# 用于治疗效应泛化的分组 *LASSO* 变量选择方法

***Chuyu Deng***[a]、***Brandon Koch***[b]、***David M. Vock***[a]、***Joseph S. Koopmeiners***[a]

[a] *University of Minnesota*、 *Minneapolis*
[b] *The Ohio State University*、 *Columbus*

关键词： 因果推断、机器学习、治疗效应泛化、治疗效应校准、选择偏差、外部效度

摘要： 在公共卫生领域，我们通常关注某项干预措施对某一人群的处理效应，而该人群与最初评估该干预措施的实验人群存在系统性差异。当随机对照试验中存在处理效应异质性时，将处理效应从实验人群推广到所关注的目标总体是一个复杂的问题；这需要刻画处理效应异质性以及两个总体之间的基线协变量不匹配。尽管这一问题十分重要，但针对该情境下变量选择的文献仍然有限。本文提出一种基于分组 LASSO 的变量选择方法，用于治疗效应泛化，并将其应用于把极低尼古丁含量卷烟的处理效应推广至美国总体吸烟人群。

arXiv:2302.03832v1 [stat.ME] 8 Feb 2023

## 1. 引言

长期以来，随机对照试验（RCT）一直被视为估计因果治疗效应的金标准，人们也投入了大量关注，致力于制定确保其内部有效性的最佳实践[1, 2]。然而，实验样本往往无法代表感兴趣的目标总体，仅有内部有效性不足以回答公共卫生领域的许多问题 [3, 4, 5]。对于许多应用，我们希望将随机对照试验的结果从较小的实验样本推广到更广泛的目标总体 [1],；当实验样本不能代表目标总体时，如何推广因果治疗效应，是研究人员和政策制定者关注的一个领域。具体而言，本研究的动机应用是将一项随机对照试验所收集的实验样本中，极低尼古丁含量卷烟对每日吸烟支数的处理效应，推广至美国整体吸烟人群。在这一应用中，推广问题至关重要：根据临床试验结果制定的广泛政策变革将影响所有美国吸烟者，但实验样本并不能代表我们的目标总体。

能否将基于实验样本估计的处理效应推广到目标人群，取决于两个主要的统计学考量：治疗效应异质性，以及实验样本与目标人群之间的基线协变量不匹配。治疗效应异质性（TEH）指同一种处理对不同个体产生不同影响的现象。如果某种处理对总体中的每个人产生的影响都相同，那么在实验样本中估计的处理效应就可直接适用于目标人群，也就不需要进行治疗效应泛化。反之，处理效应可能随一个或多个协变量而在样本中变化，从而产生亚组特异性治疗效应。在这种情况下，了解处理效应如何在样本中变化，是推广研究结果的关键步骤。实验样本与目标人群之间的基线协变量不匹配，指的是实验样本并非目标人群的代表性样本。如果实验样本与目标人群之间不存在协变量不匹配，那么在实验样本中估计的处理效应就可直接适用于目标人群。当存在治疗效应异质性和协变量不匹配时，必须同时考虑 TEH 和协变量不匹配，才能将处理效应推广到目标人群。

人们对开发用于解决治疗效应泛化问题的统计方法日益关注。 [6]。Stuart 等人提出采用三种不同的倾向评分方法（此处的倾向是指被纳入实验样本的倾向，也称为选择模型）——逆概率加权（IPW）、完全匹配和亚分类——来评估实验样本与目标人群是否足够相似，从而能够推广研究结果 [7]。倾向得分也可用于对研究中的协变量或个体重新加权，使其与目标人群相似。Tipton 等人开发了一种类似的、基于倾向评分的亚分类估计量 [8]。Kern 等人提出了一种基于结局回归建模的方法，使用贝叶斯加性回归树（BART） [9]。该方法首先使用马尔可夫链蒙特卡罗算法为实验样本拟合结局的 BART 模型。然后，根据反事实结局后验分布的抽样结果，计算目标人群中每个人的因果效应。将这些个体因果效应的平均值作为目标人群的平均处理效应。最后，Chan 等人提出了一种界限方法，用目标人群平均处理效应的区间估计取代点估计 [10]。这些区间估计取决于实验样本与目标人群之间的重叠比例、实验样本中的处理效应，以及关于可能的治疗影响最小值和最大值的假设。

在实践中，将随机对照试验的结果推广到目标总体时，可能有许多可用协变量，但并不确定应将哪些协变量纳入用于治疗效应泛化的结局模型或倾向评分模型。然而，迄今尚无论文详细评估在缺乏适当变量选择的情况下，不同协变量子集是否以及如何影响推广方法的模型性能。相比之下，潜在混杂因素的变量选择是因果效应估计的核心，已在文献中得到广泛研究。这些变量选择方法包括一类基于模型平均的方法，例如贝叶斯混杂调整（BAC） [11], 模型平均双重稳健（MA-DR）估计量 [12], ，以及用于倾向评分变量选择和模型平均的贝叶斯方法[13]。另一类基于惩罚回归的方法包括一种用于倾向评分模型的结局自适应 LASSO 方法 [14], 、一种惩罚可信区域方法（WR） [15], ，以及一种自适应组套索方法，用于同时在结局模型和协变量模型中选择协变量 [16]。然而，这些现有变量选择方法无法同时处理推广情境中存在的治疗效应异质性和基线协变量不匹配问题。我们在第 2 节中说明，在两个不匹配的总体之间进行推广时，调整与处理效应相关的基线协变量十分重要，而纳入无关协变量则会产生不利影响。因此，为推广不同总体之间的处理效应而进行变量选择是一个重要问题。对于处理效应的推广，我们希望找到一组与处理效应相关、且在两个总体中分布不同的协变量；这与混杂因素变量选择方法所识别的协变量集合不同。近期论文表明，在某些假设下，通过调整分离集——一组同时影响抽样机制和治疗效应异质性的协变量——可以识别目标总体平均处理效应（TATE） [8, 9, 17]。Egami 等人提出了一种方法，使用马尔可夫随机场对观测协变量之间的条件独立关系及已有领域知识进行编码，从而选择分离集。随后，使用该分离集，通过逆概率加权估计量（IPW）估计目标总体平均处理效应 [17]。该方法允许使用较为稀疏的目标数据集，而不必要求实验样本和目标数据集中都包含丰富的协变量数据。不过，该方法并非完全由数据驱动，因为抽样集——即决定实验样本抽样机制的变量集合——必须依据领域知识预先指定。

本文介绍 GLAVeS（组套索变量选择），这是一种治疗效应泛化方法，采用改进的自适应组套索方法，对选择模型和结局模型进行同步变量选择。随后，使用结局模型估计目标总体平均处理效应（TATE）。该方法能够识别与处理交互作用较弱的协变量：它在惩罚处理交互项时，同时考虑每个协变量被选入目标总体的概率。因此，GLAVeS 能够保留与处理交互作用较弱的协变量，而标准变量选择方法仅考虑协变量与结局的关联，因而会遗漏这些协变量。

我们首先在第 2 节中说明，在治疗效应推广的情境下，进行恰当的变量选择是必要的。在第3节中，我们提出一种新颖的数据驱动方法 GLAVeS，用于同时进行变量选择并估计目标总体平均处理效应。在第4节中，我们开展模拟研究，将 GLAVeS 的表现与其他治疗效应推广方法进行比较。随后，我们利用烟草与健康人口评估（PATH）研究——一项针对美国吸烟者、具有全国代表性的调查——将极低尼古丁含量（VLNC）香烟的效应推广至美国吸烟人群。最后，我们在第5节中讨论相关问题。

