arXiv:2302.01367v1 [stat.ML] 2 Feb 2023

基于梯度提升树的异质性处理效应增强学习

预印本

Heng Chen
生物统计学系 吉利德科学公司 福斯特城，CA 94404
Henry.Chen10@gilead.com

James Y.Dai*公共卫生科学部/生物统计学系 弗雷德·哈钦森癌症中心/华盛顿大学，美国华盛顿州西雅图
jdai@fredhutch.org

Michael L.LeBlanc公共卫生科学部/生物统计学系 弗雷德·哈钦森癌症中心/华盛顿大学，美国华盛顿州西雅图
mleblanc@fredhutch.org

2023年2月6日

摘要

基于患者遗传或临床因素的异质性治疗效应（HTE）对精准医学具有重要意义。对于包含高维预测标志物的随机临床试验，同时建模异质性治疗效应及其相应的主效应颇具挑战。受修正协变量方法启发，我们提出一种用于估计异质性治疗效应的两阶段统计学习程序，该程序具有最优效率增广，可推广至任意交互模型，并利用强大的极端梯度提升树（XGBoost）。对于定量结局，异质性治疗效应的目标估计量定义在均值差尺度上；对于二元结局，则定义在风险比尺度上；它们分别是特定损失函数的最小化解。第一阶段估计基线标志物对结局的主效应等价量，并在第二阶段的异质性治疗效应估计中将其用作增广项。所提出的两阶段程序对主效应的模型错设具有稳健性，并通过非参数函数估计（例如 XGBoost）提高异质性治疗效应估计的效率。我们提出了一种置换检验，用于整体评估异质性治疗效应的证据。我们对 SWOG癌症研究网络牵头的前列腺癌预防试验中的一项遗传学研究进行了分析，以展示该两阶段方法的特性和实用性。

关键词 增广 · 函数优化 · 标志物-治疗交互作用 · 置换检验 · 稳健性与效率

1 引言

精准医学是生物医学研究中一个新兴领域，旨在根据患者的遗传或临床背景，为其量身定制疗法或预防措施，而非采用传统的“一刀切”方法。实施精准医学的关键在于估计个体或亚组的治疗效应，这些效应会因基线“预测标志物”而异，即能够预测试验参与者之间异质性治疗效应（HTE）的标志物。 [Evans 和 McLeod，2003年，Weinshilboum 和 Wang， 2004][Simon 和 Roychowdhury， 2013]识别这些预测标志物也有助于加深对临床终点上治疗效应机制的理解。

当候选预测标志物数量庞大时，要高效且稳健地估计 e
异质性治疗效 在随机临床试验（RCT）中的效应。RCT 的样本量由主要……决定 y

*通讯作者

对总体治疗效应进行意向治疗分析，因此很少足以评估高维预测标志物。此外，同时对高维交互作用和主效应进行建模颇具挑战。令 Y 表示试验终点， T 表示随机治疗分配（通常治疗组取 1，对照组取 0）， X 表示预测标志物向量（其中第一个元素为 1，代表截距）。关于（Y， T， X）的一种常见参数工作模型为

$$g[\mathbb{E}(Y|X,T)] = \mathcal{A}(X;\boldsymbol{\alpha}) + \mathcal{F}(X;\boldsymbol{\beta})T, \tag{1}$$

其中， g 是连接函数， $\mathcal{A}$ 和 $\mathcal{F}$ 分别表示具有参数 $\boldsymbol{\alpha}$和 $\boldsymbol{\beta}$的主效应与异质性治疗效应函数。更一般的模型可能包括对 $\mathcal{A}$ 和 $\mathcal{F}$进行非参数函数建模的类别。虽然异质性治疗效应分量 $\mathcal{F}$ 是主要关注对象，但其估计依赖于对主效应分量$\mathcal{A}$进行适当处理。具体而言， $\mathcal{A}$ 设定了 Y 的基线，在此基础上定义异质性治疗效应；对 $\mathcal{A}$ 进行建模可以提高 $\mathcal{F}$的估计效率，但 $\mathcal{A}$ 的模型设定错误可能会使 $\mathcal{F}$的估计量产生偏差。 [VanderWeele等人， 2012] 为此，统计惯例是在 $\mathcal{A}$（X； $\boldsymbol{\alpha}$）与 $\mathcal{F}$（X； $\boldsymbol{\beta}$）之间施加约束：只有在模型中包含相应主效应时，才包含交互项，这被称为交互作用与主效应的“层级”或“遗传”性质。 [Cox，1984，Peixoto，1987，Choi等人， 2010] 当 X 为高维时， $\mathcal{F}$（X； $\boldsymbol{\beta}$）的模型选择和自适应估计变得具有挑战性，因为必须在层级约束下同时估计 $\mathcal{A}$（X； $\boldsymbol{\alpha}$）。例如，可将LASSO扩展到参数模型中，以估计交互作用。 [Bien等人，2013，Lim和Hastie， 2015] 已有研究提出了用于估计异质性治疗效应的非参数方法，例如通过树和随机森林进行递归分割， [Su等人，2009，Athey和Imbens，2016，Wager和Athey， 2018] 其本质是强制 $\mathcal{A}$ 和 $\mathcal{F}$采用相同的树结构。另一种方法是使用仅病例随机森林方法估计异质性治疗效应。 [Dai和LeBlanc，2019，Dai等人， 2019]

假设随机分配至治疗组或对照组的概率相等，Tian等人 [2014] 提出的修正协变量方法 $\mathcal{A}$利用随机分配机制，并去除主效应分量 $\mathcal{A}$。异质性治疗效应通过线性加性项建模 XT为 $g[\mathbb{E}(Y|X,\ T)] = \boldsymbol{\beta}^T(XT)$，其中 T编码为 $+1$或-1，分别表示以 0.5 的概率接受治疗或对照，而 XT为修正协变量。需要注意的是，除恒等函数外，修正协变量模型通常设定错误，且不一定能得到连接函数尺度上的异质性治疗效应 g。对于高维 X，Tian等人 [2014] 通过 LASSO 对 HTE 分量 $\mathcal{F}$ 进行正则化模型选择。该方法也支持采用其他非参数学习方法（如极端梯度提升树（XGBoost））来估计个体化治疗效应。 [Sugasawa和Noma, 2019]为提高效率，Tian等人 [2014] 提出在估计函数中加入增广项，其思路与半参数模型相同：通过纳入辅助协变量来提高效率。 [Tsiatis, 2007, Tsiatis等人, 2008, Zhang等人，2008] 一个特殊情形是简单线性模型，此时增广项等同于主效应分量，实质上回归到完整模型（1）。Chen等人 [2017] 将修正协变量方法推广到用于亚组识别的一般统计框架，并通过加权和 A-学习加以实现，具体方法见 Huling 和 Yu [2021]。

受修正协变量方法及上述半参数理论的启发，我们提出了一种两阶段统计学习程序，可对异质性治疗效应（HTE）的均值差和风险比估计目标进行自适应非参数建模。不同于对联合模型 (1) 进行标准最大似然估计，我们通过最小化修正协变量的专用损失函数来定义 HTE 的均值差和风险比估计目标，并将预测变量的主效应作为稳健增广项，以提高 HTE 估计的效率。设定错误的增广项不会影响 HTE 的一致估计，因此消除了 HTE 对主效应建模的依赖。我们还作出了三项创新性贡献：第一，对于任意交互模型，均可推导出均值差和风险比 HTE 的最优增广项；而 Tian等人 [2014] 和 Chen 等人 [2017] 则需要假设交互效应为零，才能推导出最优增广项。第二，我们设计了一种新颖的置换检验，用于评估在给定估计主效应分量的条件下 HTE 的显著性。第三，我们采用加权方法估计最优增广项，从而避免像 Tian等人 [2014], 、Chen 等人 [2017] 和 Huling 和 Yu [2021]那样分别估计两个条件均值。

将 HTE 与主效应增广在操作上分开，另一个好处是，对于高维协变量，我们能够利用强大的非参数学习算法，例如极端梯度提升树（XGBoost），自适应地学习主效应和 HTE；此前尚未有人尝试过这种做法。Friedman [2001] 通过函数空间中的最速下降优化提出了 GBT。Bühlmann 和 Yu [2003] 以及 Bühlmann [2006] 研究了适用于高维预测变量的 L2 提升算法，并评估了一致性。Chen 和 Guestrin[2016] 提出了带正则化的极端梯度提升树（XGBoost），通过惩罚复杂的树结构，减少了在任何预先指定的凸损失函数下训练提升树所需的计算时间。在本研究中，我们提出了一种通过 XGBoost 实现的新型两阶段梯度提升树算法，用于依次学习主效应和 HTE。

本文结构安排如下。第 2.1 节介绍异质性治疗效应估计目标和修正协变量方法。第 2.2 节分别针对

异质性治疗效应的均值差和风险比估计目标，介绍所提出的异质性治疗效应增广学习程序。第 2.2.3 节介绍用于检验异质性治疗效应的置换检验。第 2.2.4 节介绍通过改造 XGBoost 实现该程序的方法。第 3 节展示了大量模拟研究，用于比较我们的方法与现有方法的效率，并评估置换检验的有效性。第 4 节展示了所提出方法在前列腺癌预防试验（PCPT）试验）。第 5节以讨论作结。**2 方法**

2.1 异质性治疗效应估计目标

令 Y表示连续响应或二元响应， T表示治疗分配的二元指示变量，其中对照组编码为 -1，实验组编码为 $+1$。令 X 表示一个 p 维基线协变量向量，这些协变量可能会修饰治疗效应。治疗随机分配给研究参与者，且 $0 < \Pr(T = 1) < 1$，因此 T与 X相互独立。根据 Neyman [1923] 和 Rubin [1974], 的潜在结果框架，我们将 $Y^{(-1)}$和 $Y^{(1)}$ 分别记为参与者被分配到对照组或实验组时可能产生的响应。我们只能观测到一对潜在结果中的一个，即 $Y^{(t)}$，它是根据实际实现的治疗分配 $T = t$所观测到的响应 Y。假设观测数据由相互独立且同分布的随机向量 (Y_i, T_i, X_i)，构成，其中 $i = 1, \ldots, n$ 表示参与者。

