我们报告了实际运行时间的均值和标准差，以量化计算效率。

Res果。 两种混杂情境具有相同的、较小但非零的真实平均处理效应 0.056（标准差：0.000），从而能够 g dire不同方法间估计量准确率和不确定性校准的比较。

无未测量混杂。 对于平均处理效应点估计，大多数估计量的偏差都较低（图 D.7）。贝叶斯因果森林、因果森林和双重机器学习／双重稳健学习器的平均处理效应均方根误差约为 0.009–0.010，而 X 学习器变体表现出明显更大的误差（平均处理效应均方根误差 ≈ 0.024）。不同方法的平均处理效应不确定性校准存在差异：一些估计量达到了接近名义水平的覆盖率（例如贝叶斯因果森林、因果森林、线性双重机器学习），而另一些估计量主要通过夸大标准误实现了近乎完美的覆盖率（尤其是基于梯度提升回归的双重机器学习／双重稳健学习器和线性双重稳健学习器；表 D.10）。鉴于蒙特卡洛重复试验次数有限，对于名义覆盖率的轻微偏离应谨慎解读。

与平均处理效应估计相比，各估计量在条件平均处理效应推断方面的差异更加明显。贝叶斯因果森林和标准线性双重稳健学习器实现了最高的平均 CATE 覆盖率（0.91–0.92），因果森林和基于梯度提升回归的学习器的校准程度中等（0.81–0.87），而基于套索回归的双重机器学习／双重稳健学习器的条件平均处理效应覆盖率较差（0.37–0.40）。这一模式表明，尽管某些方法的平均处理效应均方根误差（表 D.10）和个体处理效应估计误差（图 D.8）具有竞争力，但其异质性效应的不确定性量化并不可靠。图 4 按年龄五分位组总结了 CVD $= 0$，查尔森 $= 0$ 亚组中的条件平均处理效应估计。贝叶斯因果森林的估计分布十分集中，并且在各五分位组中都跟随真实组均值；相比之下，因果森林表现出更高的变异性，尤其是在年龄最大的组别中。基于套索回归的估计量过度平滑并削弱了细尺度异质性，这与其较低的覆盖率一致。相比之下，梯度提升变体显示出明显更宽的分布和更厚的尾部，这一现象在年龄较大的群体中最为突出。

运行时间对比凸显了实际应用中的权衡。线性双重机器学习／双重稳健学习器的运行时间约为 2 秒，因果森林约为 25 秒，而贝叶斯因果森林以及提升式或基于套索回归的元学习器通常需要 100–270 秒（附录中的表 D.10）。
轻微未测量混杂。引入未测量混杂后，所有估计量的平均处理效应均出现系统性偏差（图 D.7），平均偏移约为 0.02，与注入的混杂程度相符。标准误较小的方法，其平均处理效应覆盖率显著下降。条件平均处理效应覆盖率的模式大体保持不变：贝叶斯因果森林的校准程度仍相对较好，而基于套索回归的双重机器学习／双重稳健学习器仍存在显著的覆盖不足（附录中的表 D.11）。

要点。 由于前列腺癌应用关注不同患者亚组之间的异质性安全性效应，选择估计量时应同时优先考虑稳定的条件平均处理效应点估计和校准的不确定性。贝叶斯因果森林是一个有力的首选方案，能够兼顾对异质性效应的准确恢复和较高的条件平均处理效应覆盖率。双重稳健和基于双重机器学习的方法，尤其是标准线性

双重稳健学习器可提供快速且校准程度较为合理的基线，适用于敏感性分析。基于 Lasso 的元学习器能够捕捉总体条件平均处理效应趋势，但由于条件平均处理效应覆盖率较低，在推断中应将其视为探索性方法。最后，由于未测量混杂会使所有估计量产生系统性偏差，因此在真实世界的前列腺癌情境中，仅凭估计量选择不足以确保有效性，还应结合残余混杂调整和敏感性分析。

4.6 应用 2：超参数优化

除了方法比较之外，CAUSALMIX 还通过在经过经验校准的数据生成过程中，针对已知的因果真实值进行直接评估，为因果估计量的超参数优化提供了一个有原则的沙盒。在本应用中，我们使用 CAUSALMIX 调整因果森林的超参数，以实现对总体层面和亚组特异性处理效应估计均具有决策相关性的准确率。合成数据集根据与第 4.5 节相同的校准机制生成，并假设存在中度重叠、线性处理效应异质性以及无未测量混杂。

目标。 我们将应用 1 中经过校准的 CAUSALMIX生成器 $\mathcal{G}$（不含残余混杂）视为替代真实值环境，并调整因果森林超参数，以优化因果性能。鉴于重点在于个性化决策和亚组决策，我们将注意力限制在控制因果森林估计平滑性与稳定性的两个关键超参数上：最小叶节点大小（min.node.size）和树的数量（num.trees）。我们使用互补的估计准确率和不确定性量化指标对超参数配置进行评估。

实验设计。 我们考虑了由以下内容定义的超参数网格：

$$\texttt{min.node.size} \in \{5, 10, 20, 40, 80\},$$
$$\texttt{num.trees} \in \{500, 1000, 2000, 4000\}.$$

对于每种配置，我们进行了50次独立重复实验。在每次重复试验中，均从CAUSALMIX生成器 $\mathcal{G}$中抽取一个合成数据集$\hat{}$，该生成器提供观测数据以及相应的真实平均处理效应和条件平均处理效应。随后，使用指定的超参数拟合因果森林，并利用袋外预测获得个体层面的条件平均处理效应估计。

我们在重复试验层面汇总性能指标，包括平均处理效应均方根误差、条件平均处理效应均方根误差（个体处理效应估计误差）、条件平均处理效应平均绝对偏差、条件平均处理效应置信区间平均宽度，以及平均处理效应和条件平均处理效应的区间覆盖率，然后对各次重复试验的结果取平均。为刻画超参数敏感性，我们共同考察了平均处理效应准确度和校准，以及条件平均处理效应点估计准确度和不确定性大小，并评估这些指标在不同配置下的变化情况。鉴于研究目标侧重于亚组分析，超参数选择优先考虑最小化个体处理效应估计误差，同时维持足够的条件平均处理效应覆盖率；平均处理效应均方根误差和覆盖率则作为总体层面的诊断指标。

结果。 在所有配置中，平均处理效应估计基本不受超参数选择的影响：平均处理效应均方根误差保持在 0.009，平均处理效应覆盖率则维持在较窄范围内（0.92–0.94）。相比之下，条件平均处理效应的表现主要受最小叶节点大小影响。增大最小叶节点大小可提高条件平均处理效应的点估计准确度，个体处理效应估计误差从约 0.036–0.037（叶节点大小为 5）降至约 0.023（叶节点大小为 80）；但这一改善是以降低条件平均处理效应区间覆盖率为代价的，覆盖率从约 0.87–0.91 降至约 0.75–0.76（见图 5）。与这一趋势一致，随着叶节点大小增加，条件平均处理效应的平均绝对偏差和平均置信区间宽度均稳步下降（图 D.9），这表明在这一合成场景中，较小的叶节点会导致过拟合和过大的方差。树的数量达到适中规模后，对平均处理效应和条件平均处理效应的准确度或覆盖率影响微乎其微，但却是运行时间的主要决定因素；运行时间大致呈线性增长（500 棵树约需 7 秒，1,000 棵约需 11 秒，2,000 棵约需 21 秒，4,000 棵约需 42 秒）。

基于这一权衡，我们选择了包含 1,000 或 2,000 棵树、最小叶节点大小为 20 的因果森林配置。该配置在保持较强平均处理效应性能（平均处理效应均方根误差 0.009；平均处理效应覆盖率 0.92）的同时，兼顾了条件平均处理效应准确度与校准（个体处理效应估计误差 0.027；平均 CATE 覆盖率 0.818），且运行时间适中（约 11 秒）。

要点。 这些结果表明，在使用因果森林进行异质性安全效应估计时，调整最小叶节点大小 比大幅增加

树的数量重要得多。较大的叶节点会产生更平滑的条件平均处理效应估计和更低的个体处理效应估计误差，但也可能导致区间过度自信、覆盖率降低。对于转移性去势抵抗性前列腺癌应用，一种实用的选择是使用约 1,000–2,000 棵树以确保稳定性，并通过敏感性分析调整最小叶节点大小（例如 10–40），优先选择既能保留临床上合理的异质性，又能避免报告亚组风险差异时不确定性覆盖不足的设置。

4.7 应用 3：前瞻性研究设计的功效分析

在亚组分析中检测异质性治疗效应往往受到样本量限制，尤其是在效应修饰较弱且协变量彼此相关时 [47]。因此，现实可行的研究规划不仅需要对效应量作出假设，还需要考虑协变量的联合分布、治疗分配和结局变异性。通过根据经验性前列腺癌队列校准 CAUSALMIX，我们生成了保留这些关键分布特征的合成数据集，从而能够开展有依据且具有实际意义的功效分析。

