arXiv:2602.05041v1 [stat.ME] 4 Feb 2026

观察性研究中将簇视为混杂因素的加权框架

Eli Ben-Michael* Avi Feller† Luke Keele‡

2026年2月6日

摘要

在观察性研究中，当个体以组为单位聚类时（例如学校中的学生或医院中的患者），研究者通常通过调整簇级协变量或簇成员身份来处理混杂。本文提出一个统一的加权框架，阐明不同估计方法如何控制两种不同来源的不平衡：全局平衡（处理组和对照组单位在不同簇之间的差异）与局部平衡（簇内差异）。我们指出，采用随机效应倾向得分模型的逆倾向得分加权（逆概率加权）——目前文献中的标准方法——仅针对全局平衡和簇间恒定的水平位移，却不对局部平衡施加任何约束。随后，我们提出两种兼顾这两种平衡形式的方法。第一，层级平衡权重通过约束优化问题直接控制全局与局部平衡。第二，基于近期提出的广义Mundlak方法，我们开发了一种新颖的Mundlak平衡权重估计量，通过调整簇级充分统计量而非簇指示变量来实现平衡；该方法可适用于所有个体均接受处理或均未接受处理的小簇。重要的是，这些方法依赖不同的假设：层级平衡权重只要求在给定协变量和簇成员身份的条件下，处理可忽略；而Mundlak方法还要求满足指数族结构。最后，我们通过模拟研究，以及教育和卫生服务研究领域的两个应用，对这些方法进行比较；这两个应用呈现出截然不同的簇结构。

关键词：部分池化、加权、固定效应、随机效应、因果推断

*卡内基梅隆大学，宾夕法尼亚州匹兹堡福布斯大道 4800 号，邮编 15213，美国，电子邮箱：eben-michael@cmu.edu

†加州大学伯克利分校，伯克利赫斯特大道 2607 号，邮编 94720，美国，电子邮箱：afeller@berkeley.edu

‡宾夕法尼亚大学，费城云杉街 3400 号，邮编 19104，美国，电子邮箱：luke.keele@gmail.com

1 引言

在标准的观察性因果推断中，研究人员假设，仅调整个体层面协变量后不存在未测量的混杂因素。在许多观察性研究中，个体会以组的形式聚类，例如嵌套在学校中的学生或嵌套在医院中的患者。在这种情况下，研究人员可以通过进一步调整簇级协变量或簇成员身份本身来放宽这一假设。请考虑一项评估新型外科手术程序的观察性研究，正如我们在第 8 节中所做的那样。除了调整患者层面协变量（如年龄、健康状况指标和合并症）之外，我们还可能担心，医院层面的因素也会同时影响治疗决策和结果。例如，在外科医生接受更高级培训的医院中，新手术程序的使用频率可能更高。为捕捉这一点，我们可能希望进一步调整医院层面的外科培训指标。或者，我们可能希望调整收治医院本身，这也可以捕捉“未测量情境”的其他重要方面（Arpino 和 Mealli，2011年）。这个简单的例子凸显了研究人员在这种情境下调整混杂因素时面临的诸多权衡。

本文综述了为在观察性因果推断中考虑簇级混杂而发展出的多种加权估计量。在若干关于此聚类设定下基于模型的逆倾向得分加权（逆概率加权）的综述（Li 等2013；Chang and Stuart 2022年；Fuentes 等2022年）基础上，我们着重将两项新的方法学进展适配到这一领域。首先，我们考虑平衡权重，这是基于模型的逆概率权重的一种流行替代方法，它转而使用约束优化，直接寻找能够平衡各组之间协变量的权重（Ben-Michael 等（2020），2023年）。我们表明，现有的平衡权重与基于模型的逆概率加权方法主要针对全局平衡——跨簇的处理组和对照组单位之间的协变量差异——与局部平衡——簇内处理组和对照组单位之间的协变量差异——的组合。带随机簇截距的基于模型的逆概率加权是文献中迄今最常见的方法，它隐含地仅针对全局平衡，并不包含对局部平衡的任何控制。相比之下，带随机系数的基于模型的逆概率加权（Arpino 和 Mealli，2011）和层级平衡权重（Ben-Michael 等人2023年）都同时针对全局与局部平衡。一个主要挑战是，由于数据稀疏性，同时实现优异的全局与局部平衡往往并不可行；由于簇样本量有限，某些簇内的局部平衡可能尤其难以实现。我们表明，若干方法需要剔除小簇，或剔除几乎全部为处理组或对照组单位的簇，这会改变目标估计量。

其次，我们考虑 广义 *Mundlak* 估计 (Arkhangelsky 和 Imbens 2024年)，该方法调整的是簇级充分统计量，而非簇成员身份本身。其关键假设是，给定个体层面信息和簇级充分统计量后，处理分配可忽略（假设3）。如果给定个体层面信息和簇成员身份后处理分配可忽略，并且额外假设个体层面协变量与处理的联合分布服从指数族分布，则可推出这一假设。与调整簇成员身份的方法不同，Mundlak 方法可以保留

方法	平衡目标	主要局限
调整簇成员身份的方法（假设*2*）		
簇特定截距	仅全局	无法控制局部平衡
分层分析	仅局部	必须剔除没有处理状态变异的簇
		（续）
随机系数	全局 +局部	在计算上 难以处理
		收敛问题
层级平衡权重	全局 + 局部	必须删除全部为否的簇
		处理组单位
调整簇级协变量的方法（假设*3*；强于假设*2*）		
广义 Mundlak 估计量	全局 +局部	依赖基于模型的倾向-
		倾向评分；要求指数
		族假设并对
		充分统计量的规定
Mundlak 平衡权重	总体 + 局部	要求指数族假
		假设与设定
		充分统计量

表1：聚类调整加权方法概述

较小的簇，但代价是必须施加更强的指数族假设，该假设要求簇级充分统计量能够完整捕捉相关的簇级信息。随后，我们提出*Mundlak* 平衡权重，这是一种同样调整簇级充分统计量的平衡权重估计量。我们概述了该估计量的两种变体，它们以不同方式约束全局平衡，并指出这两种估计量在实践中颇具吸引力。

最后，我们通过一项模拟研究和两个具有不同簇结构的实证示例，展示这些估计量之间的权衡。第一个应用是一项关于较小班级规模对考试成绩影响的观察性研究，其中学生嵌套于簇大小相对较小的学校中。第二个应用是一项关于肝胆胰疾病急诊普通外科手术的观察性研究，其中患者嵌套于拥有许多大型簇的医院中。我们展示了这两种情境下的权衡各不相同，并为实践提供了一些建议。

本文结构安排如下。在第2 节中，我们介绍符号表示、估计目标和识别假设，并回顾因果推断中常见的加权估计方法。在第3节中，我们提出用于实现全局与局部平衡的统一框架。在第4节中，我们运用该框架分析现有的、以簇成员身份为条件的方法。在第5节中，我们运用该框架分析改为以簇级协变量为条件的方法。在第6节中，我们介绍推断和偏差校正。在第7节中，我们开展模拟研究。在第8节中，我们分析两个不同的实证应用。