## 2. 用于推广的变量选择的影响

### 2.1. 符号与假设

本文采用 Kern 等人 [9] 所概述的样本平均处理效应（SATE）和目标总体平均处理效应（TATE）符号及假设。我们假设数据分析师拥有两个数据源：一个实验样本，其中包含处理、协变量和结局数据；以及一个来自目标总体的代表性样本，其中至少包含协变量信息。实验样本数据包括一个连续型结局 $Y_i^{obs}$、一个二元处理分配 $A_i$，其中 $A_i = 1$ 表示处理，$A_i = 0$ 表示对照，以及研究对象的基线协变量 $\boldsymbol{X_i} = (X_{i1}, ..., X_{ip})$，研究对象为 $i = 1, ..., n$。设 $Y_i(a)$ 为实验数据中个体 $i$ 在接受处理 $a$ 时的潜在结局，其中 $a \in \{0, 1\}$。目标总体样本仅包含研究对象的基线协变量 $\boldsymbol{X_{ij}^*}$，其中 $j = 1, ..., p$，研究对象为 $i = 1, ..., m$。设 $Y_i^*(a)$ 为目标数据中随机个体 $i$ 在接受处理 $a$ 时的潜在结局，其中 $a \in \{0, 1\}$。在本文余下部分，除非必要，我们在符号中省略 $i$。

这里需要满足若干标准因果假设，包括稳定单位处理值假设（SUTVA）、正值性以及不存在未测量混杂因素（$A \perp \{Y(1)\ Y(0)\}|X$）。随机研究中不存在未测量混杂因素，而这正是本章所考虑的情形。Stuart 等人 [7] 等人此前关于可推广性的研究已讨论并确立了这些假设。此外，我们还需要假设：给定基线协变量信息后，选择机制是强可忽略的。该假设意味着，对于个体 $i$，给定其基线协变量$X_i$后，其处理效应$Y_i(1) - Y_i(0)$的分布与其属于实验总体还是目标总体无关；结合上述其他因果假设，这意味着 $E(Y|A, \boldsymbol{X}) = E(Y^*|A, \boldsymbol{X})$。

我们将 SATE 定义为对实验人群中个体层面的因果效应取平均的因果估计目标:

$$\begin{aligned}\Delta &= E(Y(1) - Y(0)) \\ &= \int \Big(E(Y(1)|\boldsymbol{X}) - E(Y(0)|\boldsymbol{X})\Big) f(\boldsymbol{X}) d\boldsymbol{X} \\ &= \int \Big(E(Y|A = 1, \boldsymbol{X}) - E(Y|A = 0, \boldsymbol{X})\Big) f(\boldsymbol{X}) d\boldsymbol{X}\end{aligned}$$

其中，$f(\boldsymbol{X})$ 是实验人群中协变量的密度函数。第二个等式基于前述因果假设成立。我们可以分别利用实验样本中被分配接受处理和对照的个体，为 $E(Y|A = 1, \boldsymbol{X}, \boldsymbol{\beta}_{A=1})$ 和$E(Y|A = 0, \boldsymbol{X}, \boldsymbol{\beta}_{A=0})$ 建立模型，这些模型可能依赖于参数向量 $\boldsymbol{\beta}_{A=1}$ 和 $\boldsymbol{\beta}_{A=0}$。我们可以使用标准方法（例如最大似然估计量）估计这些参数。如果我们使用实验样本的经验分布ˆ来估计 $f(\boldsymbol{X})$（即 $f(\boldsymbol{X})$），那么 $\Delta$ 的插件估计量为:

$$\begin{aligned}\hat{\Delta} &= \int \Big(E(Y|A = 1, \boldsymbol{X}, \hat{\boldsymbol{\beta}}_{A=1}) - E(Y|A = 0, \boldsymbol{X}, \hat{\boldsymbol{\beta}}_{A=0})\Big) \hat{f}(\boldsymbol{X}) d\boldsymbol{X} \\ &= \frac{1}{n} \sum_{i=1}^{n} E(Y|A = 1, X_i, \hat{\boldsymbol{\beta}}_{A=1}) - E(Y|A = 0, X_i, \hat{\boldsymbol{\beta}}_{A=0})\end{aligned}$$

我们将 TATE 定义为对目标总体中个体层面的因果效应取平均的因果估计目标:

$$\begin{aligned}\Delta^* &= \int \Big(E(Y^*|A = 1, \boldsymbol{X}^*) - E(Y^*|A = 0, \boldsymbol{X}^*)\Big) f(\boldsymbol{X}^*) d\boldsymbol{X}^* \\ &= \int \Big(E(Y|A = 1, \boldsymbol{X}^*) - E(Y|A = 0, \boldsymbol{X}^*)\Big) f(\boldsymbol{X}^*) d\boldsymbol{X}^*\end{aligned}$$

我们所定义的目标总体平均处理效应，有时在其他关于推广的研究中被称为总体平均处理效应（PATE）。我们认为，目标总体平均处理效应这一术语（“目标”）更直观、更灵活，因为它承认实验估计值可以推广到多个目标人群。要获得目标总体平均处理效应，可以在选择机制强可忽略的假设下，使用实验人群中 $E(Y|A = 0, \boldsymbol{X})$ 和 $E(Y|A = 1, \boldsymbol{X})$ 的估计值来估计 $E(Y^*|A = 1, \boldsymbol{X}^*)$ 和 $E(Y^*|A = 0, \boldsymbol{X}^*)$

因此， $\Delta^*$ 的插件估计量为:

$$\begin{aligned}\hat{\Delta} &= \int \Big(E(Y|A=1,\boldsymbol{X}^*,\hat{\boldsymbol{\beta}}_{A=1}) - E(Y|A=0,\boldsymbol{X}^*,\hat{\boldsymbol{\beta}}_{A=0})\Big)\hat{f}(\boldsymbol{X}^*)d\boldsymbol{X}^* \\ &= \frac{1}{m}\sum_{i=1}^{m} E(Y|A=1,X_i^*,\hat{\boldsymbol{\beta}}_{A=1}) - E(Y|A=0,X_i^*,\hat{\boldsymbol{\beta}}_{A=0}) \end{aligned} \tag{1}$$

**2.2. 模拟研究**

根据现有的治疗效应泛化文献，我们知道，可以利用一组同时影响抽样机制和治疗效应异质性的协变量来识别目标总体平均处理效应 [8, 9, 17]。然而，尽管这些结果明确了为识别目标总体平均处理效应而必须纳入的最小协变量集合，却未能说明恰当进行变量选择对偏差和效率的影响。

我们开展了一项简单的模拟研究，以说明在治疗效应泛化模型中进行适当变量选择的重要性，尤其是在实验样本与目标人群的协变量分布不同的情况下。为简便起见，本次模拟仅使用一个协变量，但结果可以推广至多个协变量。

我们模拟了两种情景：1. 实验样本与目标人群的协变量分布相同；2. 实验样本与目标人群的协变量分布不同。在两种情景下，实验数据集均包含 300 个观测值，包括一个二元处理变量（$A$）、一个协变量（$X$）和一个结局变量（$Y$）；目标总体样本包含 900 个观测值和一个协变量（$X^*$）。在实验样本与目标人群的协变量分布相同的情景下， $X$和 $X^*$均模拟自 $N(0,\ 1)$。在实验样本与目标人群的协变量分布不同的情景下， $X$模拟自 $N(0,\ 1)$，而$X^*$模拟自 $N(1,\ 1)$。

实验样本数据集中的结局变量生成时与我们关注的 3 种主要协变量类型相匹配：1. 协变量与结局变量无关（$E(Y|A,X) = A$） 2. 协变量与结局变量相关（$E(Y|A,X) = A + X$） 3. 协变量与结局变量相关，且与处理存在交互作用（$E(Y|A,X) = A + X + AX$）。这总共构成了 6 种不同的数据生成情景。

| **实验样本** | **目标总体** |
|---|---|
| $X_i \sim \begin{cases} N(0,\ 1), & \text{对于“协变量分布相同”的情形} \\ N(1,1), & \text{对于“协变量分布不同”的情况} \end{cases}$ | $X_i^* \sim N(0,\ 1)$ |
| $A_i \sim Ber(0.5)$ | |
| $Y_i = \begin{cases} \begin{bmatrix} A & X & AX \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T + \varepsilon_i, & \text{对于“}Y\text{ 与 }X\text{无关”的情况} \\ \begin{bmatrix} A & X & AX \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^T + \varepsilon_i, & \text{对于“}Y\text{ 与 }X\text{相关”的情况} \\ \begin{bmatrix} A & X & AX \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T + \varepsilon_i, & \text{对于“}Y\text{ 与 }AX\text{相关”的情形} \end{cases}$ | |
| 其中 $\varepsilon_i \sim N(0,\ 1_{。}5)$ | |