研究目标。在本应用中，我们探讨了一个前瞻性研究设计问题：在比较治疗安全性时，检测效应修饰所需的最小样本量是多少，才能确保可靠检出？具体而言，我们的目标是检测与预先指定的二元效应修饰因子相关的非零亚组条件平均处理效应差异，

$$\Delta := \mathbb{E}[\tau(X) \mid \texttt{CVD} = 1] - \mathbb{E}[\tau(X) \mid \texttt{CVD} = 0]\,,$$

其中，CVD 表示基线心血管疾病史。为便于解释，我们假设本应用中不存在未测量混杂。

异质性设定。除非另有说明，我们采用与应用 2 相同的校准数据生成机制，包括中度重叠和线性处理效应异质性。根据研究假设的需要，可通过缩放中心化条件平均处理效应分量 $h(x) = \tau(x) - \mathbb{E}[\tau(X)]$ via $\tau_\eta(x) = \mathbb{E}[\tau(X)] + \eta\, h(x)$。参数 η 重新缩放个体相对于平均处理效应的偏差，使处理效应异质性能够系统变化，同时保持平均处理效应不变。我们报告默认设定 $\eta = 1$ 下的结果。

通过最佳线性投影检测效应修饰因子。 我们使用 grf 中实现的最佳线性投影（BLP）程序评估效应修饰。对于每个已拟合的因果森林，我们利用双重稳健评分，将估计的条件平均处理效应对基线协变量进行回归，$\hat{\tau}(X_i) \approx \beta_0 + X_{i,S}^{\top}\beta$，从而得到系数估计，以及采用异方差稳健（HC3）标准误的 Wald 检验 p 值。推断聚焦于与预先指定的效应修饰因子心血管疾病相对应的系数。对不存在效应修饰这一原假设的拒绝，基于显著性水平为 $\alpha = 0.05$ 的双侧检验。

实验设计与功效估计。 我们考虑了候选样本量 $\mathcal{N} = \{200, 500, 1000, 2000, 5000\}$。对于每个 $n \in \mathcal{N}$，我们从校准后的CAUSALMIX生成器生成了 $R = 100$ 个独立合成数据集。在每次重复试验中，我们使用应用 2 中选定的超参数拟合因果森林，并应用 BLP 程序检验心血管疾病所产生的效应修饰。若心血管疾病对应的 BLP p 值低于 α，则该重复试验层面的拒绝指示变量取值为 1。

样本量为 n 时的统计功效估计为各次重复试验中的经验拒绝频率。我们报告由此得到的功效曲线，并确定达到 80% 目标功效水平所需的最小样本量 $n^\star$。

除假设检验外，我们还记录了每次重复试验中因果森林的变量重要性得分，以及预先指定的效应修饰因子（CVD）对应的排名。这些诊断指标提供了补充信息，说明随着样本量增加，因果森林识别 CVD 为处理效应异质性驱动因素的可靠程度。

结果。 为了发现效应修饰因子，我们拟合了因果森林，使用 2,000 棵树和 80 的最小节点大小，以减少条件平均处理效应噪声并稳定异质性信号，但不以校准的不确定性为目标。图 6 总结了经验功效（检出概率）和 CVD 的平均变量重要性排名随样本量变化的情况。

统计功效随 n急剧上升：在 $n = 200$ 时较低（0.13），在 $n = 500$（0.54）和 $n = 1000$ （0.67）时升至中等水平，到 $n = 2000$ （0.94）时达到较高水平，并在$n = 5000$时接近1.0。变量重要性诊断呈现出一致的模式。当 $n \leq 1000$时，CVD未能被可靠地排在优先位置（平均排名15）；而在 $n = 2000$ 时，其平均排名提升至8；到 $n = 5000$ 时，则始终被识别为排名第一的修饰因子（平均排名1）。

要点。 在此处考察的协变量结构和异质性强度下，使用因果森林可靠地检测效应修饰，需要约 ∼2,000名患者，才能达到至少90%的统计功效。要稳定地优先识别关键修饰因子（例如将CVD列为排名靠前的变量），则需要更大的样本量，接近 $n = 5{,}000$名患者。对于前列腺癌安全性研究，这些结果提示，应将来自较小队列的修饰因子发现视为探索性结果；若目标是稳健地发现亚组，则应辅以敏感性分析或汇总数据策略。

5 讨论与结论

总结。 本研究推动了合成数据生成的发展，将其构建为一个因果推断沙盒，同时以经验真实性和明确的因果控制为目标。我们提出了 CAUSALMIX，这是一种灵活的生成框架，提供可解释的设计阶段控制参数，用于调节处理效应、未测量混杂和重叠；并展示了混合先验如何提升复杂多模态表格数据的真实性。除了生成器本身，我们还提出了一个统一的评估套件，涵盖分布保真度、因果保真度和隐私诊断，并展示了经过验证的合成数据如何指导估计量基准测试、超参数调优和功效分析。

方法论启示。 我们的结果凸显了一条核心结论：评估和调优因果合成数据时，应同时考量分布真实性与因果结构，并明确审查隐私风险。在具有挑战性的情境中，基于BGMM的潜变量抽样能够提升整体和局部的分布对齐，同时保持因果保真度。该评估套件提供了模块化的诊断视角——分布指标用于识别边缘分布和依赖结构中的不匹配，因果保真度指标用于评估目标因果量的实现情况，隐私指标则用于量化记录级风险——使 CAUSALMIX 成为因果机器学习开发和研究规划的实用测试平台。在实践中，该框架只需有限的特定任务调优：架构选择保持固定或自动推断，而目标数据生成机制仅由少量可解释的因果控制参数定义。

案例研究带来的启示。 转移性去势抵抗性前列腺癌案例研究说明，受控合成数据如何将生成器验证转化为可操作的方法学指导。在真实值已知的情况下，异质性效应估计量在三个实践中常被混为一谈的维度上各有差异：平均处理效应准确度、不确定性校准和计算成本。值得注意的是，平均处理效应准确度高，并不意味着条件平均处理效应推断可靠——一些估计量虽然给出了看似合理的点估计，却对条件平均处理效应存在严重的区间覆盖不足，因此不适合用于指导亚组结论。引入未测量混杂会使所有方法都产生系统性偏差，这进一步说明，单靠估计量选择并不能确保结果有效，还应辅以未测量混杂诊断指标（例如负对照方法）和敏感性分析。

超参数分析和功效分析分别针对异质性处理效应分析的不同方面。超参数调优侧重于估计质量：对于因果森林，最小叶节点大小决定了条件平均处理效应估计中的偏差-方差权衡；增大叶节点会使效应估计更加平滑、稳定，但会牺牲局部适应性（对更精细异质性的刻画能力）。树的数量达到足够规模后，主要影响计算稳定性。相比之下，功效分析着眼于效应修饰因子识别和研究设计。我们的结果显示，检测处理效应异质性的存在所需的样本量，远少于可靠识别并确定特定修饰因子优先级所需的样本量。中等样本量或许足以拒绝同质性假设，但要稳健地发现亚组，则需要更大的队列。这凸显了准确估计异质性效应与切实识别效应修饰因子之间的区别。

这些组件共同使 CAUSALMIX 既适合作为因果机器学习开发的测试平台，也适合作为观察性研究的决策支持工具，尤其适用于估计量选择和研究设计取决于异质性效应的情形。

局限性与未来工作。 与任何学习得到的生成器一样，CAUSALMIX 在数据支持稀疏的协变量空间区域中，会依据其神经网络架构的结构假设进行外推；因此，在总体分布指标和因果保真度指标上表现出色，并不意味着它能忠实地再现罕见事件、极端协变量组合，或经验数据中体现不足的高阶依赖关系。

此外，因果控制依赖用户指定的函数形式。尽管低维、易于解释的设定有助于提升稳定性和透明性，但当真实的处理效应异质性具有高维特征时，这些设定可能无法捕捉复杂的交互结构。这一局限反映了分布真实性与因果可控性之间更广泛的矛盾：并非所有因果目标都能同时得到高保真度的实现，尤其是在它们与数据的经验结构相冲突时。因此，如何制定规范的方法来指定、验证并对因果函数进行压力测试，仍是未来工作的重要方向。

一个实际局限涉及训练数据需求。作为一种灵活的神经生成器，CAUSALMIX 需要足够大的样本量，才能可靠地捕捉多模态协变量结构和异质性治疗效应，同时稳定因果控制约束。在较小的数据集中，可能会出现过拟合或不稳定，尤其是在维度和效应复杂度增加时。系统地刻画稳定因果合成数据生成所需的数据规模阈值，仍是未来研究的一个开放方向。