异质性治疗效应（HTE）的估计目标定义为 $X = x$的函数。对于连续响应或二分结果，我们将 HTE 定义为

$$\tau(x) = \mathbb{E}(Y^{(1)} - Y^{(-1)}|X = x) = \mathbb{E}(Y|X = x, T = 1) - \mathbb{E}(Y|X = x, T = -1), \tag{2}$$

两个治疗组中，给定 X 时 Y 的期望之差。对于二元结局，异质性治疗效应的风险比估计目标可定义为

$$\tau(x) = \frac{\mathbb{E}(Y^{(1)}|X = x)}{\mathbb{E}(Y^{(-1)}|X = x)} = \frac{\mathbb{E}(Y|X = x, T = 1)}{\mathbb{E}(Y|X = x, T = -1)}. \tag{3}$$

本文主要关注函数 $\tau(x)$ 的估计与推断。如果 $\tau(x)$ 可简化为常数，则不存在异质性，且每位参与者的治疗效应都相同——这就是检验异质性治疗效应时的零假设。

标准的最大似然方法假设采用如式 (1) 所示的概率模型。当 X 为中等维度或高维时，难以同时用符合层级约束的灵活非参数函数来建模主效应分量 $\mathcal{A}(X)$ 和异质性治疗效应函数 $\mathcal{F}(X)$。因此，我们通过函数估计定义损失函数，使其直接针对异质性治疗效应估计目标 (2) 和 (3)，同时将主效应分量作为损失函数中的增广项单独处理。这种两阶段统计学习程序的灵感来自修正协变量方法 [Tian等人， 2014], 其中 T 编码为 $+1$ 和 -1，且二者的概率相等。补充材料 S1 简要概述了修正协变量方法。

2.2 异质性治疗效应的两阶段非参数估计

2.2.1 连续响应的均值差估计目标

假设 $T = +1$ 或 $T = -1$ 的随机分配概率不等于 $_1$ $_2$。为获得式（2）中定义的异质性治疗效应均值差估计目标，我们最小化以下由随机分配概率的倒数加权的期望平方误差损失函数，

$$\underset{\mathcal{F}}{\operatorname{argmin}} \quad \mathbb{E}\left[\left\{\frac{T+1}{2\Pr(T=1)} - \frac{T-1}{2\Pr(T=-1)}\right\}\left\{Y - \mathcal{F}(X)T\right\}^2\right]. \tag{4}$$

请注意，现在将 $\mathcal{F}(X)$ 定义为不含参数 $\boldsymbol{\beta}$ 的一般函数。通过这种逆概率加权构造，我们在一个“假想总体”中定义异质性治疗效应估计目标，在该总体中，治疗标签被均等分配。这有助于推导估计目标和最优增广。式（4）最小化器的估计方程为

$$\mathbb{E}\left\{S(Y, \mathcal{F}(X)T)\right\} = 0, \tag{5}$$

其中， $S(Y, \mathcal{F}(X)T) = \left\{\frac{T+1}{2\Pr(T=1)} + \frac{T-1}{2\Pr(T=-1)}\right\}\left\{Y - \mathcal{F}(X)T\right\}$ 是损失函数 (4) 关于 $\mathcal{F}(X)$ 的一阶导数；忽略常数因子 -2，通过在 $\left\{\frac{T+1}{2\Pr(T=1)} - \frac{T-1}{2\Pr(T=-1)}\right\}$ 中翻转负号来消去因子 T，并假设导数与期望具有可交换性。对估计函数

(5) 关于在 X 条件下的 T 条件分布进行积分，可得 $\mathbb{E}(Y|X,\ T=1)\ -\mathcal{F}(X)\ -\mathbb{E}(Y|X,\ T=-1)\ -\mathcal{F}(X)\ =0$.此时，最小化解是均值差估计目标的一半，

$$\mathcal{F}^*(x)=\frac{1}{2}\Big\{\mathbb{E}(Y|X=x,T=1)-\mathbb{E}(Y|X=x,T=-1)\Big\}.$$

为提高均值差估计目标的估计效率，可以在（5）中加入逆概率加权增广项，即 $a(X)\Big\{_2{}^{T+1}{}_{\Pr(T=1)}$ $+\ _2{}^{T-1}{}_{\Pr(T=-1)}\Big\}$，其期望为零。增广估计方程为:

$$\mathbb{E}\Big\{S^{\text{aug}}(Y,\mathcal{F}(X)T)\Big\}=0, \tag{6}$$

其中，$S^{\text{aug}}(Y,\mathcal{F}(X)T)\ =S(Y,\mathcal{F}(X)T)\ -a(X)\Big\{_2{}^{T+1}\Pr(T=1)\ +\ _2{}^{T-1}\Pr(T=-1)\Big\}$。由于增广项的期望为零，（5）和（6）的解相同。为获得使估计量 $\mathcal{F}(X)$ 的方差最小的最优增广项 $a_0(X)$，渐近而言，最小化增广估计函数 $S^{\text{aug}}(Y,\mathcal{F}(X)T)$ [Robins et al., 1994] 的方差与此等价。附录中推导得出，在任意交互模型、任意随机化比例以及对HTE进行非参数学习的条件下，最优增广项为$a_0(X)\ =\frac{1}{2}\Big\{\mathbb{E}(Y|X,T=1)\ +\mathbb{E}(Y|X,\ T=-1)\Big\}$。一个特殊情形是 $\Pr(T=1)\ =\frac{1}{2}$，其中 $a_0(X)\ =\mathbb{E}(Y|X)$。为了避免在随机化比例不相等时分别构建两个模型来估计 $\mathbb{E}(Y|X,T=1)$ 和 $\mathbb{E}(Y|X,T=-1)$，我们可以通过最小化以下加权损失函数，将 $\mathbb{E}(Y|X,T=1)\ +\mathbb{E}(Y|X,T=-1)$ 作为一个整体项进行估计:

$$\underset{\mathcal{A}}{\text{argmin}}\quad\mathbb{E}\left[\left\{\frac{T+1}{2\Pr(T=1)}-\frac{T-1}{2\Pr(T=-1)}\right\}\Big\{Y-\mathcal{A}(X)\Big\}^2\right], \tag{7}$$

可通过任何参数或非参数函数，或集成学习器来实现。

备注。 我们对求解增广估计函数 *(6)* 的异质性治疗效应估计量的性质作几点重要说明。从本质上说，这种方法通过求解损失函数，将异质性治疗效应定义为目标估计量，然后在估计函数中将主效应作为增广项加回，从而消除主效应与交互作用之间的依赖关系。-1 和 $+1$ 在两个随机分配的治疗组中的特殊结构决定了 T 与 X之间的正交性，因此，无论增广项的估计是否设定正确，异质性治疗效应的估计量都具有一致性。此外，仅凭增广估计函数 *(6)* 即可对异质性治疗效应进行有效推断，并将增广项的插件估计量视为已知量。这些性质与标准交互估计的局限性形成鲜明对比 *[VanderWeele等人， 2012]*。

上述推导还意味着，可以采用两阶段程序来求解增广估计函数：首先，可以在加权损失函数 (7) 下使用任何回归模型来估计 $a_0(X)$。简单的线性模型可能并非最优，但对于实现效率增广而言可能已足够。在第二阶段，可以通过最小化来得到估计量 $\mathcal{F}(X)$

$$\underset{\mathcal{F}}{\text{argmin}}\quad\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i+1}{2\Pr(T_i=1)}-\frac{T_i-1}{2\Pr(T_i=-1)}\right\}\Big\{Y_i-\hat{a}_0(x)-\mathcal{F}(X_i)T_i\Big\}^2. \tag{8}$$

$\mathcal{F}(X)$ 是目标化异质性处理效应，可通过任意非参数函数或集成学习器进行估计。本文采用 XGBoost 来实现这一目标，具体方法将在第 2.2.4 节中介绍。

2.2.2 二元响应的风险比估计目标

为了获得式 (3) 所表示的异质性处理效应风险比估计目标，我们采用类似的逆概率加权构造，并对逻辑损失函数的一种变体进行推广。该变体最初由 Tian等人（2014年）在补充材料中提出，即

$$\underset{\mathcal{F}}{\text{argmin}}\quad\mathbb{E}\left[\left\{\frac{T+1}{2\Pr(T=1)}-\frac{T-1}{2\Pr(T=-1)}\right\}\Big\{(1-Y)\mathcal{F}(X)T+Y\exp\left(-\mathcal{F}(X)T\right)\Big\}\right], \tag{9}$$

其中，期望是对 Y、T和 X 的联合分布取值的。为求解损失函数 (9) 的最小化解，估计方程为

$$\mathbb{E}\Big\{S(Y,\mathcal{F}(X)T)\Big\}=0, \tag{10}$$

其中， $S(Y,\ \mathcal{F}(X)T)\ =\left\{{}_2{}^{T+1}{}_{\Pr(T=1)}+{}_2{}^{T-1}{}_{\Pr(T=-1)}\right\}\left\{1-Y-Y\exp\left(-\mathcal{F}(X)T\right)\right\}$ 是对损失函数 (9) 关于 $\mathcal{F}(X)$ 求一阶导数所得的估计函数。我们对给定 X 条件下 T 的随机化分布中的估计方程 (10) 进行积分，得到 $-\left[1+\exp\left\{-\mathcal{F}(X)\right\}\right]\mathbb{E}(Y|X,\ T=1)+\left[1+\exp\left\{\mathcal{F}(X)\right\}\right]\mathbb{E}(Y|X,\ T=-1)=0$。其最小化解是对数风险比估计目标，

$$\mathcal{F}^*(x)=\log\left\{\frac{\mathbb{E}(Y|X=x,T=1)}{\mathbb{E}(Y|X=x,T=-1)}\right\}.$$