2 预备知识

2.1 符号表示与问题设定

我们首先介绍符号表示和问题设定，其中大部分内容是描述采用个体化处理分配的观察性研究时的标准做法。我们假设有一个由总体中独立同分布抽取的个体样本 $i = 1, .., n$，并将 Z_i 记为二元处理，其中$Z_i = 1$ 对应处理条件， $Z_i = 0$ 对应对照条件； n_1表示分配至处理的个体数， n_0 表示分配至对照的个体数。我们用 $Y_i \in \mathbb{R}$ 表示观测到的结果，用 $X_i \in \mathcal{X}$ 表示基线协变量向量。为简化符号表示，在计算期望时，我们将省略下标 i 。

接下来，我们定义一些额外的符号，用于描述这样一种情形：组或簇成员身份被认为与处理分配相关；我们交替使用组 和簇 这两个术语来表示个体的嵌套方式。令组指示变量 $G_i \in \{1, \ldots, K\}$ 表示个体 i 是否属于该组。例如，这种符号表示可以描述学校中的学生或医院中的患者。我们用 n_{1g} 和n_{0g} 表示组 $G_i = g$中的处理组和对照组个体数量。

将倾向得分（即给定基线协变量时接受处理的概率）记为 $e(x) \equiv P(Z = 1 \mid X = x)$。并定义条件期望结果 $m(z, x) \equiv \mathbb{E}[Y \mid Z = z, X = x]$。我们还将考虑在进一步以组成员身份为条件时的倾向得分和条件期望潜在结果： $e(x, g) \equiv P(Z = 1 \mid X = x, G = g)$ 和 $m(z, x, g) \equiv \mathbb{E}[Y \mid Z = z, X = x, G = g]$。

2.2 假设与估计目标

我们概述观察性研究的一组关键假设，以及簇的存在如何改变处理效应的识别。首先，我们采用标准的稳定单元处理值假设（Rubin 1986），即不存在隐性处理形式，且某个受试者的潜在结果不受其他受试者所接受处理的影响。因此，我们排除存在簇内干扰的应用情形，即个体在组内相互作用（例如，参见Hudgens 和 Halloran 2008）。接下来，我们可以定义每个个体的两种潜在响应，$(Y_i(1)\ Y_i(0)) \in \mathbb{R}^2$，分别对应其在处理和对照条件下的响应。这些假设意味着，观测到的结果为 $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

我们关注的是潜在结果或反事实状态之间的期望差异。在观察性研究中，推断目标通常是处理组平均处理效应(ATT)， $\tau \equiv \mathbb{E}[Y(1) - Y(0) \mid Z = 1] = \mu_1 - \mu_0$，其中我们将 $\mu_1 \equiv \mathbb{E}[Y(1) \mid Z = 1]$ 和$\mu_0 \equiv \mathbb{E}[Y(0) \mid Z = 0]$分别记为。估计 μ_1 很简单：处理组单位的观测结果均值是一个无偏估计。难点在于估计 μ_0，即处理组单位在对照条件下的期望反事实结果。为简洁起见，我们聚焦于处理组平均处理效应，但我们的结果可推广至其他线性估计目标，包括平均处理效应（ATE）和重叠人群平均处理效应（ATO；Li 等，2018）。

不考虑簇成员身份的可忽略性。在不使用簇级信息、以处理组平均处理效应为目标的观察性研究中，标准假设是强可忽略性(Rosenbaum 和 Rubin， 1983)。该假设包含：(i)可忽略的处理分配——给定个体层面协变量后，处理与潜在结果相互独立；以及 (ii)重叠性——没有个体会确定性地接受处理。

Assu设 1 （强可忽略性）。 $Y(1)\ Y(0)\ \perp\!\!\!\perp Z|X$ 且 $e(x)\ <1$ 对所有 $x \in \mathcal{X}$

请注意，其他估计目标对重叠性有不同的要求。例如，对于平均处理效应，我们需要更强的假设：$0 < e(x)\ < 1$ 对所有 $x \in \mathcal{X}$。根据假设 1，我们可以将处理组单位的平均反事实结果识别为

$$\mu_0 = \mathbb{E}[m(0, X) \mid Z = 1] = \mathbb{E}\left[\frac{e(X)}{1 - e(X)} Y \mid Z = 0\right].$$

当认为簇成员身份具有相关性时，分析人员也可能有兴趣估计簇内 g的效应，例如条件平均处理组处理效应（CATT），

$$\tau_g = \mathbb{E}[Y(1) - Y(0) \mid G = g, Z = 1] = \mu_{1g} - \mu_{0g}, \tag{1}$$

其中，我们同样将 $\mu_{1g} \equiv \mathbb{E}[Y(1)\ \mid G = g,\ \ Z = 1]$ 和 $\mu_{0g} \equiv \mathbb{E}[Y(0)\ \mid G = g,\ \ Z = 1]$记为。再次，估计$\mu_{1g}$ 相对直接：我们只需使用簇 g中处理组单位的平均结果。然而，由于混杂，估计μ_{0g} 仍然更加困难。

给定簇成员身份的可忽略性。 存在簇时，我们可以将标准可忽略性假设扩展为包含簇成员身份。具体而言，我们假设处理分配在给定协变量 X 和簇成员身份 G的条件下满足强可忽略性：

假设2 （考虑簇成员身份的强可忽略性）。 $Y(1)\ Y(0)\ \perp\!\!\!\perp Z|X,\ \ G$ 且 $e(X,\ \ G)\ <1$ 对所有 $(x,\ \ g) \in \mathcal{X} \times \{1, \ldots, K\}$

考虑簇成员身份的强可忽略性假设强可忽略性在簇内成立，但不一定在簇间成立。在这种情形下，我们可以将每个簇 g 视为一个独立总体，并针对其开展一项独立的观察性研究，以估计 CATT τ_g ，然后汇总得出总体平均处理效应（ATT） τ。根据这一假设，我们可以识别 μ_0 和 μ_{0g} ：

$$\begin{aligned}
\mu_0 &= \mathbb{E}[m(0, X, G) \mid Z = 1] = \mathbb{E}\left[\frac{e(X, G)}{1 - e(X, G)} Y \mid Z = 0\right], \\
\mu_{0g} &= \mathbb{E}[m(0, X, g) \mid Z = 1, G = g] = \mathbb{E}\left[\frac{e(X, g)}{1 - e(X, g)} Y \mid Z = 0, G = g\right].
\end{aligned} \tag{2}$$