图 1：模拟研究中 6 种数据情景的数据生成。

接着，我们评估了以下 3 个用于描述结局的线性模型的表现:

$$\begin{aligned}\mathbb{E}(Y|A,X) &= \alpha_0 + \alpha_1 A \\ \mathbb{E}(Y|A,X) &= \beta_0 + \beta_1 A + \beta_2 X \\ \mathbb{E}(Y|A,X) &= \gamma_0 + \gamma_1 A + \gamma_3 X + \gamma_4 AX\end{aligned}$$

我们使用实验样本数据集拟合每个模型，然后使用目标总体中的相应样本估计各情景下的结局，并据此计算目标总体平均处理效应。针对每种数据生成情景和模型，重复此过程 1000 次，估计得到的目标总体平均处理效应的均方误差如图 1 所示。

在实验样本与目标人群的分布没有差异的情景下，正如预期，假设模型与各数据生成模型相匹配时，均方误差最低

| | | $E(Y) = A + \varepsilon$ | $E(Y) = A + X + \varepsilon$ | $E(Y) = A + X + AT + \varepsilon$ |
|---|---|---|---|---|
| Same Covariate Distributions | $Y$ unrelated to $X$ | **0.0289** | 0.029 | 0.0291 |
| | $Y$ related to $X$ | 0.0425 | **0.0284** | 0.0286 |
| | $Y$ related to $AX$ | 0.0622 | 0.0367 | **0.0319** |
| Different Covariate Distributions | $Y$ unrelated to $X$ | **0.0296** | 0.0297 | 0.0603 |
| | $Y$ related to $X$ | 0.0408 | **0.0314** | 0.0643 |
| | $Y$ related to $AX$ | 1.0693 | 1.0358 | **0.0603** |

表 1：均方误差 f 或者说，在 3 种不同线性模型下的 6 种数据情景。各列代表假设模型 ls, 各行表示 真实的数据生成模型。

对估计的目标总体平均处理效应（TATE）。我们还发现，只要模型包含正确的项，即使含有多余项（即系数真实值为零的协变量），似乎也不会显著影响其表现。在实验样本和目标总体的分布存在差异的情景下，与各数据生成模型相匹配的模型得到的均方误差最低。在这种情况下，我们发现，遗漏分布存在差异的协变量的处理交互作用，会妨碍对目标总体平均处理效应（TATE）进行准确估计；这与我们的直觉以及 Egami 等人 [17] 的研究结果一致。然而，在模型中加入多余的交互项会对均方误差产生不利影响；不必要地加入这些项会使均方误差几乎翻倍，这凸显了有针对性地进行变量选择的重要性。

尽管本模拟研究仅设置了 1 个协变量，但它有效展示了在包含处理交互作用的标准线性模型中，所有可能的纳入/排除模式对均方误差造成的惩罚。我们认为，在更贴近现实的情景中，协变量可能从我们所列举的任一情形中“抽取”，此处展示的均方误差惩罚仍然适用。

## 3. 方法

### 3.1. GLAVeS

GLAVeS 的目标是识别那些与处理效应的交互作用较弱的协变量；仅考虑协变量对结局的预测性能的标准变量选择方法不会选中这些协变量。为此，我们在结局模型中对交互项进行惩罚时，将每个协变量与进入实验样本概率之间的关联纳入考量。这样，我们就能保留那些与处理效应交互作用较弱、但在实验样本与目标人群之间分布差异显著的协变量。

GLAVeS 与 GLiDeR [16], 相似，后者采用一种改良的自适应组套索方法，同时对处理模型和结局模型进行变量选择，并估计平均处理效应。GLiDeR 能够识别与结局关联较弱、但与处理分配关联较强的协变量。在我们的情形中，遗漏那些与处理效应相关（即与处理存在交互作用）、同时也与样本选择相关的协变量会带来不利影响。因此，我们采用类似的框架，以识别那些与处理效应关联较弱、但在实验样本与目标人群之间分布差异显著的协变量。与原始组套索一样，分组协变量的系数要么全部为零，要么全部非零。GLiDeR 对这一设置进行了修改：它不再将同一回归模型中的不同协变量分组，而是将处理模型和结局模型中的相同协变量分组。我们进一步修改这一设置，以选择模型取代处理模型，并使结局模型同时包含主效应和与处理的交互作用：

结局模型：

$$E(Y|A, \boldsymbol{X}) = \sum_{j=1}^{p} \beta_j X_j + \sum_{j=1}^{p} \beta_{p+j} X_j A + \beta_{2p+1} + \beta_{2p+2} A \tag{2}$$

选择模型：

$$logit(P(S = 1|\boldsymbol{X}^C)) = \sum_{j=1}^{p} \gamma_j X_j^C + \gamma_{p+1} \tag{3}$$

其中，$\boldsymbol{X}^C = \boldsymbol{X}$ 表示受试者来自实验样本，$\boldsymbol{X}^C = \boldsymbol{X}^*$ 表示受试者来自目标样本。请注意，$Y$、$\boldsymbol{X}$ 和 $\boldsymbol{X}^*$ 已分别按各自的标准差进行缩放。 $S \in \{0, 1\}$

分别表示来自相应总体的实验样本和目标样本成员身份。令 $g(.)$和 $h(.)$分别表示上述结局模型和选择模型的条件均值，并令 $\Phi_{\text{out}}(Y, A, \boldsymbol{X}; \beta)$和 $\Phi_{\text{sel}}(S, \boldsymbol{X}; \gamma)$分别表示结局模型和选择模型的损失函数。这里，结局模型使用线性回归模型，选择模型使用逻辑回归模型。在这种情况下， $\Phi_{\text{out}}$是平方误差损失，$(Y - g(A, \boldsymbol{X}))^2$, ，而 $\Phi_{\text{sel}}$ 与二项负对数似然成正比，$S \times \log(1 + e^{-2h(\boldsymbol{X}^*)}) - (1 - S) \times \log(1 + e^{-2h(\boldsymbol{X})})$。我们将向量 $\boldsymbol{\alpha} = (\beta, \gamma)$ 划分为 $2p + 2$ 组，使得$\alpha_1 = (\beta_1)$ ..., $\alpha_p = (\beta_p)$ $\alpha_{p+1} = (\beta_{p+1}, \gamma_1)$ ..., $\alpha_{2p} = (\beta_{2p}, \gamma_p)$ $\alpha_{2p+1} = (\beta_{2p+1}, \gamma_{p+1})$ $\alpha_{2p+2} = (\beta_{2p+2})$。这些组包括 $p$ 个结果主效应系数、与选择模型中相应系数归为一组的 $p$ 个结果处理交互作用系数、两个模型中的截距组成的一组，以及单独成组的处理系数。选择模型与结局模型中的这些系数组合会迫使协变量同时进入或退出最终模型。这样分组的意图是，较弱的处理交互作用系数被纳入模型的可能性，可以通过考虑选择模型中相应的系数而提高。

因此，这一同步变量选择程序包括求解：

$$\begin{aligned}\hat{\alpha}(\lambda) = &\underset{\boldsymbol{\alpha}}{argmin}\Big(\frac{1}{n}\sum_{i=1}^{n}\Phi_{\text{out}}(Y, A, \boldsymbol{X}_i; \beta) \\ &+ \frac{1}{n+m}\Big(\sum_{i=1}^{n}\Phi_{\text{sel}}(S = 1, \boldsymbol{X}_i; \gamma) + \sum_{i=1}^{m}\Phi_{\text{sel}}(S = 0, \boldsymbol{X}_i^*; \gamma)\Big) + \lambda\sum_{k=1}^{2p+2} w_k||\alpha_k||_2\end{aligned} \tag{4}$$