为了提高估计 风险比估计目标的效率，我们构造了一个类似于 我们在第 2.2.1节中所述的增广估计方程，

$$\mathbb{E}\left\{S^{\text{aug}}(Y,\mathcal{F}(X)T)\right\}=0, \tag{11}$$

其中 $S^{\text{aug}}(Y,\mathcal{F}(X)T)\ =S(Y,\mathcal{F}(X)T)\ -a(X)\left\{{}_2{}^{T+1}\Pr(T=1)+{}_2{}^{T-1}\Pr(T=-1)\right\}$。方程（10）和（11）的解相同，因为增广部分的期望为零。在任意交互模型、任意随机化比例以及异质性治疗效应的非参数学习下，我们推导出风险比估计目标的一个最优增广项， $a_0(X)\ =1-\mathbb{E}(Y|X,T=1)\ -\mathbb{E}(Y|X,T=-1)$，详见附录。一个特殊情形是 $\Pr(T=1)\ =\frac{1}{2}$，因此 $a_0(X)\ =1-2\mathbb{E}(Y|X)$。最优增广项 $\mathbb{E}(Y|X,T=1)\ +\mathbb{E}(Y|X,\ T=-1)$ 也可以通过最小化以下加权逻辑损失函数，作为一个整体项进行估计，

$$\underset{\mathcal{A}}{\text{argmin}}\ \mathbb{E}\left[\left(\frac{T_i+1}{2\Pr(T_i=1)}-\frac{T_i-1}{2\Pr(T_i=-1)}\right)\left(Y_i\mathcal{A}(X_i)-\log\left(1+\exp(\mathcal{A}(X_i))\right)\right)\right]. \tag{12}$$

首先，可以通过最小化式（12）中的加权逻辑损失函数，采用任意参数或非参数估计方法，轻松估计最优增广项。在第二阶段，将 $\hat{a}_0(X_i)$ 代入以下函数

$$\underset{\mathcal{F}}{\text{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{T_i+1}{2\Pr(T_i=1)}-\frac{T_i-1}{2\Pr(T_i=-1)}\right\}L_i^{\text{aug}}(Y_i,\mathcal{F}(X_i)T_i)\right], \tag{13}$$

其中 $L_i^{\text{aug}}(Y_i,\mathcal{F}(X_i)T_i)\ =\left(1-Y_i-\hat{a}_0(X_i)\right)\mathcal{F}(X_i)T_i+Y_i\exp\left(-\mathcal{F}(X_i)T_i\right)$，且 $\mathcal{F}$可以通过任意非参数函数或集成学习器进行估计。

有时，临床试验中的生物标志物研究会采用病例对照抽样方案，或更一般地采用带有附加因素的回顾性抽样。利用逆概率加权法将抽样权重纳入估计函数十分简单，我们在对 PCPT 试验进行数据分析时便采用了这种方法。

2.2.3 置换检验

异质性治疗效应推断的一个主要目标，是检验是否存在任何异质性治疗效应。置换检验已被用于检验交互作用 [Foster 等人， 2016], 但本研究的一项贡献是，基于两个估计过程的正交性，针对给定主效应时均值差和风险比异质性治疗效应的非参数估计，提出了一种置换检验。在不存在异质性治疗效应的零假设下， $\tau(X)$ 为常数，H_0：$\tau(X)\ =c$；在备择假设下， $\tau(X)$ 是 X 的函数，H_a：$\tau(X)\ =f(X)$。观测到的异质性治疗效应 $\tau\text{obs}(x)$ 由非参数算法得到，例如第 2.2.4 节中的算法1。我们计算其在参与者间的样本方差，记为 $s^2\{\tau^{\text{obs}}(x)\}\ =\frac{1}{n-1}\sum_{i=1}^{n}\left(\tau\text{obs}(x_i)\ -\underline{\tau\text{obs}}\right)^2$，其中 $\tau\text{obs}\ =\frac{1}{n}\sum_{i=1}^{n}\tau\text{obs}(x_i)$，以描述治疗效应在 X [的分布上的异质性（Levy 等人， 2021]）。若个体化治疗效应中存在离群值，则可使用稳健版本的样本方差，例如中位数绝对偏差。该统计量用作以下条件置换检验的检验统计量：$\text{Var}_X\{\tau(x)\}$ 在 H_0 下为零；在 H_a 下大于零，其中 $\text{Var}_X\{\tau(x)\}$ 是目标人群分布上 HTE 的方差。请注意，也可以检验 $\tau(X)$ 的其他函数，例如相对风险为 <0.5 的参与者比例是否大于 0。

在所提出的置换检验中，我们将第一阶段估计固定为根据观测数据估计得到的 $\hat{a}_0(x)$，并对 x 中的行进行置换，以获得置换后的 x^{per}，再使用第二阶段所用的相同非参数算法估计 $\mathcal{F}(x^{per})$

。使用于第二阶段。对于置换数据，两种特定损失函数 (8) 和 (13) 的表达式分别为 $\underset{\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{a}_0(x_i) - \mathcal{F}(x_i^{per})T_i\right\}^2$ 和 $\underset{\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left[\{1 - y_i - \hat{a}_0(x_i)\}\mathcal{F}(x_i^{per})t_i + y_i \exp\left\{-\mathcal{F}(x_i^{per})t_i\right\}\right]$；为简便起见，忽略加权因子 $\left\{\frac{T_i+1}{2}\Pr(\quad) - \frac{T_i-1}{2}\Pr(T_i=-1)\right\}$。请注意，在置换数据中，$Y$ 和 T 固定不变，ITT 效应也不变。下一步是通过转换 $\mathcal{F}(x^{per})$ 得到异质性治疗效应预测值 $\hat{}\tau(x^{per})$，并计算 $s^2\{\hat{}\tau(x^{per})\}$，然后重复置换步骤足够多次，以获得 $s^2\{\hat{}\tau(x^{per})\}$ 的置换分布。我们将 p 值定义为：从置换分布中估计得到的 $s^2\{\hat{}\tau(x^{per})\}$ 等于或大于观测到的 $s^2\{\hat{}\tau \text{obs}(x)\}$ 的次数所占比例。第一阶段估计的主效应分量与第二阶段的异质性治疗效应分量正交，因此，在给定估计的主效应分量的情况下，对后者进行假设检验是有效的。我们在补充材料 S6 中，通过各种模拟设置评估了所提出的两阶段梯度提升树置换检验。

2.2.4 极端梯度提升树

原则上，任何非参数学习算法都可以用于我们的增广学习程序。极端梯度提升树（XGBoost）尤其适用，因为它允许用户自定义损失函数 [Chen 和 Guestrin,2016]。令 $\mathbb{F}$ 为一组实值弱学习器函数（即基函数），并按照 Zhang等人 [2005], $\text{span}(\mathbb{F}) = \left\{\sum_{m=0}^{M} F_m \text{的类似记法：} F_m \in \mathbb{F},\ M \in Z^+\right\}$，它构成线性函数空间。

$F_m(X) \in \mathbb{F}$ 称为弱学习器，例如分类与回归树 [Breiman 等人， 2017]。假设 $F_m(X)$ 是一棵有 J个终端节点的回归树， $F_m(x) = \omega_{q(x)}$，其中 $q(x)$：$\mathbb{R}^p \to \{1,\ldots,J\}$ 表示树结构，并将每个样本映射到叶节点索引，J为叶节点数；而 ω_q：$\{1,\ldots,J\} \to \mathbb{R}$ 表示每个叶节点的权重，例如均方误差损失下该叶节点中的样本均值。XGBoost 的具体形式见补充材料 S3。每个受试者的最终输出是各加性回归树中对应权重之和。在满足某些正则条件时，通过分类与回归树进行经验损失最小化所得的解，将收敛到指定损失函数期望在 span(𝔽) [上的最小化解（ Jiang 等人，2004 ； Zhang 等人，2005 ； Bartlett 和 Traskin， 2007] ）。

我们简要介绍 XGBoost[Chen 和 Guestrin， 2016], 这是一种近期提出的 GBT 变体，具有正则化和更快的实现，用于优化每个加性分量 F_m。损失函数 L 通过关于 $\mathcal{F}(X)$ 的二阶泰勒展开进行近似，并加入正则化项。XGBoost 以逐阶段方式进行。当 $m = 0$时，令 $F_0(X)$ 表示初始猜测。当 $m = k > 0$时，对于第 k 个独立加性分量$F_k(X)$，在固定所有先前学得的 $(k-1)$ 分量的情况下，我们通过最小化以下损失函数来估计它，

$$\mathcal{L}^{(k)} \simeq \sum_{i=1}^{n}\left\{L\left(Y_i, \mathcal{F}^{(k-1)}(X_i)T_i\right) + G_i F_k(X_i) + \frac{1}{2}H_i F_k^2(X_i)\right\} + \gamma J_k \tag{14}$$

$$= \sum_{j=1}^{J_k}\left\{\Big(\sum_{i\in I_j} G_i\Big)\omega_j + \frac{1}{2}\Big(\sum_{i\in I_j} H_i\Big)\omega_j^2\right\} + \gamma J_k, \tag{15}$$

其中， $G_i = \partial L/\partial F_k|_{F_k=0}$ 和 $H_i = \partial^2 L/(\partial F_k)^2|_{F_k=0}$分别是第 i 个样本的损失函数的梯度函数和海森函数，$\mathcal{F}^{(k-1)}(X_i) = \sum_{m=0}^{k-1}\widehat{F}_m(X_i)$ J_k 表示第 k 棵回归树的叶节点数，而 γ 是正则化参数，用于惩罚树结构的复杂度（即进一步划分所需的最小损失降低）。对于第二个方程，我们通过去掉常数项 $L\left(Y_i, \mathcal{F}^{(k-1)}(X_i)T_i\right)$，写出在固定树结构 $q_k(X)$ 下的近似损失函数 (14)，其中 $I_j = \{i|q_k(x_i) = j\}$表示第 k 棵树中第 j 个叶节点对应的集合。

XGBoost 支持自定义损失函数，我们只需提供指定损失函数的梯度和海森函数。因此，它提供了一种现成的学习技术，用于实施我们的增强学习程序以估计 $\mathcal{F}(X)$。我们采用 XGBoost 以两阶段方式估计增广项 $a_0(X)$ 和异质性治疗效应函数 $\mathcal{F}(X)$，如附录中的算法1所示，称为两阶段梯度提升树（TSGBT）。需要指出的是，第一阶段增强项不一定要通过 GBT 估计——任何参数算法和非参数算法都应可行。在该算法中，令 M_a 和 M 分别表示两个阶段中加性回归树的总数（即迭代次数）。 G， G^a 和 H， H^a 是任意凸损失函数的一阶和二阶梯度统计量。令 J_{ak} 和 J_k 表示第 k 棵回归树的叶节点数。