最后，CAUSALMIX 目前面向协变量信息完整的静态处理设定。将该框架扩展至动态处理方案和显式的缺失数据机制，将有助于提升其对纵向医疗数据的真实刻画能力，并使研究者能够在更现实的数据可用性约束下对因果估计量进行压力测试。

参考资料

[1] Paul W Holland. 统计学与因果推断。美国统计学会会刊，81(396):945–960，1986。

[2] P Richard Hahn, Vincent Dorie, and Jared S Murray. 2017年大西洋因果推断会议（ACIC）数据分析挑战。*arXiv* 预印本 *arXiv:1905.09515*，2019年。[3] Divyat Mahajan, Ioannis Mitliagkas, Brady Neal, and Vasilis Syrgkanis. 异质性因果效应估计模型选择的实证分析。*arXiv* 预印本 *arXiv:2211.01939*，2022。[4] Vasileios C Pezoulas, Dimitrios I Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, and Dimitrios I Fotiadis. 医疗保健中的合成数据生成方法：开源工具与方法综述。《计算与结构生物技术杂志》，23:2892–2910，2024年。

[5] Sarah Friedrich and Tim Friede. 基准数据集和模拟在方法比较研究中的作用。《生物计量学杂志》，66(1):2200212，2024年。[6] Juan L Gamella. 人工智能方法学的物理测试平台。博士论文，苏黎世联邦理工学院，2025年。[7] Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. REALCAUSE：真实的因果推断基准测试。*arXiv* 预印本 *arXiv:2011.15007*，2020年。[8] Susan Athey, Guido W Imbens, Jonas Metzger, and Evan Munro. 使用 Wasserstein 生成对抗网络设计蒙特卡洛模拟。《计量经济学杂志》，240(2):105076，2024年。[9] Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. 验证因果推断方法。载于国际机器学习会议，第17346–17358页。PMLR，2022年。

[10] PaulR Rosenbaum和Donald B Rubin。倾向评分在因果效应观察性研究中的核心作用。*Biometrika*，70(1):41–55，1983年。

[11] Yishai Shimoni、Chen Yanover、Ehud Karavani和Yaara Goldschmnidt。用于因果推断分析性能评估的基准评测框架。*arXiv*预印本 *arXiv:1802.05046*，2018年。

[12] Vincent Dorie、Jennifer Hill、Uri Shalit、Marc Scott 和 Dan Cervone。自动化方法与 DIY 方法在因果推断中的比较：从一场数据分析竞赛中汲取的经验。*Statistical Science*，34(1):43–68，2019年。

[13] Alicia Curth 和 Mihaela van der Schaar。关于异质性治疗效应估计的归纳偏置。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，*神经信息处理系统进展*，第 34 卷，第 15883–15894 页。Curran Associates, Inc.，2021年。[14] Alberto Abadie 和 Guido W Imbens。平均处理效应的偏差校正匹配估计量。《商业与经济统计学杂志》，29(1):1–11，2011年。[15] Jessica M Franklin、Sebastian Schneeweiss、Jennifer M Polinski 和 Jeremy A Rassen。用于评估复杂医疗保健数据库中药物流行病学方法的 plasmode 模拟。《计算统计与数据分析》，72:219–226，2014年。[16] Shao-Hsien Liu、Stavroula A Chrysanthopoulou、Qiuzhi Chang、Jacob N Hunnicutt 和 Kate L Lapane。边际结构模型中的缺失数据：比较多重插补与逆概率加权的 plasmode 模拟研究。《医疗保健》，57(3):237–243，2019年。[17] Vanessa Ress 和 Eva-Maria Wild。比较行政健康数据中因果治疗效应的估计方法：一项 plasmode 模拟研究。《健康经济学》，33(12):2757–2777，2024年。[18] Raymond B Cattell 和 Joseph Jaspers。用于因子分析练习与研究的一般 plasmode（第 30-10-5-2 号）。《多变量行为研究专论》，1967年。[19] Gary L Gadbury、Qinfang Xiang、Lin Yang、Stephen Barnes、Grier P Page 和 David B Allison。在公共数据库规模庞大的时代，使用 plasmode 数据集评估统计方法：以假发现率为例。《*PLoS*遗传学》，4(6):e1000098，2008年。[20] Laura K Vaughan、Jasmin Divers、Miguel A Padilla、David T Redden、Hemant K Tiwari、Daniel Pomp 和 David B Allison。将 plasmode 数据用作模拟的补充：评估个体混合比例估计方法的一个简单示例。《计算统计与数据分析》，53(5):1755–1766，2009年。[21] Pamela A Shaw、Susan Gruber、Brian D Williamson、Rishi Desai、Susan M Shortreed、Chloe Krakauer、Jennifer C Nelson 和 Mark J van der Laan。关于因果推断背景下 plasmode 模拟研究的警示。*arXiv* 预印本 *arXiv:2504.11740*，2025年。[22] Sandrine Boulet 和 Antoine Chambaz。为我绘制一个模拟器：利用神经网络构建更逼真的因果分析模拟方案。*Computo*，2025年。[23] Martin Arjovsky、Soumith Chintala 和 Léon Bottou。Wasserstein生成对抗网络。载于国际机器学习会议，第 214–223 页。PMLR，2017年。[24] Linying Yang、Robin J Evans 和 Xinwei Shen。节俭、灵活、忠实：通过 Frengression 模拟因果数据。*arXiv* 预印本 *arXiv:2508.01018*，2025年。[25] Anna Guo、David Benkeser 和 Razieh Nabi。前门模型下因果效应的灵活非参数推断。*arXiv* 预印本 *arXiv:2312.10234*，2023年。[26] Anna Guo 和 Razieh Nabi。含隐藏变量的 DAG 中的平均因果效应估计：超越后门和前门准则。*arXiv* 预印本 *arXiv:2409.03962*，2024年。[27] Anna Guo、David Benkeser 和 Razieh Nabi。使用“餐巾纸图”进行因果推断。*arXiv* 预印本 *arXiv:2512.19861*，2025年。[28] Patricia A Apellániz、Juan Parras 和 Santiago Zazo。一种集成 VAE-GMM 的改进型表格数据生成器。载于*2024*年欧洲信号处理会议（*EUSIPCO*）第 *32* 届会议，第 1886–1890 页。IEEE，2024年。[29] Lei Yu 和 Huan Liu。高维数据的特征选择：一种快速的基于相关性的筛选器方法。载于第 *20* 届国际机器学习会议论文集（*ICML-03*），第 856–863 页，2003年。[30] Alexander Kraskov、Harald Stögbauer 和 Peter Grassberger。互信息估计。《物理评论*E*：统计、非线性与软物质物理》，69(6):066138，2004年。

[31] Arthur Gretton、Karsten M Borgwardt、Malte J Rasch、Bernhard Schölkopf 和 Alexander Smola。核两样本检验。机器学习研究杂志，13(1):723–773，2012年。[32] Gábor J Székely 和 Maria L Rizzo。能量统计：一类基于距离的统计量。统计规划与推断杂志，143(8):1249–1272，2013年。[33] David Lopez-Paz 和 Maxime Oquab。重新审视分类器两样本检验。*arXiv* 预印本 *arXiv:1610.06545*，2016年。[34] Henry F Inman 和 Edwin L Bradley Jr. 重叠系数：作为概率分布间一致性的度量，以及正态密度重叠部分的点估计。《统计通讯：理论与方法》，18(10):3851–3874，1989年。[35] Theresa Stadler、Bristena Oprisanu 和 Carmela Troncoso。合成数据——匿名化的土拨鼠日。载于第*31*届 *USENIX* 安全研讨会（*USENIX Security 22*），第1451–1468页，2022年。[36] Amy Steier、Lipika Ramaswamy、Andre Manoel 和 Alexa Haushalter。合成数据隐私度量指标。*arXiv* 预印本 *arXiv:2501.03941*，2025年。[37] Sören R Künzel、Jasjeet S Sekhon、Peter J Bickel 和 Bin Yu。运用机器学习估计异质性治疗效应的元学习器。美国国家科学院院刊，116(10):4156–4165，2019年。[38] Heejung Bang 和 James M Robins。缺失数据和因果推断模型中的双重稳健估计。《生物统计学》，61(4):962–973，2005年。[39] Dylan J Foster 和 Vasilis Syrgkanis。正交统计学习。《统计年刊》，51(3):879–908，2023年。[40] Victor Chernozhukov、Denis Chetverikov、Mert Demirer、Esther Duflo、Christian Hansen、Whitney Newey 和 James Robins。针对处理效应和结构参数的双重/去偏机器学习，2018年。[41] Susan Athey、Julie Tibshirani 和 Stefan Wager。广义随机森林。《统计年刊》，47(2):1148–1178，2019年。[42] P Richard Hahn、Jared S Murray 和 Carlos M Carvalho。用于因果推断的贝叶斯回归树模型：正则化、混杂与异质性效应（附讨论）。《贝叶斯分析》，15(3):965–1056，2020年。[43] Ivan Henriquez、Mack Roach III、Todd M Morgan、Alberto Bossi、Junior A Gomez、Oscar Abuchaibe 和 Felipe Counago。转移性去势抵抗性前列腺癌（mCRPC）的当前及新兴疗法。《生物医学》，9(9):1247，2021年。[44] Martin W Schoen、Kenneth R Carson、Seth A Eisen、Charles L Bennett、Suhong Luo、Melissa A Reimers、Eric M Knoche、Alison L Whitmer、Yan Yan、Bettina F Drake 等。基于共病，接受恩扎卢胺和阿比特龙治疗的转移性去势抵抗性前列腺癌退伍军人的生存情况。《前列腺癌与前列腺疾病》，26(4):743–750，2023年。[45] Stephen J Freedland、Xuehua Ke、Marie-Hélène Lafeuille、Hela Romdhani、Frederic Kinkead、Patrick Lefebvre、Allison Petrilla、Zul Pulungan、Seung Kim、Denise M D’Andrea 等。利用行政健康理赔数据和实验室数据识别转移性激素敏感性或转移性去势抵抗性前列腺癌患者。《当前医学研究与观点》，37(4):609–622，2021年。[46] Jennifer L Hill。用于因果推断的贝叶斯非参数建模。计算与图形统计杂志，20(1):217–240，2011年。[47] Kosuke Imai 和 Marc Ratkovic。在随机化项目评估中估计治疗效应异质性。《应用统计年刊》，第443–470页，2013年。