通过对簇成员身份进行条件化，簇可忽略性假设可能比标准可忽略性假设更容易接受，因为它允许处理分配受到未观测到的簇级上下文混杂。例如，考虑比较

两种不同的医疗干预。如果我们以患者接受处理的医院为条件，就只比较同一家医院内接受处理的患者和对照组患者：我们的目标是控制医院中那些有助于决定结果的未观测因素，同时通过比较同一家医院内的患者，将与医院相关的所有医疗护理因素保持不变。

然而，以簇成员身份为条件可能会使重叠假设更难满足，因为这要求在每个簇内的处理组与对照组之间都满足重叠性，而不是在整个研究样本中满足重叠性。此外，在许多实证场景中，分析者会遇到样本量较小的簇，这意味着处理组单位极少或没有，从而导致（近乎）经验重叠违背。在这些场景下，通常无法对这类簇的条件平均处理组处理效应τ_g 进行非参数估计。一种解决办法是从数据中剔除小簇，尤其是在这些簇内没有处理变化的情况下。然而，这会改变目标总体以及相应的估计目标，使其变为某个特定的局部版本的处理组平均处理效应，而该效应可能定义不明确。

给定簇级充分统计量的可忽略性。另一种方法是采用一个更强的假设，使我们能够汇集不同簇的信息，同时仍允许存在簇级混杂。Arkhangelsky 和 Imbens (2024年) 表明，假设 2 等价于假设：给定协变量以及该簇中协变量和处理的经验分布 后，潜在结果与处理条件独立，即：

$$Y(1), Y(0) \perp\!\!\!\perp Z \mid X, P_G,$$

其中，P_g 是 X 和 Z 在簇 g中的经验分布。如果我们对 X 和 Z的分布施加额外假设，这一表征尤其有用。Arkhangelsky 和 Imbens (2024年证明，如果该分布属于一个指数族分布，且其充分统计量为 $S(x,\ z) \in \mathbb{R}^p$，那么假设2 意味着，在给定协变量和簇内充分统计量平均值的条件下，处理分配满足强可忽略性，$\bar{S}_g = \frac{1}{n_g} \sum_{G_i=g} S(X_i,\ Z_i)$。我们可以将这一假设形式化地写为：

假设3 (指数族簇可忽略性)。 $Y(1)\ Y(0) \perp\!\!\!\perp Z \mid X,\ S_G$ 且 $e(X,\ S_G) < 1$ 对所有 $(x,\ \bar{s}) \in \mathcal{X}$均成立。

在此假设下，我们可以识别方程（ μ_0 和μ_{0g} ）中的2，并将根据簇成员身份进行条件化的倾向得分 e（$X,\ G$）替换为 e（$X,\ S_G$），后者改为根据组级充分统计量进行条件化。另请参见He（2018），他提出了一种替代性的识别方法，同样根据簇级统计量进行条件化。在这种方法中，指数族簇可忽略性假设3 允许簇成为“匿名”的，即我们不需要知道某个簇具体属于哪个个体。相反，如果簇具有相似的充分统计量，我们就预期它们与结果的关系也相似。正如我们将看到的，在此假设下，即使簇较小，也可以估计处理组平均处理效应。然而，关键在于，假设3 比假设*2* 更强：虽然假设2 在簇成员身份方面是非参数的，

假设3 还要求簇内 (X, Z) 的联合分布¯属于某个指数族，并且所选的充分统计量 S_g 能够完整捕捉所有相关的簇级信息。如果指数族假设不成立，或充分统计量设定错误，即使假设2 成立，得到的估计值也可能存在偏差。

2.3 估计逆倾向得分加权

给定上述假设以及公式（2）中的识别公式，下一步是估计相应的处理组平均处理效应权重，$\gamma_i = e(x_i, \cdot)/(1 - e(x_i, \cdot))$，其中第二个参数可以是 g_i 或 $\bar{s}_{g_i}$。我们重点介绍估计这些权重的两种常见方法：逆倾向得分加权的传统建模方法，以及近年来广受关注的平衡权重方法。这里，我们简要回顾仅使用个体层面数据时的这些方法；在接下来的章节中，我们将探讨如何扩展这些方法，使其也能调整簇层面数据。

基于模型的逆概率权重。传统的基于模型的逆概率加权方法（Rosenbaum 1987）首先估计倾向得分 $\hat{e}(x_i)$，然后将估计得到的 $\hat{e}$ 代入公式（2）中已知的识别公式。迄今为止，最常见的方法是针对处理分配 Z 关于协变量拟合逻辑回归模型。对于仅含个体层面协变量的情况，该模型为$e(x_i) = \text{logit}^{-1}(\theta \cdot x_i)$，相应的估计倾向得分为 $\hat{e}(x_i) = \text{logit}^{-1}(\hat{\theta} \cdot x_i)$。基于模型的逆概率权重随后为：$\hat{\gamma}_i^{ipw} = \hat{e}(x_i)/(1 - \hat{e}(x_i))$。实践中的一个关键问题是，权重 $\hat{\gamma}_i^{ipw}$ 只能间接平衡协变量 X，并且常常会导致有限样本中的平衡性较差：即，$\sum(1 - Z_i)\hat{\gamma}_i^{ipw} X_i \neq \sum Z_i X_i$ (Chattopadhyay et al. 2020; Ben-Michael and Keele 2023)。正如我们将在下文看到的，当还要寻求平衡簇层面数据时，这一问题往往会进一步加剧。

平衡权重。另一种称为平衡权重的因果推断方法，以及调查研究文献中的校准方法，则致力于直接寻找能够平衡协变量的权重（Hainmueller（2011）；Zubizarreta（2015）；Zhao 和 Percival（2016）；Zhao 2019；Wang and Zubizarreta2020；Chattopadhyay et al.2020）。与传统的基于模型的逆概率权重相比，由此得到的权重通常具有更好的有限样本表现；综述请参见 Ben-Michael 等人(2020)。

这些方法首先评估一般加权估计量的偏差：

$$\text{Bias} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i m(0, X_i) - \frac{1}{n_1}\sum_{i=1}^{n}(1 - Z_i)\gamma_i m(0, X_i). \tag{3}$$

这种偏差是协变量某个特定函数的不平衡：给定协变量时，对照组下潜在结果的期望，$m(0, x)$。平衡权重方法随后会为 $m(0, x)$设定一个（参数或非参数）模型，并通过计算各种潜在模型范围内的最坏情况偏差来构造不平衡度量。最后，这些基于优化的方法会寻找一组权重，使不平衡度量（偏差的替代指标）与权重离散度的度量（方差的替代指标）两者的组合最小化。不平衡度量和离散参数有许多种可能的选择