首先，对于均方误差损失下的连续响应或逻辑似然损失下的二元响应，我们通过指定相应的梯度 G_i^a 和海森函数 H_i^a，应用 XGBoost 算法估计 $a_0(X)$。其次，我们通过其一阶和二阶梯度统计量，将第 k 个损失函数 $\mathcal{L}^{(k)}$ 与我们提出的增强损失函数联系起来，即第2.2.1节中的损失函数（8）和第2.2.2节中的损失函数（13），其各自的增广项 $a_0(X_i)$在第一阶段估计。为估计第 k 棵回归树，对于损失函数（8），$G_i = -2T_i\left[Y_i - \widehat{\mathbb{E}}(Y_i|X_i) - \mathcal{F}^{(k-1)}(X_i)T_i\right]$；或者对于损失函数（13），$-T_i\left[Y_i - 1 - 2\widehat{\mathbb{E}}(Y_i|X_i) + 1 + Y_i\exp\left(-\mathcal{F}^{(k-1)}(X_i)T_i\right)\right]$；对于损失函数（8），$H_i = 2$；或者对于损失函数（13），$Y_i\exp\left(-\mathcal{F}^{(k-1)}(X_i)T_i\right)$。当随机化比例不相等时，所有的 G_i 和 H_i 均应乘以一个权重因子 $\left\{{}_2{}^{T_i+1}_{T_i=1}\Pr(\quad) - {}_2{}^{T_i-1}\Pr(T_i=-1)\right\}$。经过$M$次迭代后，$\mathcal{F}^{(M)}(X)$即为 $\mathcal{F}(X)$的估计量。

提升算法可能会对数据过拟合，因此需要适当的正则化来确保其一致性[Bühlmann 和 Yu，2003年，Jiang等人，2004年，Zhang等人，2005年，Bühlmann，2006]。我们在每次迭代中，按照学习率（记作 η）缩减所学习到的加性分量 $A(X)$或 $F(X)$。迭代次数与学习率之间存在权衡关系。学习率较低的提升树学习速度较慢，因此需要更多迭代次数。不过，较低的学习率通常能确保更好的收敛性 [Zhang等人，2005]。超参数对梯度提升树的预测性能至关重要。除正则化参数 γ 和学习率 η外，XGBoost还有其他几个调参参数：迭代次数 M、树的最大深度（最大深度）、行采样率（子采样）、列采样率（列采样）、子叶节点所需的样本权重最小总和（即指定损失函数对每个样本的二阶导数）（子节点最小权重）。我们在补充材料 S2 中介绍这些参数的调优流程。

3 数值研究

本节在不同模拟情景下，评估所提出的两阶段方法在估计异质性治疗效应方面的表现，以及所提出的用于检验无交互效应的置换检验。为简洁起见，我们仅展示具有代表性的结果，其余结果见网络附录B，其中包括一项简单的原理验证模拟实验，用以展示两阶段程序的一致性和稳健性。

我们将所提出的两阶段梯度提升树（TSGBT）方法与完全回归法（FR）、修正协变量方法（MC）、带效率增广的修正协变量法（MC-aug）、分组梯度提升树和加权梯度提升树进行比较。完全回归法是一个多元线性模型，包含结局 Y、协变量 X、治疗指示变量 T，以及协变量与治疗指示变量之间的交互项 XT。修正协变量法是一个线性模型，包含结局 Y和修正协变量 XT [Tian等人，2014]。带效率增广的修正协变量法是一个线性模型，包含结局 Y、修正协变量 XT以及一个增广项。对于这些参数方法，即FR、MC和MC-aug，我们使用R软件包glmnet中的cv.glmnet函数，通过LASSO进行变量选择和正则化。分组梯度提升树（SGBT）方法是分别对两个处理组拟合GBT。加权梯度提升树是所提出的两阶段GBT方法中省去第一阶段后的第二阶段。我们使用XGBoost和caret软件包中的函数，并采用所提出的修正损失函数来构建梯度提升树。我们通过斯皮尔曼相关系数（sCORR）和均方误差（MSE）评估这些方法的表现。

3.1 连续型结局

假设 $Y = \sum_{j=1}^{p}\alpha_j X_j{}^2 + \mathcal{F}\ X\ \boldsymbol{\beta}\ \cdot T/2 + \sigma_0\epsilon$ = 300 $\mathcal{F}\ X\ \boldsymbol{\beta}$ $Y = \alpha_0 +$为连续 $X_1\ \ldots\ X_p\ \sim\mathcal{N}\ \mathbf{0}\ \Sigma$ 变量，我们从回归模型中生成 N 个独立高斯样本，( ) (;)，其中(;)可以是任意函数，协变量(, ,)(,)满足 Σ 为一阶自回归方差—协方差矩阵，其相关系数为 $\rho = 0.5$，$p = 50$或 1000，$\sigma_0 = 2$且 $\epsilon \sim \mathcal{N}(0,1)$。处理指示变量 T以 1/2 的概率随机分配给 $T = 1$和 $T = -1$。我们首先考察具有线性交互效应的情形，$\mathcal{F}(X;\boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{p}\beta_j X_j$。我们设计了四种参数设置，具体如下。在设置1中，$(\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\ldots,\alpha_p)$ =(0.4, 0.6,- 0.6, 0.6, 0.6, 0, . . . , 0)，$(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\ldots,\beta_p)$ =(0.8, 0.8,-0.8, 0.8, 0.8, 0, . . . , 0)。我们的目标是评估所提出的树提升方法在简单线性模型中的表现；在这类模型中，其他方法也能够捕捉真实模型。随后，我们考察另外三种具有更复杂交互效应的设定。令 $\alpha_0 = (\sqrt{3})^{-1}$，$\alpha_j = (2\sqrt{3})^{-1}$，$j = 3,\ 4,\ldots,10$，其余均为零，且 $\mathcal{F}(X;\boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{p}\beta_j(X_j + X_j^2) + \sum_{1\le i<j\le p}\beta_{ij}X_iX_j$。在设置2中，$(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\ldots,\beta_p)$ =(0.8, 1.6,-1.6, 1.6,-1.6, 0, . . . , 0) $\beta_{12} = 1.6$，$\beta_{15} = 1.6$。在设-

置3中，$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \ldots, \beta_p)$ =(0.8, 0.8,-0.8, 0.8,-0.8, 0, .. . , 0) $\beta_{12} = 0.8$， $\beta_{15} = 0.8$。在设定 4 中，$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \ldots, \beta_p)$ =(0.8, 0, 0, 0, 0, 0, .. . , 0) $\beta_{12} = 0$， $\beta_{15} = 0$。设置2和设置3旨在展示提升树算法在交互效应强和中等时的表现。设置4用于评估在不存在交互效应时这些不同方法的表现。我们在四种设定下比较了所有方法，其中 p=50 或 1000。

真实的异质性治疗效应为， $\tau(X) = \mathrm{E}[Y|X, T = 1] - \mathrm{E}[Y|X, T = -1] = \mathcal{F}(X; \boldsymbol{\beta})$。每次模拟中，我们生成了 1000 个独立的测试数据样本，以评估这些方法的表现。图 1 中，最上方的三个面板分别是设置1、设置2和设置3的斯皮尔曼相关系数箱线图；中间的三个面板分别是设置1、设置2和设置3的均方误差图；面板 (h) 是设定 4 下的均方误差图；面板 (i) 和 (g) 分别是在 p = 50 和 1000 的情况下，用于调整最佳迭代次数的诊断图。灰色箱线图对应低维设定，白色箱线图对应高维设定。

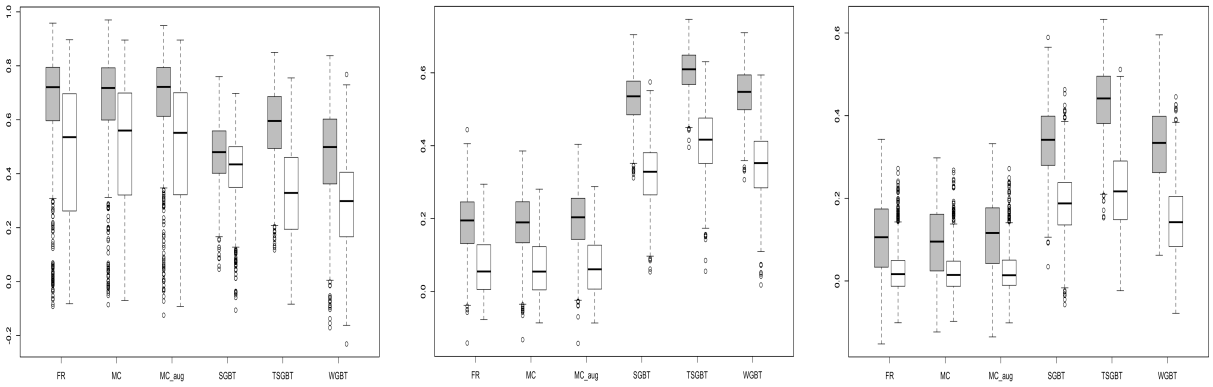
图 1 中，TSGBT 在所有设定下的表现均优于 WGBT，证实第一阶段增强提高了 HTE 估计的效率。当协变量数量从 50 增加到 1000 时，几乎所有 sCORR 都会降低，均方误差则会增大，因为高维协变量引入了更多噪声，使估计变得更加困难。图 1(a) 显示，当真实模型为线性模型时，在估计治疗效应的排序方面，所有线性方法（FR、MC、MC-aug）的表现都优于提升方法。就均方误差而言，所有提升方法的表现与线性方法相当，甚至更好（图 1(d)）。对于设置2和设置3，图 1(b)、1(c)、1(e) 和 1(f) 显示，提升方法的表现优于线性方法，因为灵活的树结构能够更好地刻画更复杂的非线性交互效应；使用 LASSO 的线性模型无法成功筛选出协变量。在设置2和设置3的低维及高维情况下，TSGBT 的表现都优于 SGBT，因为 TSGBT 利用全部数据估计两组的主效应和交互效应，而 SGBT 将数据划分给两个治疗组，导致效率损失。例如，在设置2下，当 p = 50(1000) 时，FR、MC、MC-aug、SGBT 和 TSGBT 的 sCORR 中位数分别为 0.20(0.06)、0.19(0.06)、0.20(0.06)、0.54(0.33) 和 0.61(0.42)、0.55(0.35)；在设置2下，当 p = 50(1000) 时，FR、MC、MC-aug、SGBT 和 TSGBT 的均方误差中位数分别为 26.09(27.82)、26.51(28.08)、26.15(27.83)、19.21(22.97)、16.08(20.52) 和 17.50(21.48)。