混合变量类型的对数密度比计算

本附录给出了用于计算条件对数似然的明确公式 s

$$\log p_\theta(X \mid T' = t), \qquad t \in \{0, 1\},$$

针对处理前生成器中的混合类型协变量。这些项是计算诱导重叠函数所必需的

$$\log \alpha_\theta(X) = \log p_\theta(X \mid T' = 1) - \log p_\theta(X \mid T' = 0),$$

该函数会用于训练目标中的重叠匹配惩罚项。

我们假设解码器内部具有条件独立结构:

$$p_\theta(X \mid z, T' = t) = \prod_{i=1}^{V} p_\theta(X_i \mid z, T' = t),$$

因此，总对数似然可分解为

$$\log p_\theta(X \mid z, T' = t) = \sum_{i=1}^{V} \log p_\theta(X_i \mid z, T' = t).$$

以下公式概述了二元、类别型和连续型协变量中每个变量的贡献。

A.1 二元变量

对于二元变量 $X_i \in \{0, 1\}$，解码器输出对数几率 s_i，其中

$$p_i = \sigma(s_i) = \frac{1}{1 + e^{-s_i}}.$$

伯努利对数似然贡献为

$$\log p_\theta(X_i \mid z, T' = t) = X_i \log p_i + (1 - X_i) \log(1 - p_i).$$

这等价于按样本计算的二元_交叉_熵_与_对数几率的负值。

A.2 分类变量

对于分类变量 $X_i \in \{0, \ldots, K - 1\}$，解码器输出对数几率 $s_i \in \mathbb{R}^K$。softmax 概率为

$$p_{i,c} = \frac{\exp(s_{i,c})}{\sum_{k=1}^{K} \exp(s_{i,k})}.$$

观测类别 X_i 的对数似然贡献为

$$\log p_\theta(X_i \mid z, T' = t) = \log p_{i,X_i} = s_{i,X_i} - \log\left(\sum_{k=1}^{K} e^{s_{i,k}}\right),$$

这对应于每个样本的交叉_熵负损失。

A.3 连续变量

对于连续变量 $X_i \in \mathbb{R}$，解码器输出均值 μ_i 和有界对数方差 $\log \sigma_i^2$。模型假设

$$X_i \mid z, T' = t \sim \mathcal{N}(\mu_i, \sigma_i^2),$$

因此，高斯对数密度为

$$\log p_\theta(X_i \mid z, T' = t) = -\frac{1}{2}\left[\log(2\pi) + \log \sigma_i^2 + \frac{(X_i - \mu_i)^2}{\sigma_i^2}\right].$$

A.4 对数密度比计算

将各特征项相加即可得到条件对数似然

$$\log p_\theta(X \mid T' = t) = \sum_{i=1}^{V} \log p_\theta(X_i \mid z, T' = t).$$

用于重叠惩罚的诱导对数密度比如下：

$$\log \alpha_\theta(X) = \log p_\theta(X \mid T' = 1) - \log p_\theta(X \mid T' = 0).$$

主目标函数中的重叠控制项会惩罚与用户指定目标之间的平方偏差：

$$\mathcal{L}_\alpha(\theta) = \mathbb{E}_X \left[\left(\log \alpha_\theta(X) - \log \alpha(X) \right)^2 \right].$$

这种机制能够对合成数据分布中的正值性和协变量重叠进行精细控制。

B 评估指标汇总

表 B.1： 三个维度的评估指标：分布保真度、因果结构保真度和隐私。

Axis	类别（变量类型）	指标	描述／公式
分布保真度	边缘分布（连续）	归一化瓦瑟斯坦距离	$\frac{W_1(F^{(c)}_{\text{真实数据}}, F^{(c)}_{\text{合成数据}})}{\text{标准差}^{(c)}_{\text{真实数据}} + \varepsilon}$.
	边缘分布（连续）	KS 补数 (sdmetrics)	$1 - D_{\text{KS}}$.
	边缘分布（离散）	TVComplement (sdmetrics)	$1 - \frac{1}{2}\sum_x \lvert p_{\text{real}}(x) - p_{\text{synth}}(x)\rvert$。
	成对（所有变量）	SU 相似度	$1 - \lvert \text{SU}_{\text{real}}(i,\ j) - \text{SU}_{\text{synth}}(i,\ j)\rvert$。
	成对（连续–连续）	相关性相似度（sdmetrics）	连续变量对之间皮尔逊相关的相似度 连续变量对。
	成对（离散–离散）	列联相似度 (sdmetrics)	离散列联的相似度 通过总变差的结构。
	条件（$Z \mid C$）	加权 / 归一化 MMD^2	分层规模加权混合核 $\text{MMD}^2(Z \mid C)$；归一化依据：真实数据内部基线。
	联合（所有变量）	能量距离	归一化能量距离。
	联合（所有变量）	C2ST	$\lvert 2 \cdot \text{曲线下面积} - 0.5\rvert$。
因果结构保真度	处理效应	平均处理效应误差	$\lvert \text{平均处理效应}_\theta - \text{平均处理效应}_{\text{target}}\rvert$。
	处理效应	CATE平均绝对误差	$\mathbb{E}\big[\ \lvert \tau_\theta(X') - \tau(X')\rvert\ \big]$。
	处理效应	条件平均处理效应相关性	皮尔逊相关系数 $\rho(\tau_\theta,\ \tau)$。
	处理效应	Wasserstein 距离	W_1 两种经验分布之间 $\tau_\theta(X')$ 和 $\tau(X')$。
	混杂	平均绝对误差（总体 / 按组）	$\mathbb{E}[\lvert \kappa_\theta(X',\ T') - \kappa(X',\ T')\rvert]$；按治疗组。
	混杂	Wasserstein 距离	W_1 经验混杂分布。
	重叠	MSE	$\mathbb{E}\big[(\text{对数}\ \alpha_\theta(X') - \text{对数}\ \alpha(X'))^2\big]$。
	重叠	重叠系数	$\int_0^1 \min\ \big(p_0(e)\ p_1(e)\big)\ de$.
	重叠	倾向得分AUC	逻辑回归预测的曲线下面积 T' 来自 X'。
隐私	DCR	保护比例	$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{d_{\text{syn}}(i) > d_{\text{real}}(i)\}$.
	DCR	距离比	$\rho_i = \frac{d_{\text{syn}}(i)}{d_{\text{real}}(i) + \varepsilon}$.
	DCR	DCRBaselineProtection（sdmetrics）	最小$\Big(1,\ d_{\text{synth}}/d_{\text{rand}}\Big)$。