。根据Ben-Michael 等人（2024a），我们的讨论将主要聚焦于不平衡度量的具体选择。

重要的是，这些平衡权重方法间接地估计倾向得分模型。我们可以利用以下识别结果来说明这一点： $\gamma = e(x)/(1-e(x))$，因此 $e(x) = \gamma/(1+\gamma)$。更正式地说，可以证明，基于优化的平衡权重方法与校准倾向得分估计方法互为对偶（Zhao 2019）。因此，我们可以将平衡权重方法视为使用与常见最大似然估计量略有不同的损失函数来估计倾向得分模型。此外，平衡权重通常优于通过逻辑回归估计的逆概率权重；仅使用权重时，有时也能获得双重稳健性保证；还可以通过结果模型进行去偏，从而获得双重稳健性，并在高维或非参数情形下实现渐近有效推断（Chattopadhyay et al. 2020；Wangand Zubizarreta2020；Zhou etal. 2020；Hirshbergand Wager2021；Ben-Michael and Keele 2023；Ben-Michaeletal. 2022）。

3 全局平衡和局部平衡框架

接下来，我们将分解存在簇时用于估计 ATT 的加权估计量的偏差。正如我们所讨论的，关键挑战在于估计处理组个体的均值反事实结果， $\mu_0 = \mathbb{E}[Y(0) \mid Z=1]$。对于权重 γ_i，我们通过以下方式估计 μ_0

$$\hat{\mu}_0 = \frac{1}{n_1}\sum_{i=1}^{n} \gamma_i Y_i (1 - Z_i).$$

我们可以将估计 μ_0 时的偏差分解为两个组成部分：全局平衡部分和局部平衡部分。为此，我们首先注意到，任何对照下的结果模型 $m(0, x, g)$ 都可以写成

$$m(0,x,g) = \eta_g + \underbrace{\mathbb{E}[Y - \eta_g \mid Z=0, X=x]}_{f(x)} + \underbrace{\mathbb{E}[Y \mid Z=0, X=x, G=g] - \mathbb{E}[Y \mid Z=0, X=x]}_{\delta(x,g)},$$

其中η_g 是簇特定固定效应， $f(x)$是各簇共享的模型共同分量，而 $\delta(x,g)$是簇 g特有的交互分量。借助这一符号表示，我们可以将$\hat{}\mu_0$ 中的偏差（以 X、 G和 Z为条件）分解为三项：

$$\begin{aligned}\text{Bias} = \sum_g \eta_g \underbrace{\left(\frac{1}{n_1}\sum_{G_i=g}(1-Z_i)\gamma_i - \frac{n_{1g}}{n_1}\right)}_{\text{average-to-one-within-cluster}} + \underbrace{\frac{1}{n_1}\sum_{i=1}^{n}(1-Z_i)\gamma_i f(X_i) - \frac{1}{n_1}\sum_{i=1}^{n} Z_i f(X_i)}_{\text{global balance}} \\ + \sum_g \underbrace{\frac{1}{n_1}\sum_{G_i=g}(1-Z_i)\gamma_i \delta(X_i,g) - \frac{1}{n_1}\sum_{G_i=g} Z_i \delta(X_i,g)}_{\text{local balance}}.\end{aligned} \tag{4}$$

第一项源于簇特定固定效应的存在。这可视为偏差中的簇特定截距偏移。若每个簇中对照组单位的权重之和不等于该簇中处理组单位的数量，则该项的存在会导致偏差；因此，该项与归一化密切相关。第二项刻画全局平衡 分量，即加权后结果模型的共同分量 $f(x)$在处理组和对照组单位之间的差异。全局平衡较差意味着处理组和对照组单位跨簇存在系统性差异。也就是说，偏差源于不考虑簇的处理组和对照组单位之间的系统性差异。最后，第三项刻画局部平衡：每个簇 g内处理组和对照组单位之间交互分量 $\delta(x, g)$的差异。若某一特定簇的局部平衡较差，则处理组和对照组单位在该簇内存在系统性差异。

3.1 限制函数以实现平衡

方程（4）在非参数意义上成立，因为它不要求结果模型采用特定的函数形式 $m(0,\ x,\ g)$。尽管如此，考虑结果模型的特定函数形式，或者等价地，指定需要平衡的函数，通常会有所帮助。为此，我们引入一个额外的符号表示。我们用 $\phi(x)$ 表示协变量 x的特征——即 x 的变换——用于处理非线性。可能的变换包括对 x取多项式或协变量之间的交互项。这些特征可以根据先验信息或通过数据驱动方法进行选择（Ben-Michael 等，2024a；Jin 和 Zubizarreta，2025；Shen 等，2025）。

为推进讨论，我们聚焦于这样一种情形：假设结果模型是特征的线性函数 $\phi(x)$。这一表述仍然十分一般，涵盖了广泛的灵活结果模型，包括许多树模型、森林模型和神经网络，以及核和再生核希尔伯特空间回归；其中 $\phi(x)$ 的基可以是无限维的（Ben-Michael 等（2020））。在满足线性假设时，我们可以将具有特征 $\phi(x) \in \mathbb{R}^d$ 的模型写为：

$$m(0, x, g) = \eta_g + (\beta + \delta_g) \cdot \phi(x),$$

其中，η_g 是簇特定固定效应，β 是各簇共有的共同系数，δ_g 是簇特定交互项。

现在，我们可以r 直接用（系数4）表示方程中的偏差分解 -

加权）协变量平衡：

$$\begin{aligned}\text{Bias} = \quad & \sum_g \eta_g \underbrace{\left(\frac{1}{n_1} \sum_{G_i=g} (1 - Z_i)\gamma_i - \frac{n_{1g}}{n_1} \right)}_{\text{average-to-one-within-cluster}} \quad + \quad \beta \cdot \underbrace{\left(\frac{1}{n_1} \sum_{Z_i=0}^{n} \gamma_i \phi(X_i) - \frac{1}{n_1} \sum_{Z_i=1}^{n} \phi(X_i) \right)}_{\text{global balance}} \\ + \quad & \sum_g n_{1g} \delta_g \cdot \underbrace{\left(\frac{1}{n_{1g}} \sum_{G_i=g, Z_i=0} \gamma_i \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g, Z_i=1} \phi(X_i) \right)}_{\text{local balance}}.\end{aligned} \tag{5}$$

现在我们可以看到，全局平衡项是加权后处理组和对照组单位之间的平均特征 $\phi(x)$跨簇差异，而局部平衡项是加权后处理组和对照组单位之间的平均特征 $\phi(x)$在每个簇内差异。因此，在构建旨在控制簇成员身份的加权估计量时，我们希望同时实现处理组和对照组单位在簇之间良好的全局平衡，以及在每个簇内良好的局部平衡。

局部平衡和全局平衡对于估计总体平均处理效应（ATT）的相对重要性，取决于各组之间结果过程的相似程度。在极端情况下，若结果过程在各簇之间没有差异——即 $\delta_g = 0$ 在线性模型中对所有 g 均成立——那么，控制全局平衡并确保每个簇内的权重均值为 1，就足以控制偏差。此外，如果簇级固定效应都相同——即$\eta_g = \eta$ 对所有 g 均成立——那么簇成员身份对结果完全没有影响，这对应于不考虑簇时强可忽略性假设（假设1）成立的情形。在这种情况下，无需控制学校或医院。另一个极端是，结果模型在各簇之间差异很大——例如，$\|\delta_g\|$ 在线性模型中对所有 g 都很大——这意味着结果中有重要部分与学校、医院或其他类型的簇相关。在这种情况下，我们应优先设法控制每个簇内的局部不平衡，以减少偏差。