图1(h)和图1(i)显示，当完全不存在交互效应时，GBT诊断图可以立即揭示零异质性治疗效应。这些诊断图体现了TSGBT相较于SGBT的优势：前者能够利用GBT机制评估信噪比，而后者则不能。我们还比较了相关性为零以及协方差矩阵为复合对称矩阵且 $\rho = 1/3$ 时，这些方法在设置2和设置3下的表现（结果见网络附录 B）。结论类似。

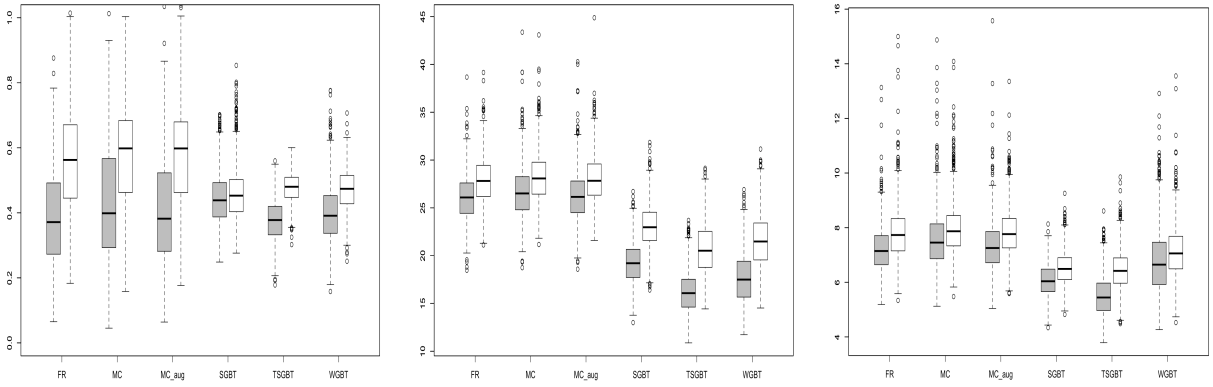
3.2 二元结局

当 Y 为二元变量时，我们根据相对风险指数模型 $\log \frac{P_1}{P_{-1}} = 2\mathcal{F}(X; \boldsymbol{\beta})$ 及其针对优势乘积的干扰模型 $\log \frac{P_1 P_{-1}}{(1-P_1)(1-P_{-1})} = -C - (\alpha_0 + \sum_{j=1}^{p} \alpha_j X_j)$[Richardson et al., 2017], 生成N个 = ~~1500 独立二元~~样本，其中 $P_1 = \mathbb{E}[Y|X, T = 1]$ 和 $P_{-1} = \mathbb{E}[Y|X, T = -1]$ 均由指定的主要模型和干扰模型推导得出，而 C 的取值则确保患病率约为0.2至0.3，且非负指数的比例小于0.01。二元结局的其他模拟方法见网络附录B。我们考虑复杂交互效应情形下的以下三种模拟设置。具体而言，设置1、2和3分别用于说明提升树算法在交互效应强、中等和不存在时的表现。令$(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_4, \ldots, \alpha_p)$ =(0.4, 0.6,-0.6, 0.6, 0.6, 0, . . . , 0)，且 $\mathcal{F}(X; \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{p}(\beta_j X_j + \gamma_j X_j^2) + \sum_{1 \le i < j \le p} \beta_{ij} X_i X_j$. 。在设置1中，$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \ldots, \beta_p)$ =(0.3, 0.3, 0.4, 0.3, 0.4, 0, . . . , 0)，$(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \ldots, \gamma_p)$ =(0.4,- 0.4, 0.4, 0.4, 0, . . . , 0) $\beta_{12} = 0.5, \beta_{15} = 0.5$, $C = 2.5$；在设置 2 中，$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \ldots, \beta_p)$ =(0.1, 0.1, 0.2, 0.1, 0.2, 0, . . . , 0)，$(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \ldots, \gamma_p)$ =(0.2,-0.2, 0.2, 0.2, 0, . . . , 0) $\beta_{12} = 0.5, \beta_{15} = 0.5$, $C = 2.5$；在设置3中，$(\beta_0, \beta_1, \ldots, \beta_p)$ =(0.3, 0, . . . , 0)，$(\gamma_1, \ldots, \gamma_p)$ =(0, . . . , 0) $\beta_{12} = 0, \beta_{15} = 0, C = 2$。

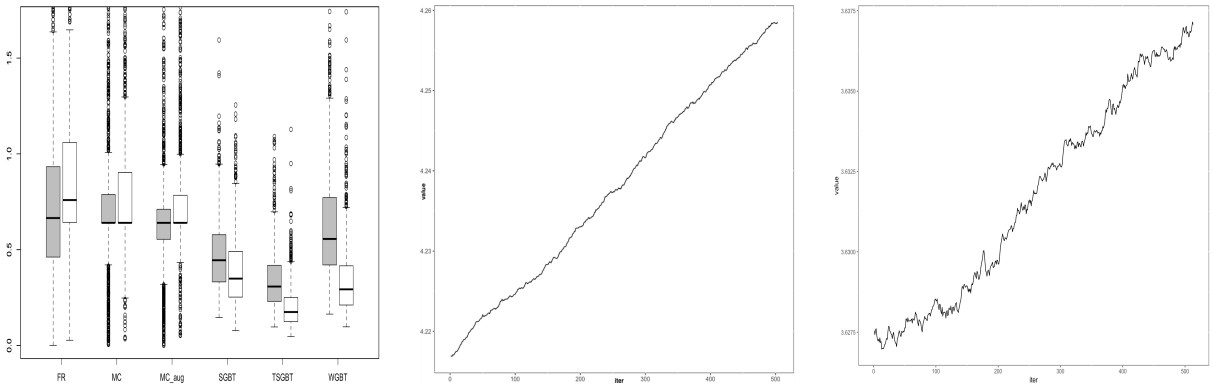
在三种设置下，我们将六种方法应用于生成的二元数据，其中 p=50 为 1000。真实的异质性治疗效应定义在风险比尺度上， $\tau(X) = \frac{\mathbb{E}[Y|X,T=1]}{\mathbb{E}[Y|X,T=-1]}$。我们通过 sCORR 和均方误差（MSE）比较估计得分与真实值~~, 其中 MSE 在~~对数风险比尺度上计算。我们生成了 1000 组相互独立的测试数据，以评估这些方法的性能。图2中的结果与连续型结局的结果一致。在设置2中，当 p = 1000 时，TSGBT 的 sCORR 表现与完全回归法相近。总体而言，如图2(a)、2(b)、2(c) 和 2(d) 所示，TSGBT 在 sCORR 和 MSE 两项指标上的表现均优于其他所有方法。例如，在 p = 50(1000) 下，设置1中完全回归法、修正协变量法、MC-aug、SGBT、TSGBT 和 GBT 的 sCORR 中位数分别为 0.67(0.64)、0.66(0.63)、0.66(0.63)、0.74(0.61)、0.75(0.68) 和 0.71(0.49)；设置2中 MSE 的中位数分别为 3.03(3.41)、2.93(3.33)、3.12(3.92)、3.59(4.08)、2.20(2.77)、2.34(3.41)，对应于



(a) sCORR，设置1 (b) sCORR，设置2 (c) sCORR，设定 3



(d) 均方误差，设置1 (e) 均方误差，设置2 (f) 均方误差，设定 3



(g) 均方误差，设定 4 (h) TSGBT 的诊断图，设定 4，p = 50 (i) TSGBT 的诊断图，设定 4，p = 1000

图 1：使用六种不同方法（FR，完全回归法；MC，修正协变量法；MC-aug，带效率增强的修正协变量法；SGBT，独立梯度提升树；WGBT，加权梯度提升树；TSGBT，两阶段梯度提升树）处理具有高斯型结局的模拟数据时，估计的异质性治疗效应与真实效应之间的秩相关系数和均方误差的箱线图。灰色和白色箱体分别表示低维和高维（p=50 和 1000）设置。设置1–3的结果如 (a)–(f) 所示，设定 4 的结果如 (g)–(i) 所示。(h) 和 (i) 展示了第二阶段梯度提升树用于选择最佳迭代次数的诊断图，其中内部测试数据的均方根误差值为纵轴。

表 1：在 $p = 50$和 $p = 1000$ 下置换检验的 I类错误

	显著性水平	p=50			p=1000		
		P1	P2	P3	P1	P2	P3
连续型	0.01	0.007	0.007	0.011	0.011	0.010	0.013
	0.05	0.051	0.051	0.055	0.044	0.055	0.053
	0.10	0.100	0.102	0.104	0.094	0.112	0.108
二元	0.01	0.009	0.009	0.007	0.008	0.012	0.011
	0.05	0.047	0.054	0.046	0.041	0.049	0.053
	0.10	0.098	0.107	0.087	0.083	0.098	0.103

在 p = 50(1000) 下，完全回归法、修正协变量法、MC-aug、SGBT 和 TSGBT 分别为。在图 2(e) 中，在无交互作用情形下，TSGBT 的均方误差几乎优于所有其他方法；但当 p = 50 时，完全回归法与 TSGBT 的表现相近。接下来，我们在第 3.3 节中进一步探讨了主效应非零且交互作用信号为零的其他无效应情形。

3.3 置换检验

我们进行了模拟，以评估所提出的置换检验在控制假阳性率方面的性能。对于一个模拟数据集，我们固定第一阶段的估计效应，并按照第 2.2.3 节中的方法，在置换后的协变量上实施第二阶段。我们重复置换 1000 次，以获得训练样本异质性治疗效应变异性的置换分布，并计算 p 值，即变异性大于或等于观测变异性的置换所占比例。为评估 I类错误，我们从不含交互效应的模型中生成了 1000 个模拟数据集。