我们强调，结局系数仅使用实验样本进行估计，而选择系数则使用实验样本和目标样本进行估计。

在组套索中， [18], 组 $k$ 的惩罚权重通常取为 $w_k = \sqrt{c_k}$，其中 $c_k$ 是组 $k$的基数（即组中的元素个数）。但在这里，权重设定如下：

$$w_k = \begin{cases}\frac{1}{|v_k|}, & \text{for } k = 1, ..., p \\ \frac{\sqrt{2}}{|v_k|}, & \text{for } k = p+1, ..., 2p \\ 0, & \text{for } k = 2p+1, 2p+2\end{cases} \tag{5}$$

对于 $k = 1, ..., 2p$，分子对应组 $k$的基数，而分母 $v_k \neq 0$ 是对协变量 $k$ 在“完整”结局模型中的处理交互作用回归系数的估计；该模型包含主效应和处理交互作用，且不进行变量选择。分母的这一设定会对那些与结局无关的处理交互作用协变量（即 $|v_k|$较小）施加强惩罚，即使它们与选择过程密切相关。对于截距 $k = 2p + 1$, $2p + 2$，我们将权重设为 $w_k = 0$ ，使其不受惩罚。这里采用所提权重的目的，是选出与结局相关的强处理交互作用协变量，或是与选择过程密切相关且对结局具有处理交互作用的协变量。

由于我们同时考虑结局和选择损失函数，但希望仅将广义交叉验证（GCV）应用于结局模型，因此使用广义交叉验证（GCV）的修改版本来求解 $\hat{\alpha}(\lambda)$。这与 GLiDeR [16]所使用的 GCV 相同。在一般的组套索中， $\lambda$ 的特定取值下的 GCV 为 $\frac{RSS}{(1-df/n)^2}$，其中 RSS 是残差平方和，$df = \sum_{k=1}^{K} I(||\hat{\xi}_k|| > 0) + \sum_{k=1}^{K}\frac{||\xi_k||}{||\tilde{\xi}_k||}(d_k - 1)$，其中 $\hat{\xi}_k$和 $\tilde{\xi}_k$分别是 $k = 1$、...、 $K$组系数的自适应组套索估计量和最小二乘估计量，每组大小为 $d_k$。由于我们只希望针对结局进行模型选择，因此分子和分母都只使用 $\hat{\alpha}(\lambda)$ 中与结局模型系数相对应的部分：

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_{2p+1}(\lambda) - \hat{\beta}_{2p+2}(\lambda)A_i - \sum_{k=1}^{p}\hat{\beta}_k(\lambda)X_{ki} + \hat{\beta}_{p+k}(\lambda)X_{ki}A_i)^2}{(1 - (2 + \sum_{k=1}^{2p} I(||\hat{\beta}_k(\lambda)|| > 0) + \sum_{k=1}^{2p}\frac{|\hat{\beta}_k(\lambda)|}{|v_k|})/n)^2} \tag{6}$$

$\hat{\beta}(\lambda)$, $\hat{\gamma}(\lambda)$}，我们可以将 $\hat{\alpha}(\lambda) = \{$给定解 $\hat{\beta}(\lambda)$ 代入结局模型，以获得估计的目标总体平均处理效应：

$$\Delta^* = \hat{\beta}(\lambda)_{2p+2} + \frac{1}{m}\sum_i\sum_j \hat{\beta}(\lambda)_{p+j}X_{ij}^* \tag{7}$$

置信区间可以通过自助法计算。由于GLAVeS模拟中的TATE估计值尽管具有较高的变量选择准确率，却显示出显著偏差，因此我们还引入了OLSGLAVeS：先使用GLAVeS进行推广的变量选择，然后使用简化模型的普通最小二乘估计获得TATE估计值。

### 3.2. 调查权重

在公共卫生领域，我们的目标总体通常是一个广泛的人群（例如，美国成年人口总体），很难通过随机抽样获得该人群的代表性样本。然而，许多大型调查能够提供有关目标总体的大量信息。需要注意的是，这些原始分析性调查数据集并不是目标总体的代表性样本；它们源自复杂的调查设计，例如对某些人口统计群体进行过度抽样或抽样不足、调整无应答等。为了将这些调查样本校准到所关注的目标总体，最终的分析数据集中通常会包含调查权重，并利用这些权重获得目标总体参数的无偏估计量。通过下面列出的额外调整，可以很容易地将调查权重纳入GLAVeS。

令 $r_i$ 表示来自目标总体样本的原始调查权重，该样本包含 $i = 1, \ldots\ldots, m$ 名研究对象。在 GLAVeS 中，我们定义权重 $\Omega_i$，使实验样本中的研究对象权重为 1，并将目标总体中的研究对象的调查权重标准化为均值 1：

$$\Omega_i = \frac{m}{\sum_i r_i} r_i, \qquad \text{for } i = 1, ..., m \tag{8}$$

随后可以将该权重纳入方程 4：

$$\begin{aligned} \hat{\alpha}(\lambda) = \underset{\boldsymbol{\alpha}}{argmin}\Big(&\frac{1}{n}\sum_{i=1}^{n}\Phi_{\text{out}}(Y, A, \boldsymbol{X}; \beta) \\ &+ \frac{1}{n+m}\Big(\sum_{i=1}^{n}\Phi_{\text{sel}}(S=1, \boldsymbol{X}_i; \gamma) + \sum_{i=1}^{m}\Omega_i\Phi_{\text{sel}}(S=0, \boldsymbol{X}_i^*; \gamma)\Big) + \lambda\sum_{k=1}^{2p+2} w_k||\alpha_k||_2 \end{aligned} \tag{9}$$

其中，调查权重仅纳入选择模型，并在最终目标总体平均处理效应的方程 7 中使用：

$$\hat{\Delta}^* = \hat{\beta}(\lambda)_{2p+2} + \frac{1}{m}\sum_i\sum_j \Omega_i\hat{\beta}(\lambda)_{p+j}X_{ij}^* \tag{10}$$

## 4. 模拟

我们开展了一项模拟研究，以考察第 3.1 节所述变量选择方法的统计性质。在模拟中，我们研究了两种总体问题设定：一种包含 8 个协变量，另一种包含 15 个协变量。在每种设定下，我们改变了处理交互作用系数的大小以及协变量之间的相关性。下面将对我们的模拟研究作出完整说明。

### 4.1. 数据生成

我们考虑了 8 种不同情景，涵盖协变量数量、相关性、处理交互作用强度，以及实验样本与目标总体之间的不相似程度等方面的变化。对于每个重复实验样本集，我们模拟了 600 名个体，索引为 $i$；情景 1–4 中协变量总数为 8，情景 5–8 中协变量总数为 15。在不存在相关性的情景（情景 1、2、5 和 6）中，协变量 $\boldsymbol{X}_i = (X_{1i}, X_{2i}, \ldots\ldots, X_{pi})$ 服从多元标准高斯分布。在指定相关性的情景中，$\boldsymbol{X}_i = (X_{1i}, X_{2i}, \ldots\ldots, X_{pi})$ 服从多元高斯分布 $N(0, \Sigma)$，其中 $\Sigma$ 是一个对角线元素均为 1 的矩阵（即相关矩阵），使得表2中指定的每组相关协变量均具有复合对称协方差结构，相关系数为 0.6。我们还为每位个体生成了一个二元处理变量 $A_i$，其服从概率为 0.5 的伯努利分布。每位个体的结局 $Y_i$ 则根据 $N(A_i + g(\boldsymbol{X}_i), 1)$ 生成，其中 $g(\cdot)$ 为表2中指定的结局模型。