这种偏差分解为观察性研究中的簇提出了一项总体目标：我们希望在全局和局部层面都消除处理组和对照组单位之间的不平衡。然而，即使有可能在全局层面实现个体的精确平衡，受数据稀疏性影响，在实践中实现局部平衡仍可能很困难。

4 调整簇成员身份的方法

我们首先考虑仅以簇成员身份为条件进行调整，而非针对簇级协变量进行调整。在这种方法下，研究者可以仅以全局平衡、仅以局部平衡，或同时以全局与局部平衡为目标。这里，我们回顾各项策略，并概述常规做法如何归入各类，以阐明其中采用的平衡假设。

4.1 基于模型的逆概率权重：全局与局部平衡

我们首先回顾基于模型的逆概率权重的实现方式，这些方式（间接地）以全局与局部平衡的某种组合为目标。这里重点介绍三种主要方法：纳入簇特异性截距、按簇成员身份分层开展分析，以及纳入随机系数。接下来，我们将讨论平衡权重的对应方法。

还有许多其他逆概率加权方法无法恰当地纳入这一框架。许多论文探讨了在纳入簇效应的同时，采用更灵活的方法估计倾向得分，包括使用 GRF方法、BART、GBM 和潜在类别建模（Kim andSteiner2015；Suk等人 2021；Chang等人 2022；Salditt and Nestler 2023）。另一种思路是，Lee等人（2021）提出将相似的小簇部分合并为较大的组。

仅针对全局平衡：簇特定截距。允许不同簇之间存在水平差异、同时针对全局平衡的主要加权方法，是在倾向得分模型中纳入簇特定截距： $e(x,g) = f(\alpha_g + \theta \cdot \phi(x))$，其中链接函数为 f；参见 Arpino and Mealli (2011)；Thoemmes and West (2011)。回到方程（4）和（5）中的偏差分解，全局方法假定不存在交互分量，即对于所有 x 和 g，$\delta(x,g) = 0$；在线性模型中，这相当于假设 $m(0,x,g) = \eta_g + \beta \cdot \phi(x)$。因此，这种全局方法并不试图在每个簇内平衡个体层面协变量，任何此类不平衡都未得到处理。因此，该方法仅针对偏差的两个方面：(i) 每个簇内的平均值为一项（源自簇特定固定效应）；[1] 以及 (ii) 全局平衡项（源自共同系数）。

在倾向得分模型中实现簇特定截距有两种方法：固定效应和随机效应。 固定效应 方法是在逻辑回归中为每个簇添加一个虚拟变量或固定项。遗憾的是，估计带有簇特定截距的逻辑回归倾向得分模型（例如，纳入 $G-1$ 指示变量）往往会遇到附带参数问题（Neyman and Scott1948；Lancaster 2000），并可能导致有限样本表现不佳。 随机效应 或 随机截距 方法则在层级模型中对组特定截距进行部分汇聚（Arpino 和 Mealli，2011；Thoemmes 和 West，2011）；方差较大的簇，其簇截距会比方差较小的簇更多地向样本均值收缩。[2] 重要的是，这种方法对公式（4）中的簇固定效应进行正则化，因此近似满足每个簇内的平均值为一约束。我们将这种方法称为随机截距逆概率加权（RI-IPW）。这种方法在应用研究中最为常用，也已在文献中得到频繁研究和综述（Li 等，2013；Saldittand Nestler 2023；ChangandStuart 2022）

值得注意的是，包括 纳入簇特定截距并不能实现局部平衡，因为倾向评分模型不包含簇成员身份与个体层面协变量之间的交互项，

[1]未标准化的逻辑回归倾向得分逆概率加权（IPW）仅能在样本量趋于无穷大时弱保证平衡。然而，在每个簇内对权重进行标准化，可确保满足簇内平均值为一约束，并消除固定效应对偏差的贡献。[2]关于其他实现方法，例如，请参见 Yang (2018)，他们采用了这种协变量平衡倾向评分方法。

因此，即使在非常大的样本中，也无法控制簇内的局部不平衡。因此，如果簇内存在严重不平衡，仅采用全局方法将无法控制局部平衡，因而不足以控制偏差。

仅针对局部平衡：分层分析。另一个极端是，我们可以估计一个包含所有簇特异性交互项的倾向得分模型，即 $e(x,\ g) = f(\alpha_g + \zeta_g \cdot \phi(x))$。也就是说，倾向得分模型需要纳入簇固定效应以及每个特征与簇指示变量之间的交互项。这相当于通过完全分层分析，在每个簇内分别估计倾向得分；在实践中，使用逻辑回归倾向得分模型通常不可行，因为这通常会出现“完全分离”（Gelman 等人，2008）。

即使完全交互模型可以估计，这种方法也有两个主要缺点。首先，样本量较小的簇中局部平衡可能较差：有些簇的对照组单位可能很少，或者在某些组内很难实现良好的协变量平衡。其次，这种分析完全忽略了全局平衡：如果每个簇内的局部平衡都较差，就会导致全局平衡较差，即使原则上能够实现良好的全局平衡也是如此。另一方面，在许多应用中，可能存在处理组单位很少或为零的簇。可以舍弃这些簇，但这会改变估计目标，而且未必对应于一个定义明确的目标总体。

同时实现局部平衡与全局平衡：随机系数。最后，我们可以在倾向得分模型中通过随机截距和随机系数对各簇进行部分汇聚，从而同时实现局部平衡与全局平衡：$e(x,\ g) = f(\alpha_g + (\theta + \zeta_g) \cdot \phi(x))$，其中$\alpha_g$和 ζ_g具有层级成分（Arpino 和 Mealli，2011；Thoemmes 和 West，2011）。也就是说，现在每个个体的倾向得分估计都将包含一个特定于协变量的成分，该成分随簇而变化，但会在簇之间进行正则化。

同样，这种方法也有缺点。通常，目标是对所有个体层面的协变量实现局部平衡，这可能需要在模型中加入许多随机系数项。估计大量随机系数项往往在计算上不可行，或需要专用算法（Gao 和 Owen，2020）。另一种做法是只加入少数随机系数项，但限制随机系数的数量意味着只能针对有限的一组协变量实现局部平衡。在实践中，即使样本较大，估计包含单个随机系数的倾向得分模型也可能很困难。此外，对于许多数据集，这些模型难以收敛，甚至无法收敛。