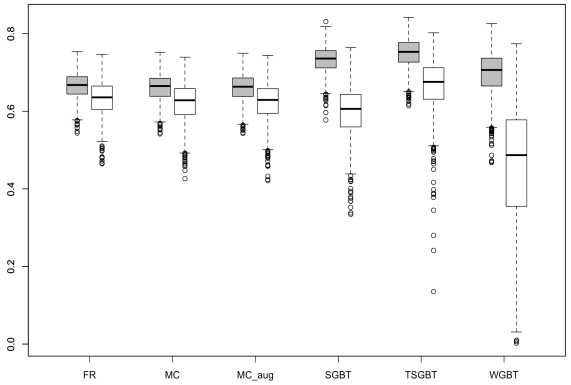
我们根据模型生成了 N=300， $Y = (\alpha_0 + \sum_{j=1}^{p} \alpha_j X_j)^2 + \beta_0 \cdot T + \sigma_0 \epsilon$，其中 $\beta_0 = 0.8$、 $\sigma_0 = 2$ 和 $\epsilon \sim \mathcal{N}(0, 1)$，并根据相对风险的主要指数模型生成了 N=1500，$\log \frac{P_1}{P_{-1}} = \beta_0$，同时采用其优势乘积干扰模型 $\log \frac{P_1 P_{-1}}{(1-P_1)(1-P_{-1})} = -C - (\alpha_0 + \sum_{j=1}^{p} \alpha_j X_j)^2$ [Richardson 等人， 2017]，，其中 $P_1 = \mathbb{E}[Y|X, T = 1]$ 和 $P_{-1} = \mathbb{E}[Y|X, T = -1]$ 均由指定的主要模型和干扰模型推导得出，C 的选取方式与第 3.2 节类似，且 $\beta_0 = 0.3$。我们改变主效应参数，同时保持其他参数不变。我们考虑主效应分别为强、弱和零的三种情形：在 P1 中， $\alpha_0 = (\sqrt{3})^{-1}$， $\alpha_j = (2\sqrt{3})^{-1}$， $j = 3, 4, \ldots, 10$，其他 α 均为零， $\rho = 0.5$；在 P2 中， $\alpha_0 = (\sqrt{6})^{-1}$， $\alpha_j = (2\sqrt{6})^{-1}$， $j = 3, 4, \ldots, 10$，其他 α 均为零， $\rho = 0.5$；在 P3 中， $\boldsymbol{\alpha} = (0, 0, \ldots, 0)$， $\rho = 0.5$。如表 1 所示，在主效应强、弱或为零的所有情形下，所提出的 TSGBT 置换检验都能将假阳性率控制在正确水平。

4 PCPT 数据分析

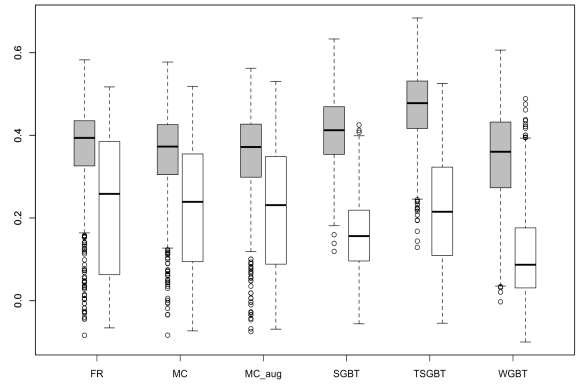
前列腺癌预防试验（PCPT）由 SWOG 开展，旨在检验非那雄胺能否预防前列腺癌（PCa）（ClinicalTrials.gov NCT00288106）。患者以相同概率被随机分配至安慰剂组或非那雄胺组 [Thompson 等人， 2003]。主要终点是在 7 年期间经活检检出的前列腺癌，这些男性被定义为病例。对照组定义为在研究结束时完成前列腺活检且未发现前列腺癌的男性。研究人员开展了一项病例对照研究，以探讨遗传因素与前列腺癌风险之间的关系；对照组按治疗组、年龄以及一级亲属患前列腺癌的阳性家族史分布与病例进行频数匹配抽样 [Goodman 等人， 2010]。

在我们的分析中，我们重点关注临床意义重大的终点——高级别前列腺癌风险，并从各项研究中提取了 306 名高级别病例和 1,365 名对照组的基因型数据；这些受试者均有足够的白细胞 DNA 可用于基因分型。相关研究及基因分型详情此前已发表 [Price 等人，2016年，Chen 等人，2016年，Chu 等人， 2018]。基因型数据共包含来自 85 个候选基因的 444 个单核苷酸多态性，其中 44 个单核苷酸多态性因未通过哈代 - 温伯格检验的质量控制指标而被剔除。非那雄胺的异质性治疗效应定义为某一基因型亚组发生高级别前列腺癌的风险比。我们的目标是利用这些基因型开发用于个体化治疗效应的预测特征。我们采用两阶段梯度提升树方法分析病例对照抽样权重。对照组的抽样权重为 $W_i = 7$， $110/1$， $365 \approx 5.21$，病例的抽样权重为 $W_i = 1$。我们将逆抽样概率 W_i 纳入一阶和二阶梯度统计量，定义见 G^a， H^a 以及第 2.2.4 节中的 G， H。

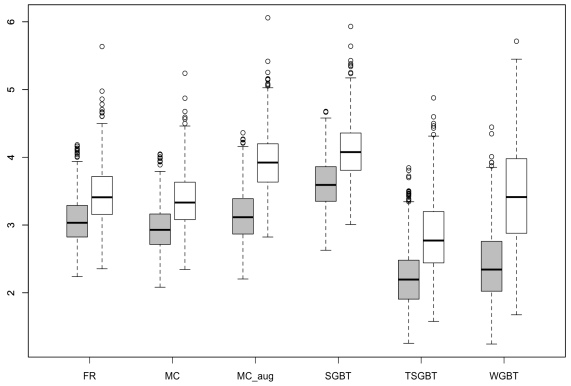
在第一阶段，我们从常用范围中选择了超参数值，其中 *max-depth* = 6, $\eta = 0.1$,$\gamma = 4$, *colsample* = 0.7, *subsample* = 0.6, *min-child-weight* = 2. 迭代次数 M_a 通过十折交叉验证并采用提前停止进行调优，最终确定为 43；其诊断图见图3(a)，显示主效应较弱



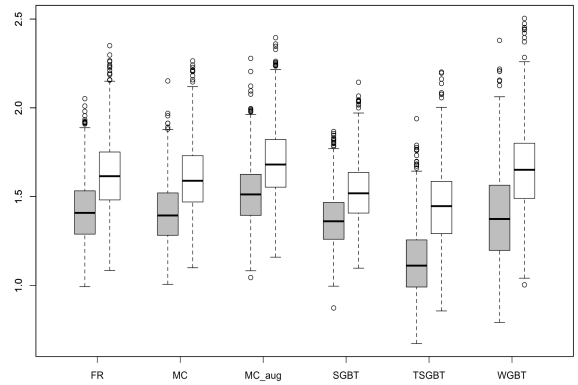
(a) sCORR，设置1



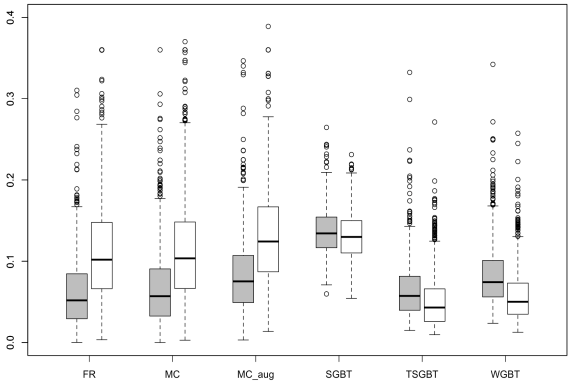
(b) sCORR，设置2



(c) 均方误差，设置1



(d) 均方误差，设置2



(e) 均方误差，设定 3

图2：在二元结局中，采用六种不同方法（FR，完全回归法；MC，修正协变量法；MC-aug，带效率增强的修正协变量法；SGBT，独立梯度提升树；WGBT，加权梯度提升树；TSGBT，两阶段梯度提升树）时，设定 1、2 和 3 下估计的异质性治疗效应与真实效应之间的秩相关系数和均方误差的箱线图。灰色和白色箱线图分别表示低维（p=50 ）和高维（1000）情形。

在前列腺癌中单核苷酸多态性的影响。在第二阶段，我们通过连续十折交叉验证调优了这些超参数，具体为 $\eta = 0.01$、最大深度 = 4、列采样 = 0.7、子采样 = 0.6，并选择较大的 $\gamma = 8$、最小子节点权重= 12 以保持保守。随后，我们通过结合提前停止的十折交叉验证，将 M 调优至391。图3(b)中的诊断图表明，这些单核苷酸多态性可能与非那雄胺相互作用，从而影响高级别前列腺癌的风险。图3(c)显示，所提出的TSGBT方法将14.7%的男性分类为服用非那雄胺后高级别前列腺癌风险降低的人群。我们在图3(d)中展示了按相对变量重要性排序的前20个遗传变异。图3(e)展示了所有选定的350个单核苷酸多态性的原始方差重要性增益图，顺序从最大到最小。补充材料 S3介绍了XGBoost中变量重要性的推导过程。前100个单核苷酸多态性的增益迅速衰减，后续单核苷酸多态性的贡献极小。

使用全部数据进行估计可能会导致过拟合。为缓解这一潜在问题，我们将数据划分为90%的训练集和10%的测试集，并将训练好的模型应用于测试数据，以估计相应比例。我们多次重复这一交叉验证程序，发现服用非那雄胺后高级别前列腺癌发病风险降低的患者比例，与使用整个数据集进行估计所得的比例一致。

我们实施了所提出的个体化治疗效应不存在异质性的置换检验。我们使用中位数绝对偏差（MAD）衡量风险比尺度上估计的个体化治疗效应的稳健方差。置换检验所得的p值为0.051，表明非那雄胺预防高级别前列腺癌风险的治疗效应存在中等程度的显著异质性。

5 讨论

对于连续型和二元型试验终点，我们提出了一种灵活的两阶段统计学习程序，用于估计异质性治疗效应（HTE），并通过增广提高效率。该程序允许在针对HTE进行一致的非参数学习时，采用为任意交互模型推导出的最优增广项。因此，我们提出了一种简单且高效的两阶段梯度提升树算法来估计HTE。由于两阶段估计量具有正交性（这是由随机治疗的结构所决定的），该算法对第一阶段模型设定错误具有稳健性。这一性质使得我们能够在给定增广项的条件下，开发一种置换检验，以全局评估是否存在HTE的证据。

我们研究的一个自然延伸方向是生存结局；当存在删失且随访期间发生事件时，问题会更加复杂。此时，增广项不再具有原先的简洁性，因此，要像我们为线性结局和二元结局所构建的那样实施两阶段程序，将是一项颇具挑战的任务。另一个延伸方向是观察性研究，在这类研究中，我们需要用倾向评分替代随机分配概率，并估计倾向评分或进行结局加权，以实现双重稳健性质 [Nie and Wager, 2021]。在强可忽略性假设（即给定已观测协变量后不存在未测量混杂因素）下，我们提出的两阶段提升程序可以轻松适用于观察性研究。

致谢

作者谨 向 Cathy Tangen 致谢，感谢她提供 PCPT遗传数据，并感谢 Lu Tian 提供宝贵意见ts.