缩写和符号：*cont.* = 连续型；*disc.* = 离散型（二元或类别型）；*TE* = 处理效应；*dist.* = 距离；C = 条件变量；Z =除 C 外的所有变量。对于连续型特征 c：$F(c)$真实数据 $F(c)$合成数据= 经验累积分布函数；W_1 = *1-Wasserstein* 距离；D_{KS} = 柯尔莫哥洛夫–斯米尔诺夫统计量。对于离散型 x：$p_{\text{real}}(x)$ $p_{\text{synth}}(x)$ = 经验概率质量函数。$SU_{\text{real}}(i, j)$、$SU_{\text{synth}}(i, j)$ = 对称不确定性（真实数据与合成数据）。MMD^2 = 最大均值差异。(X', T') = 合成样本；$ATE_{\text{target}} = \mathbb{E}[\tau(X')]$, 平均处理效应$_\theta = \mathbb{E}[\tau_\theta(X')]$。$e = \mathbb{P}(T' = 1 \mid X')$；$p_0(e)$ $p_1(e)$ = 组特定密度。*DCR* = 到最近$\tilde{}$ $\tilde{}$真实记录的距离（编码空间）；$d_{\text{syn}}\ i$ $d_{\text{real}}\ i$ = $d_{\text{synth}}\ d_{\text{rand}}\ (\ )\ (\ )$ 到合成记录和另一条$\bar{}$真实记录的最近邻距离；，*DCR*中位数（合成、随机）。标记为 *sdmetrics* 的指标使用 *sdmetrics*库。

C 隐私评估

最近记录距离（DCR）是一种基于最近邻的准则，用于量化合成数据在多大程度上避免复现或近似真实个体 [35； 36]。设 $\mathcal{R} = \{r_i\}_{i=1}^{n}$ 和 $\mathcal{S} = \{s_j\}_{j=1}^{m}$ 分别表示经过连续变量标准化和离散变量独热编码后的真实数据集和合成数据集。

对于每条真实记录 r_i，我们计算两个距离：

$$d_{\text{syn}}(i) = \min_{s_j \in \mathcal{S}} \|r_i - s_j\|, \qquad d_{\text{real}}(i) = \min_{\substack{r_k \in \mathcal{R} \\ k \neq i}} \|r_i - r_k\|.$$

r_i 的保护准则定义如下：

$$\mathbf{1}_{\text{protected}}(i) = \mathbf{1}\{d_{\text{syn}}(i) > d_{\text{real}}(i)\},$$

that 也就是说，如果一条真实记录的最近合成近邻比它 s
clos更远的其他 真实近邻还要远，则认为该真实记录受到保护。因此，DCR保护比例为

$$\text{DCR Protection} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\text{protected}}(i),$$

数值越大表示隐私保护越强。

为提供更精细的分辨率，我们计算距离比：

$$\rho(i) = \frac{d_{\text{syn}}(i)}{d_{\text{real}}(i) + \varepsilon},$$

并报告包括均值、中位数和选定分位数（例如第 5、50、95 百分位数）在内的汇总统计量。比值显著大于 1 表明合成数据集在特征空间中与真实数据分离良好。

我们还报告 sdmetrics库中的 DCRBaselineProtection得分。该得分将合成数据的 DCR 中位数按随机数据的 DCR 中位数进行归一化（截断至 $[0, 1]$）；数值接近 1 表示其隐私水平与随机噪声相当。

$$\text{DCRBaselineProtection} = \min\left(1,\ \frac{\text{median}_{x \in \mathcal{D}_{\text{synth}}}\ d(x, \mathcal{D}_{\text{real}})}{\text{median}_{r \in \mathcal{D}_{\text{rand}}}\ d(r, \mathcal{D}_{\text{real}})}\right)$$

D 其他结果

表 D.1：使用贝叶斯高斯混合模型和标准高斯先验时情景1中的分布保真度。

类别	指标	方向	BGMM	高斯
边缘分布（连续）	归一化瓦瑟斯坦距离（均值）	↓ 更好	0.071（0.001）	0.336（0.010）
边缘分布（连续）	KS 补数（均值）	↑ 更好	0.939 (0.001)	0.853 (0.006)
边缘分布（离散）	总变差补数（均值）	↑ 越高越好	0.973 (0.001)	0.939 (0.001)
成对（连续–连续）	相关性相似度	↑ 越高越好	0.987 (0.002)	0.915 (0.005)
成对（所有变量）	SU 相似度（均值）	↑ 更好	0.994 (0.002)	0.992 (0.000)
成对（离散–离散）	列联相似度（均值）	↑ 更好	0.953 (0.001)	0.899 (0.002)
条件（除 C 外的所有项）[a]	加权 MMD^2	↓ 更好	0.004 (0.001)	0.007 (0.002)
条件（除 C 外的所有项）[a]	归一化 MMD^2 比率与真实数据的对比	↓ 更好	1.533 (0.154)	2.531 (0.601)
联合（所有变量）	归一化能量距离	↓ 更好	0.004 (0.001)	0.025 (0.001)
联合（所有变量）	分类器两样本检验（AUC 补数）	↑ 更好	0.748 (0.009)	0.551 (0.013)

[a] 条件变量（C）为暴露。

表 D.2：场景 2 中使用贝叶斯高斯混合模型和标准高斯先验的分布保真度 .

类别	指标	方向	BGMM	高斯
边缘分布（连续）	归一化瓦瑟斯坦距离（均值）	↓ 更好	0.111 (0.004)	0.360 (0.011)
边缘分布（连续）	KS 补数（均值）	↑ 更好	0.957 (0.001)	0.858 (0.004)
边缘分布（离散）	总变差补数（均值）	↑ 更好	0.978 (0.001)	0.928 (0.002)
成对（连续–连续）	相关性相似度	↑ 更好	0.985 (0.004)	0.950 (0.007)
成对（所有变量）	SU 相似度（均值）	↑ 更好	0.993 (0.002)	0.990 (0.000)
成对（离散–离散）	列联相似度（均值）	↑ 更好	0.956 (0.001)	0.881 (0.003)
条件（除 C 外的所有项）[a]	加权 MMD^2	↓ 更好	0.006 (0.001)	0.016 (0.001)
条件（除 C 外的所有项）[a]	归一化 MMD^2 与真实数据的比值	↓ 越好	2.244 (0.446)	5.981 (0.358)
联合（所有变量）	归一化能量距离	↓ 更好	0.005 (0.000)	0.033 (0.002)
联合（所有变量）	C2ST（AUC 补数）	↑ 更好	0.699 (0.016)	0.510 (0.016)

[a] 条件变量 (C) 是暴露。

表 D.3：在场景 3 中使用贝叶斯高斯混合模型和标准高斯先验的分布保真度。

类别	指标	方向	BGMM	高斯
边缘分布（连续）	归一化瓦瑟斯坦距离（均值）	↓ 更好	0.075 (0.001)	0.679 (0.008)
边缘分布（连续）	KS 补数（均值）	↑ 更好	0.954 (0.001)	0.720 (0.005)
边缘分布（离散）	总变差补数（均值）	↑ 更好	0.961 (0.001)	0.898 (0.002)
成对（连续–连续）	相关性相似度	↑ 更好	0.984 (0.002)	0.883 (0.009)
成对（所有变量）	SU 相似度（均值）	↑ 更好	0.991 (0.002)	0.990 (0.000)
成对（离散–离散）	列联相似度（均值）	↑ 更优	0.928 (0.001)	0.841 (0.003)
条件（除 C 外的全部）[a]	加权 MMD^2	↓ 更好	0.023 (0.002)	0.029 (0.002)
条件（除 C 外的所有项）[a]	归一化 MMD^2 比率（相对于真实数据）	↓ 更好	8.563 (0.604)	10.863 (0.724)
联合（所有变量）	归一化能量距离	↓ 更优	0.011 (0.000)	0.080 (0.002)
联合（所有变量）	C2ST（AUC 补数）	↑ 更好	0.625 (0.008)	0.333 (0.010)