4.2 全局与局部平衡的平衡权重

Ben-Michael 等人（2024a）提出了一种替代性的平衡权重方法，用于估计权重，使其同时实现全局与局部平衡，但允许在这两个目标之间进行权衡。在这一

框架中，目标是找到能够直接控制方程（5）中偏差分解三个项的权重。重要的是，该方法包含多层校准估计量，例如Yang（2018）和Kim 等人（2017）（参见Fuentes 等 2022年的综述），这些估计量采用精确平衡约束，而非我们在此考虑的近似约束。

第一步是为共享共同成分 $f(x)$ 和交互成分 $\delta(x,\ g)$ 指定一个模型类，即对所有 $g,\ f \in \mathcal{F}$ 和 $\delta(\cdot,\ g) \in \Delta$。然后，我们可以找到权重，使这两个模型类的最坏情形偏差最小化。由此衡量的全局与局部平衡项如下：

$$\text{global balance} = \sup_{f\in\mathcal{F}} \left| \frac{1}{n_1} \sum_{Z_i=0}^{n} \gamma_i f(X_i) - \frac{1}{n_1} \sum_{Z_i=1}^{n} f(X_i) \right|,$$

$$\text{local balance}(g) = \sup_{\delta(\cdot,g)\in\Delta} \left| \frac{1}{n_{1g}} \sum_{G_i=g,Z_i=0} \gamma_i \delta(X_i,g) - \frac{1}{n_{1g}} \sum_{G_i=g,Z_i=1} \delta(X_i,g) \right|.$$

为使其具备可操作性，我们可以采用第 3.1 节讨论的线性形式 $\phi(x)$ 模型。为此，我们增加一项约束，要求系数有界：共享共同成分的模型类为 $\mathcal{F} = \{\beta \cdot \phi(x) \mid \|\beta\|_2 \le B\}$，簇特定交互成分的模型类为 $\Delta = \{\delta_g \cdot \phi(x) \mid \|\delta_g\|_2 \le D\}$。由此得到的不平衡项为：

$$\text{global balance} = B \left\| \frac{1}{n_1} \sum_{Z_i=0}^{n} \gamma_i \phi(X_i) - \frac{1}{n_1} \sum_{Z_i=1}^{n} \phi(X_i) \right\|_2,$$

$$\text{local balance}(g) = D \left\| \frac{1}{n_{1g}} \sum_{G_i=g,Z_i=0} \gamma_i \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g,Z_i=1} \phi(X_i) \right\|_2,$$

其中，B 是全局分量系数的最大幅度，D 是交互分量系数的幅度。[3]

通常，我们可能会认为结果模型中的共同分量远大于交互分量，即 $B \gg D$。也就是说，许多局部平衡项可能很小或接近于零。在这种情况下，自然应优先考虑全局平衡项，而非局部平衡项。例如，Ben-Michael 等人(2023年) 和Ben-Michael 等人(2024a找到的权重能够在确保整个样本实现精确全局平衡的同时，优化局部平衡。

[3]这里我们使用了欧几里得范数，但也可以使用其他范数，而且很容易实现。

由此得到以下二次规划目标函数：

$$\begin{aligned}
\min_{\gamma} \quad & \sum_{g=1}^{K} \left\| \frac{1}{n_{1g}} \sum_{G_i=g, Z_i=0} \gamma_i \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g, Z_i=1} \phi(X_i) \right\|_2^2 + \frac{\lambda}{n_{1g}^2} \sum_{G_i=G, Z_i=0} \gamma_i^2 \\
\text{subject to} \quad & \frac{1}{n_1} \sum_{Z_i=0} \gamma_i \phi(X_i) = \frac{1}{n_1} \sum_{Z_i=1} \phi(X_i) \\
& \frac{1}{n_{1g}} \sum_{G_i=G, Z_i=0} \gamma_i = 1, \qquad \gamma_i \geq 0 \quad \forall i = 1, \ldots, n.
\end{aligned} \tag{6}$$

该目标函数会最小化每个簇 g 内的局部不平衡，同时约束权重，使其在整个样本中实现精确的全局平衡，并满足每个簇内的平均值为 1 的约束。此外，还有一个正则化项，用于惩罚较大的权重，进而惩罚高方差权重。正则化程度由超参数 λ控制： λ 值越大，允许的权重复杂度越高，施加的平衡约束越宽松； λ值越小则相反。

这种方法的一大难点是，很难在每个簇内实现良好的局部平衡，尤其是在簇较小时。问题本质上在于维度：纳入交互项后，平衡约束的数量会从 d（协变量的数量）增加到 $d \times K$（协变量数量乘以簇的数量）。也就是说，尽管这种方法确实以有原则的方式针对局部平衡，但要在如此多的项上实现良好平衡，可能并不容易。最后，与我们上文讨论的其他方法一样，小簇中更有可能出现实际的重叠性违反：某些小簇内几乎所有个体可能都接受了处理，或者都未接受处理。同样，解决这一问题的主要办法是从分析中剔除小簇，但这往往并不可取。

5 种调整簇级协变量的方法

另一种方法是利用假设3 ，通过使用簇级充分统计量¯来降低局部平衡约束的维度 S_G ，以取代针对簇成员身份的固定效应。例如，Arkhangelsky 和 Imbens (2024年) 提议对个体层面协变量的簇级均值、每个簇中接受处理的比例，以及这两项指标之间的交互项进行平衡，而不使用簇固定效应。他们将这种方法称为广义 *Mundlak* 估计，以此指代Mundlak (1978) 提出的固定效应估计量。正如我们将看到的，当充分统计量的维度相对于簇的数量较小时，这种方法会减少平衡约束的数量，使实现良好局部平衡更为可行。我们认为，这种方法本质上是对大量簇特定交互项进行正则化的另一种形式。

这很重要　　强调 Mundlak 方法的实际优势是以承担　　e

调用假设3的代价，而该假设严格强于假设2。第4节中对簇成员身份进行调整的方法，在较弱的假设2 下仍然有效，因此无需对簇内协变量的分布或簇级上下文的形式作出假设。然而，由于簇较小，这些方法可能难以实现良好的局部平衡。即使簇较小，广义 Mundlak 方法也能实现更好的局部平衡，但前提是额外的指数族结构成立。研究人员应仔细考虑这一额外假设在其应用中是否合理。

准确刻画充分统计量的函数形式尤为重要 $S(x,\ z)$。包含交互项、非线性变换或高阶项的高维充分统计量会使假设3更弱，因为由此得到的指数族分布将更加灵活，能够涵盖更广泛的情形。然而，正如我们将看到的，更高维的充分统计量也会带来更多局部平衡约束，这在实践中可能使实现良好的局部平衡更加困难。因此，充分统计量 $S(x,\ z)$的选择，与第 $\phi(x$节讨论的特征3.1)的选择类似；类似的数据驱动方法或许也有助于选择这些充分统计量。