对于每个重复目标集，我们仅模拟了 300 名个体的协变量，$\boldsymbol{X}_i^* = (X_{1i}^*, X_{2i}^*, \ldots\ldots, X_{pi}^*)$。除非表2中的不同情景另有说明，这些协变量均再次从多元标准高斯分布中生成。

| 情景 | $p$ | 结局模型 $g$（。） | 相关<br>协变量 | 协变量分布<br>$N$（1，1）位于目标集 |
|---|---|---|---|---|
| 1 | 8 | $X_1 + X_3 + 0.1AX_1 + 0.1AX_3$ | | $X_1, X_2, X_3$ |
| 2 | 8 | $X_1 + X_3 + 0.05AX_1 + 0.05AX_3$ | | $X_1, X_2, X_3$ |
| 3 | 8 | $X_1 + X_3 + 0.1AX_1 + 0.1AX_3$ | $\{X_1, X_2\}, \{X_3, X_4\}$ | $X_1, X_2, X_3$ |
| 4 | 8 | $X_1 + X_3 + 0.05AX_1 + 0.05AX_3$ | $\{X_1, X_2\}, \{X_3, X_4\}$ | $X_1, X_2, X_3$ |
| 5 | 15 | $X_1 + X_2 + 0.1AX_1 + 0.1AX_3 + 0.1AX_5$ | | $X_1, \ldots, X_7$ |
| 6 | 15 | $X_1 + X_2 + 0.05AX_1 + 0.05AX_3 + 0.05AX_5$ | | $X_1, \ldots, X_7$ |
| 7 | 15 | $X_1 + X_2 + 0.1AX_1 + 0.1AX_3 + 0.1AX_5$ | $\{X_1, X_2\}$, $\{X_3, X_4\}$, $\{X_5, X_6, X_7\}$, $\{X_8, X_9, X_{10}\}$ | $X_1, \ldots, X_7$ |
| 8 | 15 | $X_1 + X_2 + 0.05AX_1 + 0.05AX_3 + 0.05AX_5$ | $\{X_1, X_2\}$, $\{X_3, X_4\}$, $\{X_5, X_6, X_7\}$, $\{X_8, X_9, X_{10}\}$ | $X_1, \ldots, X_7$ |

表2：考虑的8种情景。

### 4.2. 实施的方法

我们按照方法部分的说明实施了 6 种不同的方法，以及 GLAVeS。我们实施的第一种方法是一种朴素方法，计算实验样本中处理组与对照组的平均结局之差：

$$\Delta_{\text{S.Diff}} = \frac{\sum_{i=1}^{n} Y_i I(A_i = 1)}{\sum_{i=1}^{n} I(A_i = 1)} - \frac{\sum_{i=1}^{n} Y_i I(A_i = 0)}{\sum_{i=1}^{n} I(A_i = 0)}$$

该方法忽略了实验样本与目标总体之间的任何分布差异，并表明若不针对不同总体进行调整，可能会出现什么情况。在下文的模拟结果中，该方法记为“S.Diff”。下一组方法使用不同的变量选择模型来拟合结局模型。所有方法均使用公式1中的目标总体平均处理效应插件估计量。

第二种方法使用实验样本中的所有协变量，分别为处理组和对照组拟合线性回归模型。不使用变量选择程序。在模拟结果中，该方法记为“OLS”：

$$\hat{\Delta}_{\text{OLS}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{X}_i^{*T} \hat{\boldsymbol{\beta}}_{\text{OLS}, A=1} - \boldsymbol{X}_i^{*T} \hat{\boldsymbol{\beta}}_{\text{OLS}, A=0}$$

第三种方法使用实验样本，通过带有最小绝对收缩和选择算子（即 L1）惩罚的线性模型拟合所有协变量及可能的处理交互作用。我们使用交叉验证来选择使均方预测误差最小的惩罚参数 $\lambda$。我们使用 R语言中的 glmnet 软件包 [19] 实现了该方法。在模拟结果中，该方法记为“IntLASSO”：

第四种方法在实验样本中，使用所有协变量，分别为处理组和对照组拟合最小绝对收缩和选择算子模型。我们再次使用交叉验证来选择使每个模型中的均方误差最小的惩罚参数 $\lambda$ 。我们使用 R语言中的 glmnet 软件包 [19] 实现了该方法。在下方的模拟结果中，该方法记为“LASSO”：

我们实施的最后两种方法是第3节所述的 GLAVeS，以及“OLSGLAVeS”。在后者中，GLAVeS 仅用作变量选择方法，随后使用所选协变量和处理交互作用拟合普通最小二乘回归模型。然后将该线性回归模型应用于目标总体，以获得目标总体平均处理效应估计。

### 4.3. 模拟结果

对于每种情景，我们计算了各方法在1000次模拟中的平均偏差和均方误差。各种方法的偏差如表3所示。在所有考察的情景中，S.Diff均未针对不同总体进行调整，因此始终产生最大的绝对偏差。在执行变量选择的方法中，OLSGLAVeS在所有测试情景下的绝对偏差均最低。尽管在具有15个协变量的情景5至8中，OLS的偏差最低，但它不执行变量选择。总体而言，OLSGLAVeS的普通最小二乘估计仅比OLS估计的偏差略大；后者不执行变量选择，因此预期具有较低偏差。

| 情景 | S.Diff | 普通最小二乘法 | IntLASSO | 最小绝对收缩和选择算子 | GLAVeS | OLSGLAVeS |
|---|---|---|---|---|---|---|
| 1 | -0.202 | 0.003 | -0.103 | 0.003 | 0.123 | **-0.001** |
| 2 | -0.102 | 0.003 | -0.079 | 0.003 | 0.136 | **0.001** |
| 3 | -0.196 | 0.006 | -0.094 | 0.006 | 0.129 | **0.003** |
| 4 | -0.096 | 0.006 | -0.072 | 0.006 | 0.146 | **0.003** |
| 5 | -0.304 | 0.000 | -0.198 | -0.066 | 0.118 | **-0.017** |
| 6 | -0.154 | 0.000 | -0.151 | -0.050 | 0.134 | **-0.015** |
| 7 | -0.307 | 0.001 | -0.192 | -0.064 | 0.069 | **-0.018** |
| 8 | -0.157 | 0.001 | -0.153 | -0.049 | 0.090 | **-0.016** |

表2所述情景1–8的偏差

各种方法的均方误差如表4所示。在总共有8个协变量的情景中，当治疗交互作用的真实系数为0.1时（情景1和3），OLSGLAVeS的均方误差最低；当治疗交互作用的真实系数为0.05时（情景2和4），IntLASSO的均方误差最低。对于所有考察的情景，S.Diff始终产生最高的均方误差。在具有15个协变量的情景中，除协变量相关且真实治疗交互作用系数为0.1的情景（情景7）外，最小绝对收缩和选择算子在所有情景下的均方误差均最低；在情景7中，GLAVeS的均方误差最低。在所有情景中，最小绝对收缩和选择算子和OLSGLAVeS的均方误差始终最低：在所有方法和所有情景下，它们的均方误差要么最低，要么接近最低。附录??中还可以找到包含偏差和均方误差结果的其他模拟情景。

| 情景 | S.Diff | 普通最小二乘法 | IntLASSO | 最小绝对收缩和选择算子 | GLAVeS | OLSGLAVeS |
|---|---|---|---|---|---|---|
| 1 | 0.062 | 0.027 | 0.030 | 0.025 | 0.036 | **0.024** |
| 2 | 0.031 | 0.027 | **0.022** | 0.025 | 0.043 | 0.026 |
| 3 | 0.061 | 0.025 | 0.028 | 0.025 | 0.038 | **0.024** |
| 4 | 0.031 | 0.025 | **0.021** | 0.025 | 0.042 | 0.023 |
| 5 | 0.112 | 0.061 | 0.063 | **0.044** | 0.055 | 0.052 |
| 6 | 0.043 | 0.061 | 0.040 | **0.037** | 0.055 | 0.047 |
| 7 | 0.125 | 0.036 | 0.061 | 0.037 | **0.032** | 0.036 |
| 8 | 0.055 | 0.036 | 0.041 | **0.029** | 0.034 | 0.033 |