经费

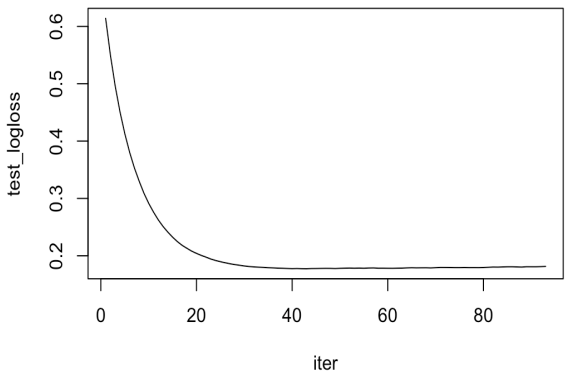
本研究得到了美国国立卫生研究院资助（项目编号 R01 CA222833）。我们衷心感谢美国国立卫生研究院项目（编号 S10OD028685）提供的计算支持。

数据可用性声明

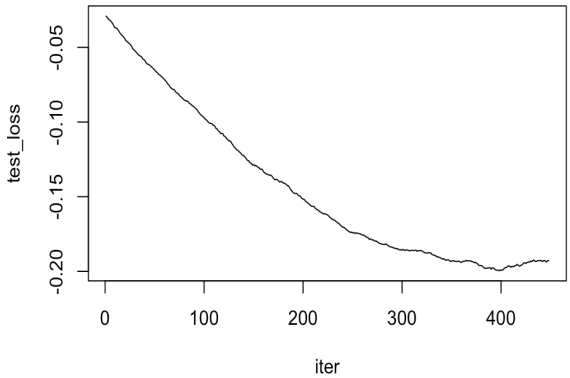
支持本研究发现的数据可向通讯作者索取。由于隐私或伦理方面的限制，这些数据不公开。

补充信息

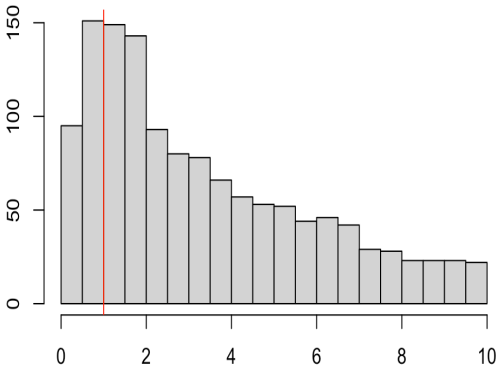
更多补充信息可在本文末尾的“补充信息”部分在线查阅。补充信息简要概述了修正协变量方法（补充材料 S1）、XGBoost 参数的调优过程（补充材料 S2）以及变量重要性的表达式（补充材料 S3），



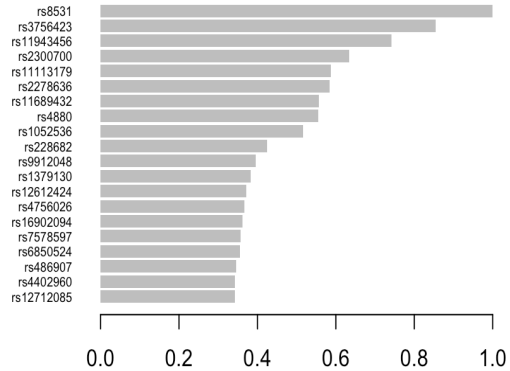
(a) 主效应诊断图



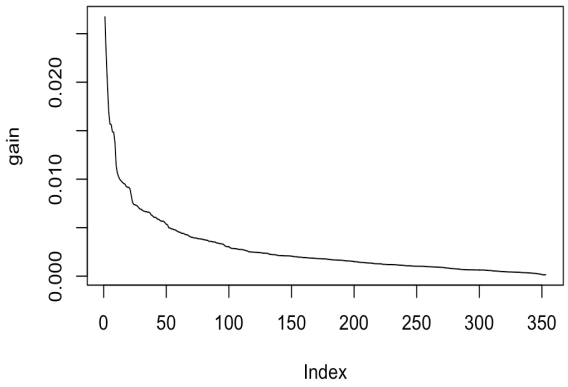
(b) 异质性治疗效应诊断图



(c) 个体化治疗效应



(d) 按变量重要性排序的前 20 个变量



(e) 变量重要性增益

图3：基于这400个单核苷酸多态性（SNPs）评估高级别前列腺癌发病风险的主效应/XGBoost模型中迭代次数与测试对数损失/损失之间的诊断图；基于这400个SNPs，针对这306例高级别病例和1,365名对照组绘制的风险比异质性治疗效应直方图；对估算异质性治疗效应贡献最大的前20个SNPs的相对变量重要性图；以及全部400个SNPs的变量重要性损失增益图。

增广估计量与最大似然估计量之间的联系（补充材料 S4）、展示效率和稳健性的参数模拟（补充材料 S5），以及更多不同场景下的额外模拟结果（补充材料 S6-7）。

第 2.2 节最优增广的推导

为了推导出使 $\mathcal{F}(X)$ 对应的估计量方差最小的最优增广项，只需推导出使增广估计函数 $S^{\text{aug}}(Y,\ \mathcal{F}(X)T)$ 方差最小的增广项，如公式（6）所示。换言之，$a_0(X)$ 应使下式最小:

$$\mathbb{E}\left[\left\{S^{\text{aug}}(Y,\mathcal{F}(X)T)\right\}^{\otimes 2}\right],$$

对 $a(X)$ 求一阶导数并令其等于零，即可通过求解下列方程得到 $a_0(X)$:

$$\mathbb{E}\Bigg[\left\{S(Y,\mathcal{F}(X)T)-a(X)\left(\frac{T+1}{2\text{Pr}(T=1)}+\frac{T-1}{2\text{Pr}(T=-1)}\right)\right\}\times \left(\frac{T+1}{2\text{Pr}(T=1)}+\frac{T-1}{2\text{Pr}(T=-1)}\right)\Bigg|X\Bigg]=0.$$

对治疗的随机化分布进行积分后，我们便可得到最优增广项的一般形式

$$a_0(X)=\left[\mathbb{E}\left\{S(Y,\mathcal{F}(X))\Big|X,T=1\right\}-\mathbb{E}\left\{S(Y,-\mathcal{F}(X))\Big|X,T=-1\right\}\right]\times \text{Pr}(T=1)\text{Pr}(T=-1).$$

随后，我们将增广项的一般形式应用于均值差估计目标和风险比估计目标。

A.1 均值差估计目标的最优增广

For 对于均值差估计目标，我们采用第 2.2 节中的估计函数 $S(Y,\ \mathcal{F}(X)T)$。 1,

$$S(Y,\mathcal{F}(X)T)=\left\{\frac{T+1}{2\text{Pr}(T=1)}+\frac{T-1}{2\text{Pr}(T=-1)}\right\}\left(Y-\mathcal{F}(X)T\right)$$

将附录中推导出的一般形式 $a_0(X)$ 代入，即可得到均值差估计目标在平方误差损失函数下的最优增广项,

$$a_0(X)=\left\{\mathbb{E}(Y|X,T=1)\text{Pr}(T=-1)+\mathbb{E}(Y|X,T=-1)\text{Pr}(T=1)\right\} -\mathcal{F}(X)\left\{\text{Pr}(T=-1)-\text{Pr}(T=1)\right\}$$

如第2.2.1节所示，$\mathcal{F}(X)$ 定义为 $\frac{1}{2}\left\{\mathbb{E}(Y|X,\ T=1)-\mathbb{E}(Y|X,\ T=-1)\right\}$。如果 $\mathcal{F}(X)$ 被正确设定且估计一致，最优增广项则变为 $a_0(X)=\frac{1}{2}\left\{\mathbb{E}(Y|X,\ T=1)+\mathbb{E}(Y|X,\ T=-1)\right\}$。一种特殊情况是 $\text{Pr}(T=1)=1/2$，$a_0(X)=\mathbb{E}(Y|X)$。

A.2 风险比估计目标的最优增广

我们将第2.2.2节中的估计函数应用于,

$$S(Y,\mathcal{F}(X)T)=\left\{\frac{T+1}{2\text{Pr}(T=1)}+\frac{T-1}{2\text{Pr}(T=-1)}\right\}\left\{1-Y-Y\exp\left(-\mathcal{F}(X)T\right)\right\}$$

转换为一般形式 $a_0(X)$ 在附录A中推导得出，用于获得风险比估计目标的最优增广项 ,

$$a_0(X)=1-\left[1+\exp\left\{-\mathcal{F}(X)\right\}\right]\mathbb{E}(Y|X,T=1)\text{Pr}(T=-1) -\left[1+\exp\left\{\mathcal{F}(X)\right\}\right]\mathbb{E}(Y|X,T=-1)\text{Pr}(T=1)$$

如第2.2.2节所示， $\mathcal{F}(X)$ 定义为 $\log\left\{\frac{\mathbb{E}}{\mathbb{E}}(Y|X,\ T=1)\,(Y|X,\ T=-1)\right\}$。当方程（11）中的 $\mathcal{F}(X)$ 得到正确设定并被一致估计时，最优增广项变为 $a_0(X) = 1 - \overline{\mathbb{E}(Y|X,\ T=1)} - \mathbb{E}(Y|X,\ T=-1)$，其中当 $\Pr(T=1) = 1/2$ 时， $a_0(X) = 1 - 2\mathbb{E}(Y|X)$。