5.1 Mundlak 框架

在假设3下，我们考虑给定处理、协变量、¯以及簇级充分统计量的期望结果，$m(z,\ x,\ s)$ $= \mathbb{E}[Y \mid Z = z, X = x, S_G = s]$，并且可以像第3节那样进行类似的分解，得到以下方程：

$$m(0, x, s) = \eta(s) + f(x) + \tilde{\delta}(x, s),$$

其中，$\eta(s)$ 是簇级充分统计量的函数，$f(x)$ 是˜模型在各簇之间共享的共同分量，而 $\delta(x, s)$ 是同时取决于簇级充分统计量和协变量的交互分量。这里，函数 $\eta(s)$ 起到上一节中簇特定固定效应的作用，但现在它是簇级˜充分统计量的函数，而不是每个簇各自独有的函数。同样，交互项 $\delta(x, s)$ 现在是簇级充分统计量的函数，而不是簇成员身份的函数。接下来，我们可以类似于公式（4）分解估计 μ_0 时的偏差：

$$\begin{aligned}\text{Bias} = &\underbrace{\frac{1}{n_1}\sum_{i=1}^{n}(1 - Z_i)\eta(\bar{S}_{G_i})\gamma_i - \frac{1}{n_1}\sum_{i=1}^{n} Z_i\eta(\bar{S}_{G_i})}_{\text{global cluster balance}} + \underbrace{\frac{1}{n_1}\sum_{i=1}^{n}(1 - Z_i)\gamma_i f(X_i) - \frac{1}{n_1}\sum_{i=1}^{n} Z_i f(X_i)}_{\text{global individual balance}} \\ &+ \underbrace{\frac{1}{n_1}\sum_{i=1}^{n}(1 - Z_i)\gamma_i\tilde{\delta}(X_i, \bar{S}_{G_i}) - \frac{1}{n_1}\sum_{i=1}^{n} Z_i\tilde{\delta}(X_i, \bar{S}_{G_i})}_{\text{local balance}}.\end{aligned} \tag{7}$$

将这一分解与方程（4）中的分解进行比较，我们可以看到，簇成员身份 G_i 不再¯直接与偏差相关。相反，簇级充分统计量S_{G_i} 本身就能捕捉

簇级信息。在这种方法中，根据假设3，簇被视为“匿名”的，也就是说，我们无需知道哪些簇包含哪些特定个体。相反，如果各簇在根据簇内所有个体测得的充分统计量方面相似，那么我们预计它们与结果之间的关系也会相似。我们还可以看到只需确保权重在不同簇的处理组和对照组单位之间平衡簇级充分统计量 S_{G_i} ，而无需要求权重在每个簇内的平均值为 1：根据假设 3，除了簇级充分统计量所捕捉到的内容外，簇成员身份并无独特之处。下文将考虑另一种约束，即权重在每个簇内的平均值为 1。

因此，与按簇成员身份进行条件化相比，按簇级充分统计量进行条件化的主要优势在于，我们能够实现更好的局部平衡。交互项 $\delta(x, s)$ 是簇级充分统计量 S_G 的函数，而不是簇成员身份 G。因此，我们可以将个体层面协变量 X 与簇级充分统计量之间的交互 S_G 视为一个新的协变量，并针对这个新的交互协变量实现局部平衡，而不必试图在每个簇内针对所有协变量实现局部平衡。然而，只有在更强的假设3成立时，这种简化才有效；如果充分统计量未能完全捕捉簇级混杂，即使这些项实现了极佳的平衡，仍可能存在偏差。

5.2 Mundlak 平衡权重

我们现在概述一种平衡权重估计量，用于将第5.1节中的 Mundlak 设定付诸实践。这不同于 Arkhangelsky 和 Imbens (2024年) 提出的方法：他们采用基于模型的倾向得分进行估计，并使用纳入结果模型的增广方法来进一步校正偏差。[4]

首先，令 $\psi(x,\ s)$ 表示个体层面协变量 x 与簇级充分统计量 s 之间相关的交互项。然后，我们可以扩展第 4.2 节中的层级平衡权重设置，以平衡 $\psi(x,\ s)$：

$$
\begin{aligned}
\min_{\gamma} \quad & \left\| \frac{1}{n_1} \sum_{i=1}^{n} (1 - Z_i) \gamma_i \psi(X_i, \bar{S}_{G_i}) - \frac{1}{n_1} \sum_{i=1}^{n} Z_i \psi(X_i, \bar{S}_{G_i}) \right\|_2^2 + \frac{\lambda}{n_1^2} \sum_{Z_i=0} \gamma_i^2 \\
\text{subject to} \quad & \frac{1}{n_1} \sum_{Z_i=0} \gamma_i \bar{S}_{G_i} = \frac{1}{n_1} \sum_{Z_i=1} \bar{S}_{G_i} \\
& \frac{1}{n_1} \sum_{Z_i=0} \gamma_i \phi(X_i) = \frac{1}{n_1} \sum_{Z_i=1} \phi(X_i) \\
& \gamma_i \geq 0 \quad \forall i = 1, \ldots, n.
\end{aligned}
\tag{8}
$$

[4]正如我们在上文讨论的，我们也可以拟合包含随机截距或随机系数项的倾向得分模型，不过Arkhangelsky 和 Imbens（2024年）并未明确说明此类步骤。

与方程（6）中的优化问题相比，我们可以看到，个体层面协变量仍然满足精确的全局平衡约束，同时权重上也施加了方差惩罚项。和之前一样， λ 是一个超参数，用于控制偏差降低与权重稳定性之间的权衡。不过，局部平衡项现在是¯个体层面协变量 X_i 与簇级充分统计量 S_{G_i}之间交互作用的函数。当簇内出现接近经验重叠性违背的情况时，这种局部平衡度量尤其有用；小簇中就可能出现这种情况。此外，这种形式的局部平衡不要求每个簇内都达到平衡，因此适用于某些簇中只有对照组单位或处理组单位的情形。因此，无需剔除较小的簇，或处理组或对照组单位较少的簇。

更广泛地说，当簇数量较多且每个簇内的个体较少时，这种局部约束通常更容易满足。如果不采用假设 3，局部平衡意味着存在 $d \times K$ 个平衡约束，其中 d 是 $\phi(x)$ 的维度，而 K 是簇的数量。相比之下，采用假设3后，我们可以将平衡约束的数量减少到 $\psi(x, s)$ 的维度，这通常小于 $d \times K$。这种约束数量的减少是 Mundlak 方法在实践中的主要优势。不过，这种优势也有代价：该方法要求采用更强的假设3，而不仅仅是采用假设2。具体而言，研究者必须假设：(i) 簇内 (X, Z) 的联合分布属于¯指数族；(ii) 所选的充分统计量 S_g 能够捕捉所有簇级混杂。如果其中任一条件不成立，即使直接对簇成员身份进行条件化的方法仍然有效，Mundlak 方法也可能产生有偏估计。