T表4：表2中概述的情景1-8的均方误差　　表2

对于表2中的每种情景，我们还通过1000次模拟，计算了各方法识别正确处理交互作用的平均敏感度和特异度。对于S.Diff，由于未进行调整且不使用处理交互作用，其敏感度始终为0，特异度始终为1。相反，对于OLS，模型中包含所有可能的处理交互作用，因此其敏感度始终为1，特异度始终为0。如果某个处理交互作用被LASSO模型的处理模型或对照模型选中，则认为该交互作用已被“选中”。由于OLSGLAVeS使用GLAVeS选出的变量，其敏感度和特异度始终相同。表5和表6列出了情景1-8的这些敏感度和特异度指标，以及1000次模拟所得的均方误差、偏差和标准差。

在表5所示的协变量总数为8的情景中，GLAVeS/OLSGLAVeS 在敏感度和特异度之间取得了最佳平衡，敏感度为 0.9 至 0.95，特异度为 0.68 至 0.78。虽然最小绝对收缩和选择算子在所有情景中的敏感度均为 1，但其特异度要低得多，仅为 0.35 至 0.38，这意味着它在纳入真实的处理交互作用的同时，也纳入了过多无关变量。在这些情景中，IntLASSO 的敏感度和特异度均低于 GLAVeS/OLSGLAVeS。

在表6所示的协变量总数为15的情景中，GLAVeS/OLSGLAVeS 再次在敏感度和特异度之间取得了最佳平衡，敏感度为 0.57 至 0.75，特异度为 0.74 至 0.76。最小绝对收缩和选择算子仍然纳入了过多无关变量，特异度为 0.43 至 0.55；尽管如此，它以 0.66 至 0.88 的敏感度识别出了更多真实的处理交互作用。在所有情景中，IntLASSO 的敏感度和特异度均与 GLAVeS/OLSGLAVeS 相近或更低。

烟草（商业品牌的典型水平）降至 0.4 mg/g，主要结局为第6周每日吸烟支数。六周时，随机分配至 0.4 mg/g 组的吸烟者吸烟支数显著少于随机分配至 15.5 mg/g 组的吸烟者 [20]。CENIC项目2是一项双盲、

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.062 | -0.202 | 0.145 | 0 | 1 |
| IntLASSO | 0.03 | -0.103 | 0.141 | 0.8 | 0.67 |
| 最小绝对收缩和选择算子 | 0.025 | 0.003 | 0.159 | 1 | 0.38 |
| GLAVeS | 0.036 | 0.123 | 0.146 | 0.95 | 0.78 |
| OLSGLAVeS | 0.024 | -0.001 | 0.156 | 0.95 | 0.78 |
| OLS | 0.027 | 0.003 | 0.166 | 1 | 0 |

(a) 场景1

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.031 | -0.102 | 0.143 | 0 | 1 |
| IntLASSO | 0.022 | -0.079 | 0.127 | 0.6 | 0.68 |
| 最小绝对收缩和选择算子 | 0.025 | 0.003 | 0.159 | 1 | 0.38 |
| GLAVeS | 0.043 | 0.136 | 0.158 | 0.9 | 0.74 |
| OLSGLAVeS | 0.026 | 0.001 | 0.161 | 0.9 | 0.74 |
| OLS | 0.027 | 0.003 | 0.165 | 1 | 0 |

(b) 情景2

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.061 | -0.196 | 0.15 | 0 | 1 |
| IntLASSO | 0.028 | -0.094 | 0.139 | 0.8 | 0.66 |
| 最小绝对收缩和选择算子 | 0.025 | 0.006 | 0.157 | 1 | 0.36 |
| GLAVeS | 0.038 | 0.129 | 0.145 | 0.93 | 0.71 |
| OLSGLAVeS | 0.024 | 0.003 | 0.154 | 0.93 | 0.71 |
| OLS | 0.025 | 0.006 | 0.158 | 1 | 0 |

(c) 情景3

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.031 | -0.096 | 0.147 | 0 | 1 |
| IntLASSO | 0.021 | -0.072 | 0.125 | 0.59 | 0.68 |
| 最小绝对收缩和选择算子 | 0.025 | 0.006 | 0.157 | 1 | 0.35 |
| GLAVeS | 0.042 | 0.146 | 0.145 | 0.9 | 0.68 |
| OLSGLAVeS | 0.023 | 0.003 | 0.152 | 0.9 | 0.68 |
| OLS | 0.025 | 0.006 | 0.158 | 1 | 0 |

(d) 情景4

表5：在总协变量数为8的情况下，情景1至4中各种方法的均方误差、偏差、标准差（SD）、敏感度（SE）和特异度（SP）。

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.112 | -0.304 | 0.143 | 0 | 1 |
| IntLASSO | 0.063 | -0.198 | 0.154 | 0.7 | 0.73 |
| 最小绝对收缩和选择算子 | 0.044 | -0.066 | 0.2 | 0.88 | 0.43 |
| GLAVeS | 0.055 | 0.118 | 0.203 | 0.75 | 0.75 |
| OLSGLAVeS | 0.052 | -0.017 | 0.228 | 0.75 | 0.75 |
| OLS | 0.061 | 0 | 0.248 | 1 | 0 |

(a) 情景 5

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.043 | -0.154 | 0.141 | 0 | 1 |
| IntLASSO | 0.04 | -0.151 | 0.129 | 0.44 | 0.76 |
| 最小绝对收缩和选择算子 | 0.037 | -0.05 | 0.185 | 0.73 | 0.47 |
| GLAVeS | 0.055 | 0.134 | 0.193 | 0.6 | 0.74 |
| OLSGLAVeS | 0.047 | -0.015 | 0.216 | 0.6 | 0.74 |
| OLS | 0.061 | 0 | 0.248 | 1 | 0 |

(b) 情景 6

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.125 | -0.307 | 0.176 | 0 | 1 |
| IntLASSO | 0.061 | -0.192 | 0.156 | 0.64 | 0.75 |
| 最小绝对收缩和选择算子 | 0.037 | -0.064 | 0.181 | 0.82 | 0.51 |
| GLAVeS | 0.032 | 0.069 | 0.166 | 0.7 | 0.75 |
| OLSGLAVeS | 0.036 | -0.018 | 0.19 | 0.7 | 0.75 |
| OLS | 0.036 | 0.001 | 0.19 | 1 | 0 |

(c) 情景7

| | MSE | Bias | SD | SE | SP |
|---|---|---|---|---|---|
| S.Diff | 0.055 | -0.157 | 0.175 | 0 | 1 |
| IntLASSO | 0.041 | -0.153 | 0.131 | 0.37 | 0.79 |
| 最小绝对收缩和选择算子 | 0.029 | -0.049 | 0.165 | 0.66 | 0.55 |
| GLAVeS | 0.034 | 0.09 | 0.162 | 0.57 | 0.76 |
| OLSGLAVeS | 0.033 | -0.016 | 0.18 | 0.57 | 0.76 |
| OLS | 0.036 | 0.001 | 0.19 | 1 | 0 |

(d) 情景8

表6：在总协变量数为15的情景5至8中，各种方法的均方误差、偏差、标准差（SD）、敏感度（SE）和特异度（SP）。

这是一项随机平行设计研究，于2014年7月至2017年3月期间在美国10个地点开展，旨在评估立即降低尼古丁含量与逐步降低尼古丁含量的影响。一组自愿参与、且无意在30天内戒烟的每日吸烟者，先完成了为期2周的基线吸烟期，随后被随机分配至以下3种治疗条件之一：1. 立即降低尼古丁含量：参与者使用尼古丁含量为每克烟草0.4毫克的香烟。2. 逐步降低尼古丁含量：参与者逐步吸食尼古丁含量越来越低的香烟，每种含量持续一个月，直至研究最后一个月，使用与立即降低组相同的每克烟草含0.4毫克尼古丁的极低尼古丁含量香烟。3. 常规尼古丁对照组：参与者获发指定香烟，其尼古丁含量为每克烟草15.5毫克，与市售香烟的平均尼古丁含量大致相当。参与者在为期20周的研究期间使用研究用香烟；评估的结局指标包括烟雾暴露生物标志物，并记录参与者的吸烟行为指标。该数据集包含46项人口统计学和基线协变量，例如年龄、性别、种族、教育程度、每日吸烟量、总尼古丁当量（TNE，一种常用的体内尼古丁剂量指标，包括尼古丁及其代谢物）、一氧化碳水平等。与随机分配至逐步降低组或对照组的吸烟者相比，随机分配至立即降低尼古丁含量组的吸烟者，其烟草暴露生物标志物随时间显著降低 [21]。