[a] 条件变量（C）是暴露。

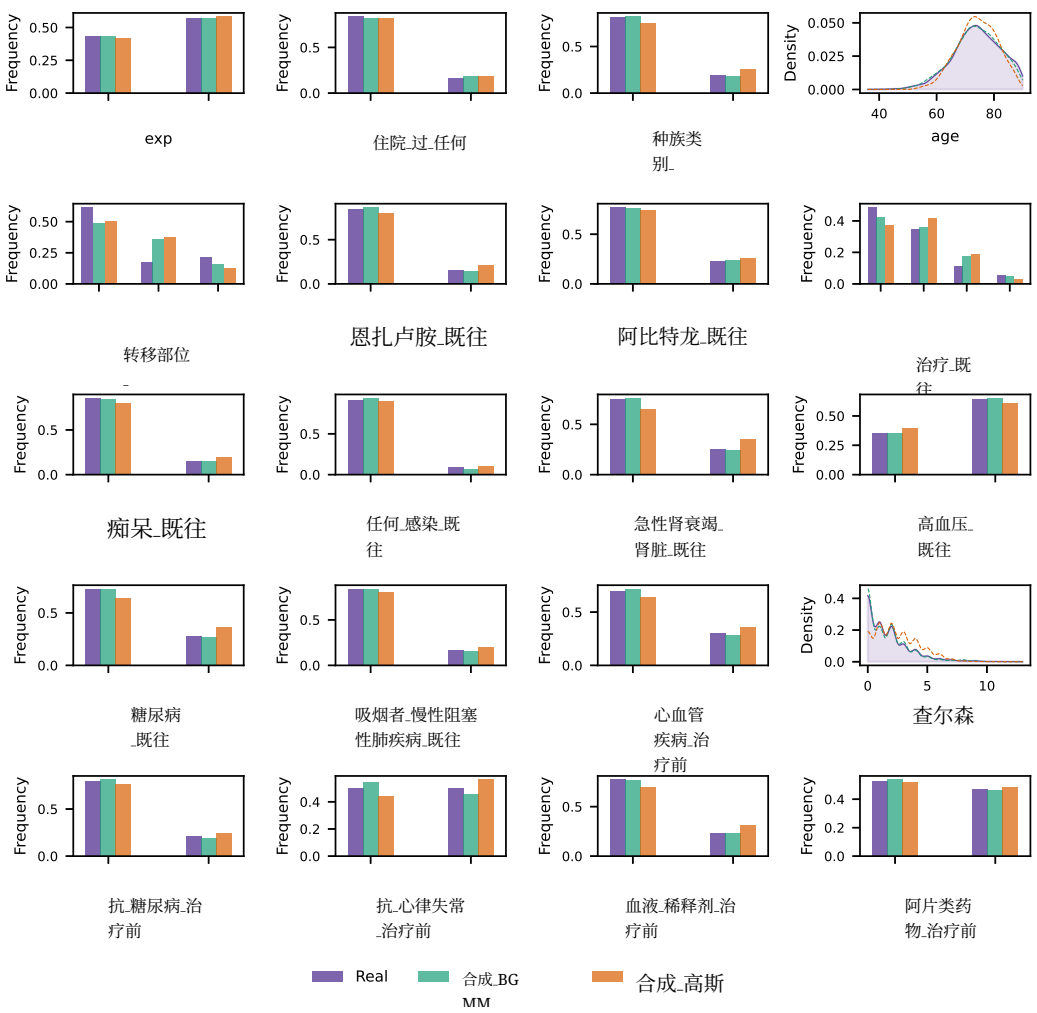


图 D.1：比较贝叶斯高斯混合模型和高斯先验下合成数据中所有变量的边际分布与真实 mCRPC 数据集（场景 3）。

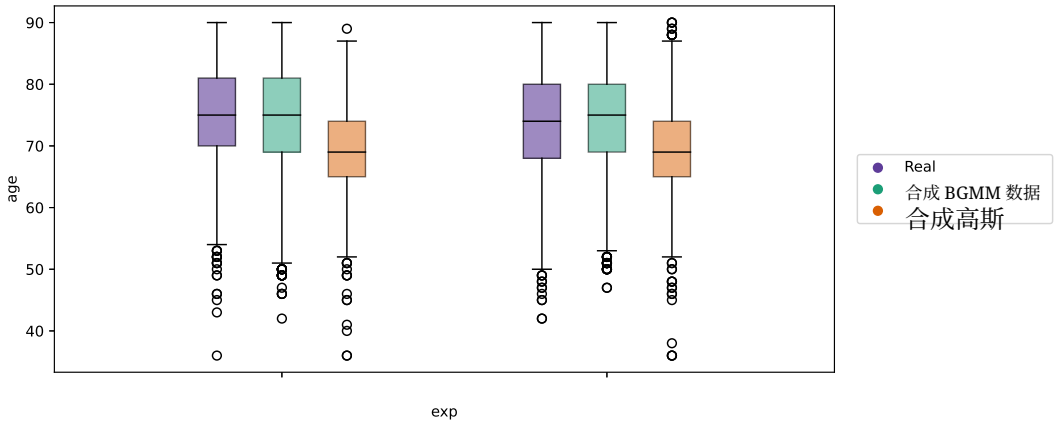


图 D.2：比较贝叶斯高斯混合模型和高斯先验下合成数据中暴露和年龄的成对分布与真实 mCRPC 数据集（场景 3）。

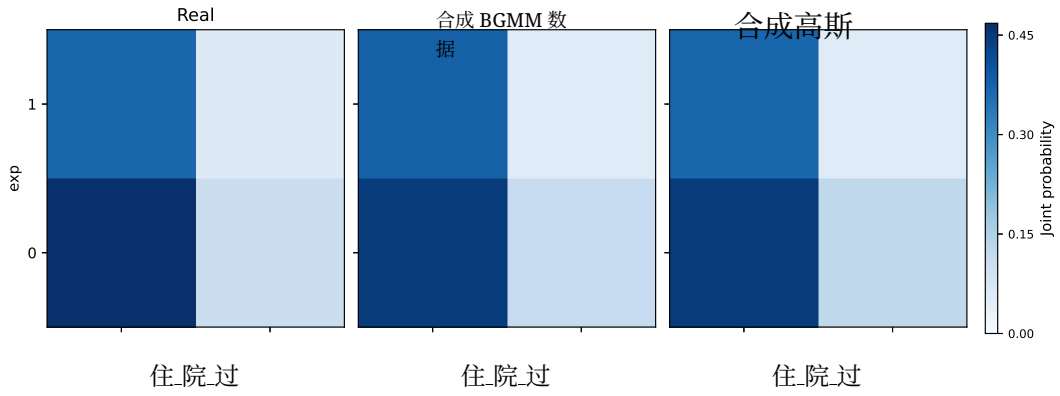


图 D.3：场景 3 中，采用贝叶斯高斯混合模型和高斯先验的合成数据中暴露与结局的成对分布，与真实 mCRPC 数据集的对比。

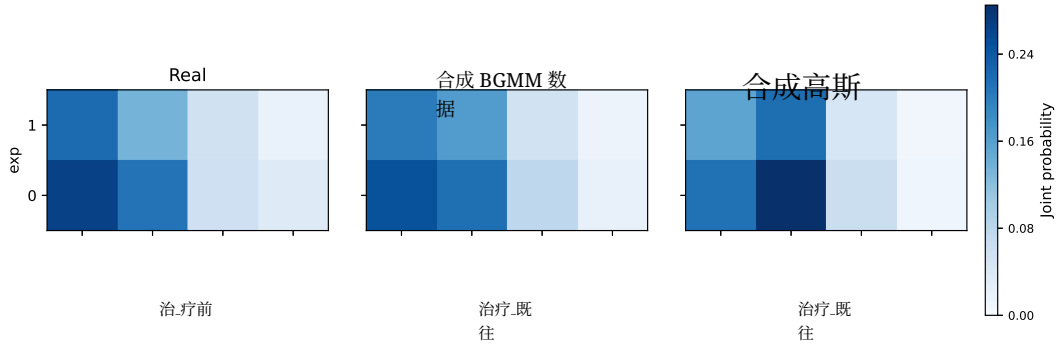


图 D.4：场景 3 中，采用贝叶斯高斯混合模型和高斯先验的合成数据中暴露与治疗史的成对分布，与真实 mCRPC 数据集的对比。

表 D.4：情景1中使用贝叶斯高斯混合模型和标准高斯先验生成的合成数据的因果结构保真度。

类别	指标	方向	BGMM	高斯
处理效应	条件平均处理效应/个体处理效应平均绝对误差	↓ 更好	0.003 (0.000)	0.004 (0.000)
处理效应	条件平均处理效应相关性	↑ 更好	NA	NA
处理效应	平均处理效应误差	↓ 更好	0.000 (0.000)	0.002 (0.000)
处理效应	TE分布距离（W1）	↓ 更好	0.003 (0.000)	0.004 (0.000)
混杂	混杂MAE	↓ 更好	0.000 (0.000)	0.000 (0.000)
混杂	分组平均绝对误差 (T=0)	↓ 更好	0.000（0.000）	0.000（0.000）
混杂	分组平均绝对误差（T=1）	↓ 更好	0.000 (0.000)	0.000 (0.000)
混杂	混杂分布 (W1)	↓ 更好	0.000 (0.000)	0.000（0.000）
重叠度（解码器）	MSE	↓ 更好	0.000（0.000）	0.000（0.000）
重叠度（解码器）	容差范围内的比例	↑ 更好	1.000 (0.000)	1.000 (0.000)
重叠度（倾向评分）	倾向得分AUC[a]	NA	0.539 (0.007)	0.536 (0.008)
重叠度（倾向评分）	直方图重叠系数	↑ 更好	0.940 (0.017)	0.947 (0.015)
重叠度（倾向评分）	共同支持比例	↑ 更好	1.000 (0.000)	1.000 (0.000)
重叠度（倾向评分）	共同支持比例 (T=0)	↑ 更好	1.000 (0.000)	1.000 (0.000)
重叠度（倾向评分）	共同支持比例 (T=1)	↑ 更好	1.000 (0.000)	1.000 (0.000)