B 第 2.2.4 节中的两阶段GBT算法

算法1： 两阶段梯度提升树

第一阶段;

为 $i = 1$设置初始值 $k = 1$和 $\mathcal{A}_i^{(0)}$， ....., n;

当 $1 \le k \le M_a$ **执行时** $\sum_{i=1}^{n}\left\{G^a\left[y_i, \mathcal{A}_i^{(k-1)}\right]\mathrm{A}_k(x_i) + \frac{1}{2}H^a\left[y_i, \mathcal{A}_i^{(k-1)}\right]\mathrm{A}_k^2(x_i)\right\}/\Pr(T=t_i) + \gamma J_{ak}$;

$\hat{\mathrm{A}}_k = \underset{\mathrm{A}_k \in \text{空间 an}(\mathbb{F})}{\text{使.....最小}} \frac{1}{n}$ –

$\mathcal{A}_i^{(k)} = \mathcal{A}_i^{(k-1)} + \hat{}$ <style id="4">$A_k(x_i)$ 对于 $i = 1, \ldots, n$;

end

$\widehat{\mathcal{A}}(x_i) = \mathcal{A}^{(M_a)\,i}$， 并变换 $\widehat{\mathcal{A}}(x_i)$， 以获得 $i = 1, \ldots, n$ 的 $\hat{a}_0(x_i)$;

第二阶段;

重置初始值 $k = 1$和 $\mathcal{F}$ $_i^{(0)}$ 对于 $i = 1$、...、 n;

当 $1 \le k \le M$ **时执行**

$\hat{\mathrm{F}}_k = \underset{\mathrm{F}_k \in \mathrm{span}(\mathbb{F})}{\mathrm{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left\{G\left[y_i, \left(\mathcal{F}_i^{(k-1)}\right)t_i\right]\mathrm{F}_k(x_i) + \frac{1}{2}H\left[y_i, \left(\mathcal{F}_i^{(k-1)}\right)t_i\right]\mathrm{F}_k^2(x_i) - \hat{a}_0(x_i)\mathrm{F}_k(x_i)t_i\right\}/\Pr(T = t_i) + \gamma J_k$;

$\mathcal{F}_i^{(k)} = \mathcal{F}_i^{(k-1)} + \hat{\mathrm{F}}_k(x_i)$ for $i = 1, \ldots, n$;

end

获取 $\widehat{\mathcal{F}}(x_i) = \mathcal{F}_i^{(M)}$， 其中 $i = 1$、...、 n;

结果： 异质性治疗效应是以下变量的函数 $\widehat{\mathcal{F}}(x_i)$ 对于 $i = 1, \ldots, n$

参考文献

William E Evans 和 Howard L McLeod。药物基因组学——药物处置、药物靶点与副作用。《新英格兰医学杂志》，348(6):538–549，2003年。Richard Weinshilboum 和 Liewei Wang。药物基因组学：从实验室到临床。《自然·药物发现评论》，3(9): 739–748，2004。Richard Simon 和 Sameek Roychowdhury。在临床试验中实施个体化癌症基因组学。《自然·药物发现评论》，12(5):358–369，2013。Tyler J VanderWeele、Bhramar Mukherjee 和 Jinbo Chen。未测量混杂情况下交互作用的敏感性分析。《医学统计学》，31(22):2552–2564，2012。David R Cox。交互作用。《国际统计评论/*Revue Internationale de Statistique*》，第1–24页，1984。Julio L Peixoto。多项式回归模型中的层次变量选择。《美国统计学家》，41(4): 311–313，1987。Nam Hee Choi、William Li 和 Ji Zhu。具有强遗传性约束及其 Oracle 性质的变量选择。《美国统计学会会刊》，105(489):354–364，2010。Jacob Bien、Jonathan Taylor 和 Robert Tibshirani。用于层次交互作用的套索。《统计学年鉴》，41(3): 1111，2013。Michael Lim 和 Trevor Hastie。通过层次组套索正则化学习交互作用。《计算与图形统计学杂志》，24(3):627–654，2015。Xiaogang Su、Chih-Ling Tsai、Hansheng Wang、David M Nickerson 和 Bogong Li。通过递归分割进行亚组分析。《机器学习研究杂志》，10(2009):141–158，2009。Susan Athey 和 Guido Imbens。针对异质性因果效应的递归分割。《美国国家科学院院刊》，113(27):7353–7360，2016。

Stefan Wager 和 Susan Athey。使用随机森林估计并推断异质性治疗效应。《美国统计学会会刊》，113(523):1228–1242，2018。James Y Dai 和 Michael LeBlanc。病例专属树和随机森林：在具有二元终点的随机临床试验中探索基因型特异性治疗效应。《英国皇家统计学会学报：*C*辑（应用统计）》，68(5):1371–1391，2019。James Y Dai、Michael LeBlanc、Phyllis J Goodman、M Scott Lucia、Ian M Thompson 和 Catherine M Tangen。病例专属方法识别出可预测一组男性非那雄胺治疗后高级别前列腺癌风险降低的遗传位点。《癌症预防研究》，12(2):113–120，2019。Lu Tian、Ash A Alizadeh、Andrew J Gentles 和 Robert Tibshirani。一种估计治疗与大量协变量之间交互作用的简单方法。《美国统计学会会刊》，109(508): 1517–1532，2014。Shonosuke Sugasawa 和 Hisashi Noma。使用梯度提升树估计个体治疗效应。《医学统计学》，38(26):5146–5159，2019。Anastasios Tsiatis。半参数理论与缺失数据。Springer Science & Business Media，2007。Anastasios A Tsiatis、Marie Davidian、Min Zhang 和 Xiaomin Lu。随机临床试验中两组治疗比较的协变量调整：一种有原则且灵活的方法。《医学统计学》，27(23): 4658–4677，2008。Min Zhang、Anastasios A Tsiatis 和 Marie Davidian。使用辅助协变量提高随机临床试验推断的效率。《生物统计学》，64(3):707–715，2008。Shuai Chen、Lu Tian、Tianxi Cai 和 Menggang Yu。用于亚组识别和比较治疗评分的通用统计框架。《生物统计学》，73(4):1199–1209，2017。Jared D Huling 和 Menggang Yu。使用 personalized 软件包进行亚组识别。《统计软件杂志》，98，2021年。Jerome H Friedman。贪婪函数逼近：梯度提升机。《统计学年鉴》，第1189–1232页，2001年。Peter Bühlmann 和 Bin Yu。采用 l 2 损失的提升法：回归与分类。《美国统计学会会刊》，98(462):324–339，2003年。Peter Bühlmann。用于高维线性模型的提升法。《统计学年鉴》，34(2):559–583，2006年。Tianqi Chen 和 Carlos Guestrin。Xgboost：一种可扩展的树提升系统。《第*22*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集》，第785–794页，2016年。Jersey Neyman。概率论在农业试验中的应用：原理探讨。《农业科学年鉴》，10:1–51，1923年。Donald B Rubin。估计随机和非随机研究中的治疗因果效应。《教育心理学杂志》，66(5):688，1974年。James M Robins、Andrea Rotnitzky 和 Lue Ping Zhao。当部分回归变量并非总能观测到时回归系数的估计。《美国统计学会会刊》，89(427):846–866，1994年。Jared C Foster、Bin Nan、Lei Shen、Niko Kaciroti 和 Jeremy MG Taylor。用于治疗与协变量交互作用及亚组识别的置换检验。《生物科学统计学》，8(1):77–98，2016年。Jonathan Levy、Mark van der Laan、Alan Hubbard 和 Romain Pirracchio。治疗效应异质性的一个基本度量。《因果推断杂志》，9(1):83–108，2021年。Tong Zhang、Bin Yu 等。带提前停止的提升法：收敛性与一致性。《统计学年鉴》，33(4): 1538–1579，2005年。Leo Breiman、Jerome H Friedman、Richard A Olshen 和 Charles J Stone。分类与回归树。Routledge，2017年。Wenxin Jiang 等人。AdaBoost 的过程一致性。《统计学年鉴》，32(1):13–29，2004年。Peter L Bartlett 和 Mikhail Traskin。AdaBoost 具有一致性。《机器学习研究杂志》，8(2007): 2347–2368，2007年。Thomas S Richardson、James M Robins 和 Linbo Wang。相对风险和风险差的建模与估计。《美国统计学会会刊》，112(519):1121–1130，2017年。

Ian M Thompson, Phyllis J Goodman, Catherine M Tangen, M Scott Lucia, Gary J Miller, Leslie G Ford, Michael M Lieber, R Duane Cespedes, James N Atkins, Scott M Lippman 等。非那雄胺对前列腺癌发生的影响。《新英格兰医学杂志》，349(3):215–224，2003年。Phyllis J Goodman, Catherine M Tangen, Alan R Kristal, Ian M Thompson, M Scott Lucia, Elizabeth A Platz, William D Figg, Ashraful Hoque, Ann Hsing, Marian L Neuhouser 等。临床试验向转化研究的过渡：前列腺癌预防试验的经验。《癌症预防研究》，3(12):1523–1533，2010年。Douglas K Price, Cindy H Chau, Cathee Till, Phyllis J Goodman, Robin J Leach, Teresa L Johnson-Pais, Ann W Hsing, Ashraful Hoque, Howard L Parnes, Jeannette M Schenk 等。雄激素代谢基因多态性与前列腺癌风险及雄激素浓度的关联：来自前列腺癌预防试验的结果。《癌症》，122(15):2332–2340，2016年。Haitao Chen, Xu Liu, Charles B Brendler, Donna P Ankerst, Robin J Leach, Phyllis J Goodman, M Scott Lucia, Catherine M Tangen, Li Wang, Fang-Chi Hsu 等。将遗传风险评分纳入家族史评估，可识别出数量翻倍的前列腺癌高风险男性：前列腺癌预防试验的结果。《前列腺》，76(12): 1120–1129，2016年。Lisa W Chu, Cathee Till, Baiyu Yang, Catherine M Tangen, Phyllis J Goodman, Kai Yu, Yong Zhu, Summer Han, Ashraful M Hoque, Christine Ambrosone 等。前列腺癌预防试验中的昼夜节律基因与前列腺癌风险。《分子致癌》，57(3):462–466，2018年。Xinkun Nie 和 Stefan Wager。异质性治疗效应的拟神谕估计。《生物统计学》，108(2):299–319，2021年。