在烟草监管科学领域，目标总体是美国全部吸烟人群。如果不存在治疗效应异质性（即干预效果不会因参与者特征而异），随机对照试验人群与目标总体之间存在系统性差异并不会造成问题；但如果存在治疗效应异质性，随机对照试验估计的处理效应相对于目标总体中的效应可能存在偏差。在烟草监管科学领域，人们尤其关注这种偏差，因为任何监管规定都会影响所有美国吸烟者，而不仅是参与干预的人。将 CENIC研究中 VLNC 的处理效应推广至烟草使用与健康人口评估研究（PATH）的人群后，我们便能够估计潜在尼古丁减量政策对美国整体吸烟人群的影响。

我们利用 PATH（一项针对美国烟草使用者和非使用者开展的前瞻性队列研究，旨在调查烟草使用模式、风险认知以及对当前烟草产品的态度），将随机对照试验背景下评估的潜在烟草产品监管规定的影响推广至美国全部吸烟人群 [**?** ]。这项纵向研究由美国食品药品监督管理局烟草制品中心和美国国立卫生研究院支持，获资助每年收集数据，直至2024年。基线数据于2013年9月至2014年12月期间收集，涵盖45,971名成年人和青少年 [**?** ]。该研究收集的许多调查数据和生物标志物协变量，与 CENIC研究中的相同或类似协变量相对应。

| **变量** | **中位数[四分位距] 或计数（%）** | |
|---|---|---|
| | **CENIC** | **PATH** |
| n | 538 | 8351 |
| Age | 48 [35, 56] | 38 [26, 52] |
| 性别 = 男性 | 290 (53.9) | 4333 (51.9) |
| 基线体重（kg） | 83 [70, 96] | 77 [66, 91] |
| Race | | |
| 白人 | 336 (62.5) | 6375 (76.3) |
| 黑人 | 164 (30.5) | 1215 (14.5) |
| 其他 | 38 (7) | 761 (9.1) |
| 教育程度 | | |
| 高中及以下 | 49 (9.1) | 2490 (29.8) |
| 高中毕业 | 170 (31.6) | 2202 (26.4) |
| 大学或以上学历 | 319（59.3） | 3659（43.8） |
| 薄荷醇香烟的使用 =薄荷醇 | 243（45.2） | 3256（39） |
| 基线每日香烟总量 | 16 [11, 22] | 10 [5, 15] |

Ta表7：CENIC和PATH第一波参与者的人口统计学特征及基线特征。

表7呈现了两个数据集中均可获得的协变量汇总。总体而言，与CENIC人群相比，PATH第一波人群年龄更小、体重更轻、白人比例更高、受教育程度更低、使用的薄荷醇香烟更少，并且基线时吸烟量更低。CENIC人群中极低尼古丁含量香烟的样本平均处理效应估计为-7.65支/日，这意味着在CENIC人群中，使用极低尼古丁含量香烟的吸烟者平均每天比使用对照香烟的吸烟者少吸7.65支香烟。PATH数据集包含

我们用于将不同的目标总体平均处理效应估计推广至总体美国吸烟人群的调查权重。采用治疗效应泛化后，使用第4节所述的全部方法计算目标总体平均处理效应估计值，并在图2中呈现了100个自助法抽样样本的点估计值及其相应的百分位数95%置信区间。表8呈现了所选协变量及其系数的比较。

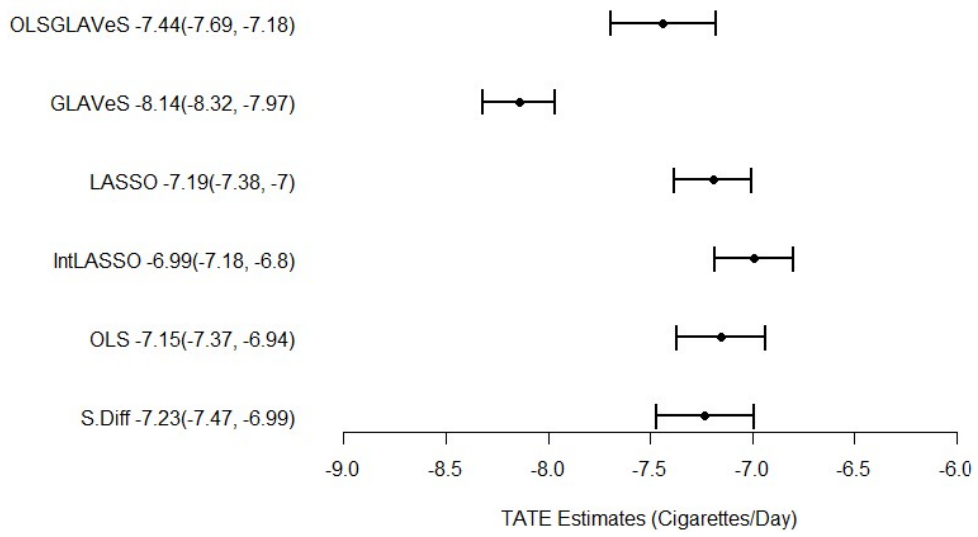


图2：使用各种估计方法将极低尼古丁含量香烟的目标总体平均处理效应推广至美国整体吸烟人群后，其目标总体平均处理效应估计值的森林图，以及相应的自助法百分位数95%置信区间。

从各种方法的目标总体平均处理效应估计值可以看出，将结果从CENIC推广至美国整体吸烟人群后，大多数估计值显示，极低尼古丁含量香烟的处理效应在每日7至7.5支香烟之间。只有GLAVeS给出的估计值超出此范围，差异为每日8.14支香烟。

就推广中的变量选择而言，GLAVeS/OLSGLAVeS 的选择最为严格：在对协变量与结局的关联以及 CENIC 人群和 PATH人群之间的分布差异进行惩罚后，仅将年龄选为处理交互项。IntLASSO 纳入了所有可能的处理交互项，而最小绝对收缩和选择算子仅排除了种族交互项。由于 S.Diff 不会调整人群间的分布差异，因此不纳入任何处理交互项。普通最小二乘法不进行任何变量选择，所有可能的处理交互项均纳入线性回归模型。

| 变量 | S.Diff | IntLASSO | 最小绝对收缩和选择算子 | GLAVeS | OLSGLAVeS | 普通最小二乘法 |
|---|---|---|---|---|---|---|
| Age | 0.11 | 0.12 | 0.14 | 0.17 | 0.14 |
| 性别 = 男性 | -0.12 | -1.36 | | | -0.15 |
| 基线体重（kg） | -0.04 | -0.01 | | | -0.05 |
| 种族 = 白人 | 0.87 | | | | 2.26 |
| 种族 = 黑人 | 0.55 | | | | 1.42 |
| 教育程度 = 高中及以下 | 0.28 | -0.18 | | | 0.33 |
| 教育程度 = 高中毕业 | -0.08 | -0.62 | | | -0.27 |
| 薄荷醇香烟的使用 = 薄荷醇 | 2.38 | 0.72 | | | 3.15 |
| 基线总量 每日香烟数量 | -0.15 | -0.22 | | | -0.17 |