[a]曲线下面积越接近 0.5，表示重叠程度越高。

表 D.5：场景 2 中使用贝叶斯高斯混合模型和标准高斯先验生成的合成数据的因果结构保真度。

类别	指标	方向	BGMM	高斯
处理效应	条件平均处理效应/个体处理效应平均绝对误差	↓ 更好	0.003（0.000）	0.004（0.000）
处理效应	条件平均处理效应相关性	↑ 更好	0.971 (0.002)	0.915 (0.004)
处理效应	平均处理效应误差	↓ 更好	0.000 (0.000)	0.001 (0.000)
处理效应	TE分布距离（W1）	↓ 更好	0.001 (0.000)	0.001 (0.000)
混杂	混杂MAE	↓ 更好	0.002 (0.000)	0.002 (0.000)
混杂	分组平均绝对误差 (T=0)	↓ 更好	0.002 (0.000)	0.002 (0.000)
混杂	分组平均绝对误差 (T=1)	↓ 更好	0.002 (0.000)	0.002 (0.000)
混杂	混杂分布 (W1)	↓ 更好	0.002 (0.000)	0.002 (0.000)
重叠度（解码器）	MSE	↓ 更好	0.004 (0.000)	0.005 (0.000)
重叠度（解码器）	容差范围内的比例	↑ 更好	0.997 (0.001)	0.993 (0.001)
重叠度（倾向评分）	倾向得分AUC[a]	NA	0.561 (0.007)	0.563 (0.008)
重叠度（倾向评分）	直方图重叠系数	↑ 更好	0.850 (0.021)	0.855 (0.015)
重叠度（倾向评分）	共同支持比例	↑ 更好	1.000 (0.000)	1.000 (0.000)
重叠度（倾向评分）	共同支持比例（T=0）	↑ 更好	1.000 (0.000)	1.000 (0.000)
重叠度（倾向评分）	共同支持比例 (T=1)	↑ 更好	1.000 (0.000)	1.000 (0.000)

[a]曲线下面积越接近 0.5，表示重叠程度越高。

表 D.6：场景 3 中使用贝叶斯高斯混合模型和标准高斯先验生成的合成数据的因果结构保真度。

类别	指标	方向	BGMM	高斯
处理效应	条件平均处理效应/个体处理效应平均绝对误差	↓ 更好	0.004 (0.000)	0.004 (0.000)
处理效应	条件平均处理效应相关性	↑ 更好	0.956 (0.003)	0.931 (0.005)
处理效应	平均处理效应误差	↓ 更好	0.001 (0.000)	0.002 (0.000)
处理效应	TE分布距离（W1）	↓ 更好	0.002 (0.000)	0.002 (0.000)
混杂	混杂MAE	↓ 更好	0.004 (0.000)	0.005 (0.000)
混杂	分组平均绝对误差 (T=0)	↓ 更好	0.005 (0.000)	0.005 (0.000)
混杂	分组平均绝对误差 (T=1)	↓ 更好	0.003 (0.000)	0.004 (0.000)
混杂	混杂分布 (W1)	↓ 更好	0.002 (0.000)	0.002 (0.000)
重叠度（解码器）	MSE	↓ 更好	0.004 (0.000)	0.300 (0.091)
重叠度（解码器）	容差范围内的比例	↑ 更好	0.996 (0.001)	0.934 (0.004)
重叠度（倾向评分）	倾向得分AUC[a]	NA	0.889 (0.006)	0.859 (0.005)
重叠度（倾向评分）	直方图重叠系数	↑ 更好	0.243 (0.012)	0.309 (0.008)
重叠度（倾向评分）	共同支持比例	↑ 更好	0.998 (0.003)	1.000 (0.000)
重叠度（倾向评分）	共同支持比例 (T=0)	↑ 更好	1.000 (0.000)	1.000 (0.000)
重叠度（倾向评分）	共同支持比例 (T=1)	↑ 越好	0.996 (0.004)	1.000 (0.000)

[a]曲线下面积越接近 0.5，表示重叠程度越高。

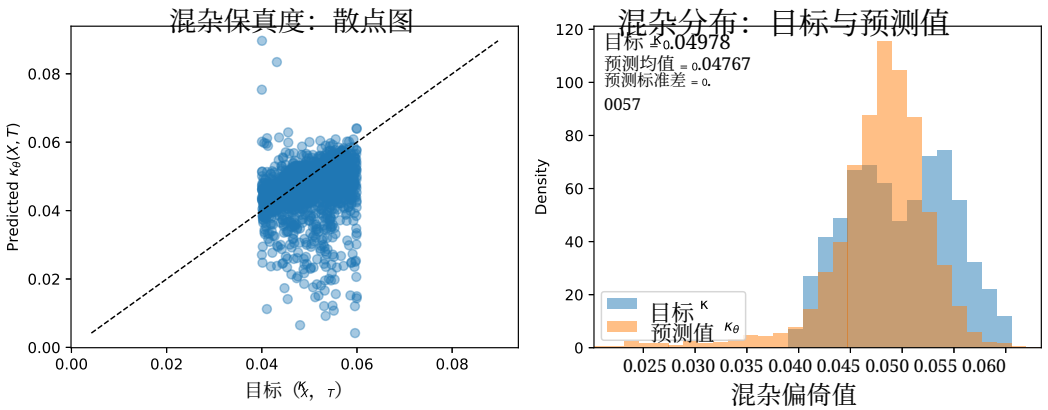
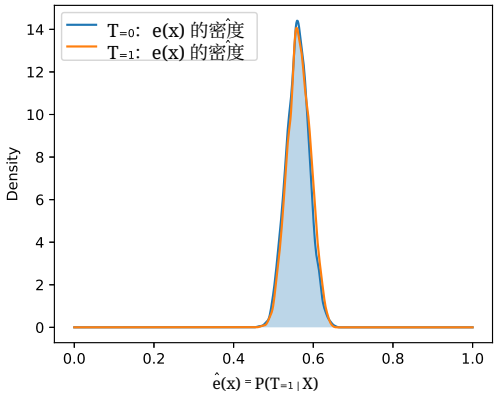
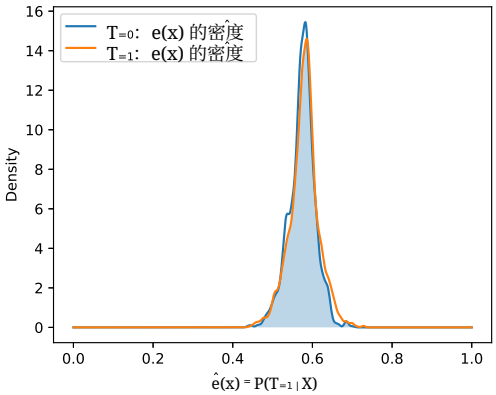


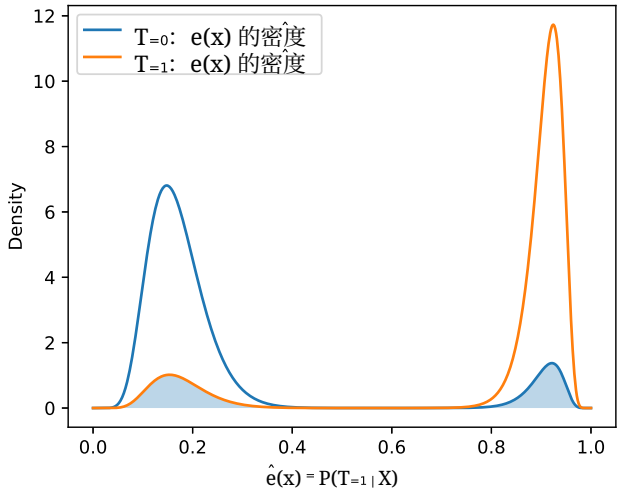
图 D.5：在场景 3 中，使用 BGMM 先验的 CAUSALMIX 生成的 mCRPC 数据集未测量混杂与目标未测量混杂的比较。



(a) 情景1



(b) 场景 2



(c) 场景 3

图 D.6：CausalMix 使用 BGMM 先验在不同因果情景下生成的 mCRPC 数据集重叠情况。情景1：完全重叠 $\log \alpha(X) = 0$；情景 2：中等恒定重叠 $\log \alpha(X) = 1$；情景 3：协变量依赖型重叠 $\log \alpha(X) = 2(2 \text{阿比特龙_prev} - 1)$。

表 D.7：情景中使用 BGMM 和标准高斯先验的合成数据隐私 1.