表8：各种估计方法所选变量及处理交互项的估计系数。

## 5. 讨论

在当前因果推断文献中，针对潜在混杂因素的数据驱动变量选择方法有很多。然而，尽管治疗效应泛化对许多公共卫生问题都很重要，针对治疗效应泛化的变量选择仍鲜有研究，也没有公认的解决方案。标准变量选择方法足以捕捉较强的处理交互作用，但往往会遗漏较弱的交互作用；当协变量存在分布差异时，这会显著影响目标总体平均处理效应估计的准确性。GLAVeS 旨在捕捉这些较弱的处理交互作用：它采用改进的自适应组套索方法，在惩罚处理交互项的同时，纳入个体被选入目标总体的概率。

在我们考察的所有模拟情景中，与我们实现的其他任何方法相比，GLAVeS 均能以更高的综合敏感度和特异度筛选出治疗交互作用。识别正确治疗交互作用的准确性体现在 OLSGLAVeS 的 TATE 估计值偏差较低。不过，OLSGLAVeS 估计值的低偏差未必意味着均方误差最小，因为与我们实现的其他方法相比，其估计值的标准差更高。尽管如此，我们认为 TATE 估计的偏差是更重要的指标，因为偏差较大将导致估计值的 95%置信区间覆盖率较低。总的来说，对于可能存在许多较弱治疗交互作用的治疗效应泛化应用，我们建议使用 GLAVeS 选择协变量，并将这些协变量纳入普通最小二乘回归模型，以获得无偏的 TATE 估计。

未来，可以在不同设置下重复模拟，这将很有意义。例如，减少实验样本中的个体数量，以考察当推广基于少于 600 人的实验样本时各种方法的表现。另一种可能的扩展是，除了 GLiDeR 之外，调整其他用于混杂因素变量选择的方法，使其适用于治疗效应泛化框架，并将这些方法的表现与 GLAVeS 进行比较。

**致谢**

作者（们）声明，就本研究、论文撰写和/或本文发表，已获得以下经费支持：本研究由美国国家癌症研究所（资助编号：R01CA214825 和 R01CA225190）、美国国家药物滥用研究所（资助编号：R01DA046320、R03DA041870 和 U54-DA031659）以及美国国家转化科学促进中心（资助编号：UL1TR002494）资助。本文内容完全由作者负责，并不一定代表美国国立卫生研究院或美国食品药品监督管理局的官方观点。

# 参考文献


[1] W. R. Shadish, T. D. Cook, D. T. Campbell, Experimental and quasi-experimental designs for generalized causal inference, Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin and Company Boston MA ,,,, US , 2002. , V. C. Hayden, V. P. Pontes, Selection criteria and generalizability within the counterfactual framework: [2] H. I. Weisberg explaining the paradox of antidepressant-induced suicidality? , Clinical trials 6 (2) (2009) 109–118.,,,,, [3] A. Gheorghe T. E. Roberts J. C. Ives B. R. Fletcher M. Calvert, Centre selection for clinical trials and the generalisability of results: a mixed methods study , PloS one 8 (2).,, [4] H. Allcott S. Mullainathan External validity and partner selection bias, National Bureau of Economic Research Cambridge, MA , 2012. , [5] A. S. Deaton Instruments of development: Randomization in the tropics and the search for the elusive keys to economic,,, development Report 0898-2937 National bureau of economic research (2009).,, [6] E. Tipton R. B. Olsen A review of statistical methods for generalizing from evaluations of educational interventions Edu-, cational Researcher 47 (8) (2018) 516–524. doi:10.3102/0013189x18781522. URL https://journals.sagepub.com/doi/abs/10.3102/0013189X18781522 [7] E. A. Stuart S. R. Cole C. P. Bradshaw P. J. Leaf ,,,, The use of propensity scores to assess the generalizability of results from randomized trials , Journal of the Royal Statistical Society: Series A (Statistics in Society) 174 (2) (2011) 369–386.,,, [8] E. Tipton Improving generalizations from experiments using propensity score subclassification: Assumptions properties and contexts , Journal of Educational and Behavioral Statistics 38 (3) (2013) 239–266.,,,, [9] H. L. Kern E. A. Stuart J. Hill D. P. Green, Assessing methods for generalizing experimental impact estimates to target URL https://doi.org/10.1080/19345747.2015.1060282 populations Journal of Research on Educational Effectiveness 9 (1) (2016) 103–127. doi:10.1080/19345747.2015.1060282.,[10] W. Chan Partially identified treatment effects for generalizability,, URL https://doi.org/10.1080/19345747.2016.1273412 Journal of Research on Educational Effectiveness 10 (3)[11] C. Wang F. Dominici G. Parmigiani C. M. Zigler ,,,, (2017) 646–669. doi:10.1080/19345747.2016.1273412. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models , Biometrics 71 (3) (2015) 654–665. Biometrics 73 (2) (2017) [12] M. Cefalu F. Dominici N. Arvold G. Parmigiani Model averaged double robust estimation 410–421. [13] C. M. Zigler F. Dominici ,, Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects Journal of the American Statistical Association 109 (505) (2014) 95–107 ,, URL https://doi.org/10.1080/01621459.2013.869498 pMID: 24696528. arXiv:,[14] S. M. Shortreed A. Ertefaie ,, https://doi.org/10.1080/01621459.2013.869498 doi:10.1080/01621459.2013.869498. Outcome-adaptive lasso: Variable selection for causal inference Biometrics 73 (4) (2017), 1111–1122. doi:10.1111/biom.12679. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12679 [15] A. Wilson B. J. Reich Confounder selection via penalized credible regions ,,, Biometrics 70 (4) (2014) 852–61. doi:10.1111/[16] B. Koch, D. M. Vock, J. Wolfson, biom.12203. Covariate selection with group lasso and doubly robust estimation of causal effects, Biometrics 74 (1) (2018) 8–17. doi:10.1111/biom.12736. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12736 [17] N. Egami E. Hartman Covariate selection for generalizing experimental results,,, arXiv preprint arXiv:1909.02660). Journal of the Royal Statistical Society: [18] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables Series B (Statistical Methodology) 68 (1) (2006) 49–67. [19] N. Simon J. Friedman T. Hastie R. Tibshirani ,,,, Regularization paths for cox's proportional hazards model via coordinate URL http://www.jstatsoft.org/v39/i05/ descent Journal of Statistical Software 39 (5) (2011) 1–13. ,[20] E. C. Donny R. L. Denlinger J. W. Tidey ,,, J. S. Koopmeiners, N. L. Benowitz, R. G. Vandrey, M. al'Absi, S. G. Carmella, P. M. Cinciripini S. S. Dermody D. J. Drobes S. S. Hecht J. Jensen T. Lane C. T. Le F. J. McClernon I. D.,,,,,,,,,,,,,,,, Montoya S. E. Murphy J. D. Robinson M. E. Stitzer A. A. Strasser H. Tindle D. K. Hatsukami ,of reduced-nicotine standards for cigarettes , Randomized trial URL https://doi.org/10.1056/NEJMsa1502403 New England Journal of Medicine 373 (14) (2015) 1340–1349. doi:10.1056/[21] D. K. Hatsukami X. Luo J. A. Jensen M. al'Absi,,, NEJMsa1502403. S. S. Allen S. G. Carmella M. Chen P. M. Cinciripini R. Denlinger-,,,,,,,,,,,,,, Apte D. J. Drobes J. S. Koopmeiners T. Lane C. T. Le S. Leischow K. Luo F. J. McClernon S. E. Murphy V. Paiano J. D. Robinson H. Severson C. Sipe A. A. Strasser L. G. Strayer M. K. Tang R. Vandrey S. S. Hecht N. L. Benowitz ,,,,,,,,, E. C. Donny Effect of immediate vs gradual reduction in nicotine content of cigarettes on biomarkers of smoke exposure: A' randomized clinical trial JAMA 320 (9) (2018) 880–891. doi:10.1001/jama.2018.11473. ,URL https://doi.org/10.1001/jama.2018.11473