指标	方向	BGMM	高斯
保护比例	↑ 更好	0.595 (0.009)	0.696 (0.008)
距离比（均值）	↑ 更好	2.535 (0.083)	3.254 (0.163)
距离比（第5百分位数）	↑ 更好	0.233 (0.028)	0.501 (0.003)
距离比（第50百分位数）	↑ 更好	1.039 (0.006)	1.122 (0.006)
距离比（第95百分位数）	↑ 更好	4.841 (0.263)	6.150 (0.304)
标准化距离比	↑ 更好	0.276 (0.003)	0.333 (0.003)

表 D.8：场景中的贝叶斯高斯混合模型与标准高斯先验合成数据的隐私性 2.

指标	方向	BGMM	高斯
保护比例	↑ 更好	0.607 (0.007)	0.747 (0.007)
距离比（均值）	↑ 更好	2.787 (0.193)	3.723 (0.249)
距离比（第5百分位数）	↑ 更好	0.255 (0.021)	0.624 (0.036)
距离比（第50百分位数）	↑ 更好	1.057 (0.007)	1.195 (0.009)
距离比（第95百分位数）	↑ 更优	4.623 (0.207)	7.681 (0.622)
标准化距离比	↑ 更好	0.282 (0.002)	0.373 (0.007)

表 D.9：场景 3 中使用贝叶斯高斯混合模型和标准高斯先验生成的合成数据的隐私。

指标	方向	BGMM	高斯
保护比例	↑ 更好	0.720 (0.006)	0.839 (0.004)
距离比（均值）	↑ 更好	3.298 (0.245)	4.287 (0.089)
距离比（第5百分位数）	↑ 更好	0.396 (0.030)	0.788 (0.014)
距离比（第50百分位数）	↑ 更好	1.169 (0.007)	1.337 (0.009)
距离比（第95百分位数）	↑ 更好	6.046 (0.055)	11.740 (0.493)
标准化距离比	↑ 更好	0.297 (0.003)	0.559 (0.005)

表 D.10：无未测量混杂情况下因果估计量的准确率和覆盖率

估计量	运行时间（秒，标准差）	ATE 均方根误差	平均 ATE 标准误	ATE 覆盖率	平均 CATE 覆盖率
贝叶斯因果森林	258.7 (9.5)	0.010	0.010	0.90	0.91
因果森林	25.1 (1.7)	0.009	0.009	0.94	0.87
标准线性双重机器学习	2.2 (0.1)	0.009	0.009	0.96	0.90
Lasso 双重机器学习	229.9 (10.1)	0.009	0.008	0.88	0.37
梯度提升回归双重机器学习	268.1 (10.7)	0.010	0.113	1.00	0.82
标准线性双重稳健学习器	2.3 (0.1)	0.010	0.046	1.00	0.92
Lasso 双重稳健学习器	234.5 (10.6)	0.010	0.009	0.88	0.40
梯度提升回归 双重稳健学习器	274.0 (10.8)	0.010	0.125	1.00	0.82
Lasso X 学习器	118.4 (4.1)	0.024	0.036	1.00	0.82
梯度提升回归 X 学习器	160.4 (4.4)	0.024	0.085	1.00	0.82

表 D.11：轻微未测量混杂下因果估计量的准确率和覆盖率

估计量	运行时间（秒，标准差）	ATE 均方根误差	平均 ATE 标准误	ATE 覆盖率	平均 CATE 覆盖率
贝叶斯因果森林	260.6 (9.6)	0.023	0.011	0.48	0.86
因果森林	25.3 (1.8)	0.027	0.009	0.18	0.80
标准线性双重机器学习	2.1 (0.1)	0.026	0.009	0.24	0.84
Lasso 双重机器学习	225.0 (7.9)	0.025	0.010	0.28	0.26
梯度提升回归双重机器学习	262.8 (7.7)	0.026	0.125	1.00	0.83
标准线性双重稳健学习器	2.2 (0.1)	0.028	0.046	1.00	0.85
Lasso 双重稳健学习器	230.5 (8.4)	0.028	0.008	0.14	0.25
GBR 双重稳健学习器	269.7 (7.3)	0.028	0.132	1.00	0.81
Lasso X 学习器	116.4 (3.8)	0.008	0.035	1.00	0.86
梯度提升回归 X 学习器	158.6 (4.3)	0.009	0.088	1.00	0.83

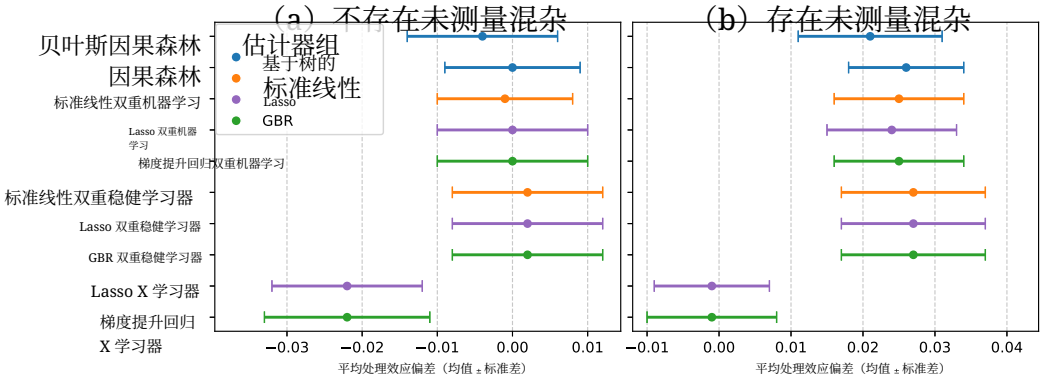


图 D.7：比较无未测量混杂（$\kappa = 0$）和中度未测量混杂（$\kappa = 0.02$）情况下各估计量的平均处理效应偏差。数据点表示 50 次独立重复实验的平均偏差，误差线表示一个标准差。估计器组根据条件平均处理效应模型定义。

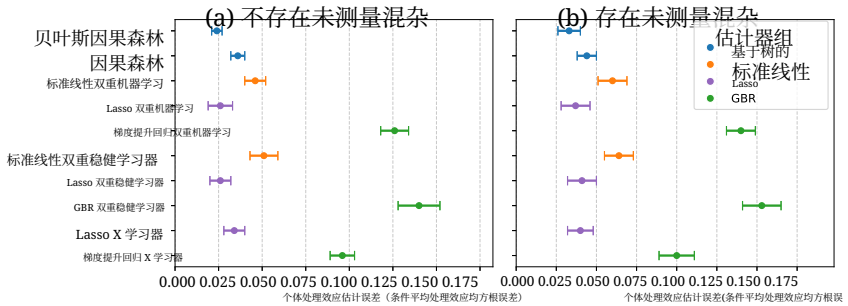


图 D.8：在无未测量混杂（$\kappa = 0$）和中度未测量混杂（$\kappa = 0.02$）条件下，各估计量的个体处理效应估计误差比较。点表示 50 次独立重复实验中的均方根误差均值，误差线表示一个标准差。估计器组根据条件平均处理效应模型定义。

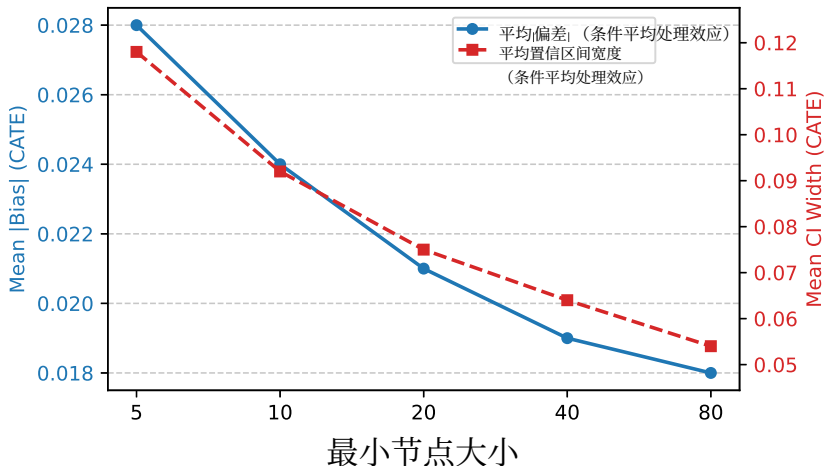


图 D.9：因果森林（2,000 棵树）中最小叶节点大小对条件平均处理效应准确度和不确定性的影响。实线表示条件平均处理效应偏差绝对值的均值，虚线表示条件平均处理效应置信区间宽度的均值，均值均基于 50 次重复试验计算。