平均值为一约束与全局平衡。 正如我们在第 5.1 节所讨论的，全局平¯衡约束对 S_{G_i} 在方程（8）中的设置，可以控制方程（7）中全局簇平衡项的偏差。不过，只需约束每个簇内的权重平均值为一，也足以达到这一目的。为说明这一点，请注意，我们可以将全局簇平衡项写成与方程（4）中的平均值为一项相同的形式：

$$\frac{1}{n_1}\sum_{i=1}^{n}(1-Z_i)\eta(\bar{S}_{G_i})\gamma_i - \frac{1}{n_1}\sum_{i=1}^{n} Z_i\eta(\bar{S}_{G_i}) = \frac{1}{n_1}\sum_{g} n_{1g}\eta(\bar{S}_g)\left(\frac{1}{n_{1g}}\sum_{G_i=g}(1-Z_i)\gamma_i - 1\right).$$

因此，我们可以考虑对方程（8）中的优化问题作出变体，将针对簇级充分统计量的平衡约束替换为这种簇内平均值为一约束。需要注意的是，这种簇内平均值为一约束通常会施加额外的约束：共有¯个约束——每个簇一个——而不是像方程（ S_{G_i} 8）中那样，针对簇级充分统计量的每个维度各设一个约束 K 。尽管这些额外约束可能是冗余的，但将权重限制为在每个簇内的平均值为一，有助于增强对簇特定项模型设定错误的稳健性¯(S_{G_i})。不过，采用这一方法时，我们仍必须剔除仅包含处理组单位或仅包含对照组单位的簇。在接下来的实证分析中，我们采用这两种方法，分别称为全局平衡 η

（广义平衡）和平均值为一（*AvTO*）。

5.3 估计方法总结

本节最后对内容作简要总结。如上所述，在需要调整簇的影响时，我们可以同时以全局与局部平衡为目标。全局平衡反映所有簇中处理组和对照组单位之间的差异，而局部平衡反映每个簇内处理组和对照组单位之间的差异。最常见的方法是通过随机截距估计倾向得分，但该方法不包含任何局部平衡约束。随机系数可以用于实现局部平衡，但由于计算方面的问题，通常难以实施。Ben-Michael 等人（2024a）提出了一种策略，在确保全局平衡成立的同时，对局部平衡项进行正则化。据我们所知，他们提出的方法是目前唯一一种专门用于直接实现全局与局部平衡的方法。

另一种方法是平衡簇级充分统计量。在这种方法中，我们假设协变量相似的簇与结果之间存在相似的关系。该假设使我们能够克服局部平衡约束中的稀疏性。我们概述了一个平衡权重框架，用于平衡这些簇级充分统计量及其与个体层面特征的交互项。

总结一下关键权衡：以簇成员身份为条件的方法（第 4 节）只需要假设 2，该假设对簇效应采用非参数设定，但在簇较小或存在经验重叠性违反时可能难以奏效。Mundlak 方法（本节）能够更妥善地处理小簇，但需要更强的假设 3，该假设对数据生成过程施加了额外结构。研究者应根据指数族假设是否合理，以及簇大小是否允许直接以簇成员身份为条件，在这两种方法之间作出选择。

下文将开展两项实证研究，回顾如何实施这些方法，并帮助读者理解这些不同选项之间的权衡。

6 统计推断与偏差校正

现在我们转向不确定性量化和统计推断。我们还将介绍偏差校正，即通过结果模型扩展加权估计量。当加权未能实现足够好的平衡时，偏差校正很有用；此外，它也是定义方差估计量所必需的。我们仍然侧重于估计处理组平均处理效应（ATT），并指出，将其扩展到其他估计目标也很直接。

方差估计和偏差校正都依赖于 $\hat{}m_i$，它是对对照条件下预期结果的估计。关键在于，根据识别假设的不同，存在三组不同的协变量集。具体而言，在假设 1 下，$\hat{}m_i = \hat{}m(0,\ X_i)$；在假设 2 下，$\bar{}\hat{}m_i = \hat{}m(0,\ X_i,\ G_i)$；在假设 3 下，$\hat{}m_i = \hat{}m(0,\ X_i,\ S_{G_i})$。在 AVTO 方法中，我们还会在结果模型中纳入簇特定固定效应。使用结果模型

估计值 $\hat{}m_i$，我们可以构造如下偏差校正估计量：

$$\hat{\mu}_0^{\text{bc}} = \frac{1}{n_1}\sum_{Z_i=1}\hat{m}_i + \frac{1}{n_1}\sum_{Z_i=0}\hat{\gamma}_i(Y_i - \hat{m}_i).$$

在附录中，我们表明偏差校正对我们的实证估计值影响甚微，因此这里不再赘述。未来的研究可以更明确地考察在什么条件下，这种偏差校正可被视为具有双重稳健性。

接下来，我们回顾一种同样依赖于 $\hat{}m_i$ 的方差估计量。首先，我们将估计处理效应的平方标准误写作 $V_{\text{diff}} = V_1 + V_0$，其中 V_1 和 V_0 分别表示处理组和对照组内的估计方差。对于处理组，我们可以使用均值的常见方差 $\hat{}V_1 = \frac{1}{n_1^2}\sum_{Z_i=1}(Y_i - \hat{}\mu_1)^2$,，其中 $\hat{}\mu_1$ 是处理组观测结果的简单平均值。对于对照组，这个方差项为 —

$$\hat{V}_0 = \frac{1}{\left(\sum_{Z_i=0}\hat{\gamma}_i\right)^2}\sum_{Z_i=0}\hat{\gamma}_i^2\,(Y_i - \hat{m}_i)^2 = \sum_{Z_i=0}\hat{\gamma}_i^2\,(Y_i - \hat{m}_i)^2\,.$$

$\hat{}$我们沿用Ben-Michael 等人(2024a) 的术语，并称 V_0 为残差化方差估计量（RVE）。Hirshberg 等人(2019年) 和Ben-Michael 等人(2021) 为估计值 $\hat{}\mu_0$ 渐近正态分布以及该方差估计值相合提供了技术条件。关键的是，这些结果要求不平衡所导致的偏差相对于标准误渐近可忽略。具体而言，当仅靠加权无法实现足够好的平衡时，估计量就需要通过结果模型进行额外的偏差校正。[5]有了这些方差估计值，我们便可通过 $\hat{}\,1-\alpha\,\hat{}\mu_1 - \hat{}\mu_0 \pm z_{\alpha/2}\sqrt{\hat{}V_1 + \hat{}V_0}$构造平均处理效应的近似$^{z_{\alpha/2}}$置信区间，其中 $\alpha/2$ 表示标准正态分布的上分位数。

一个自然的问题是，是否应在组层面计算标准误。尽管文献中的一般建议是，应通过聚类稳健方差估计量或聚类自助法来考虑聚类因素（Cafri 等人 2019年），但在这种情况下并无必要。正如 Abadie 等人(2023年) 所讨论的，聚类校正应针对分配处理的数据层级进行。在当前情境中，处理被分配给簇内的个体，因此无需对标准误进行聚类处理。也就是说，如果处理不是分配给簇内的学生或患者，而是分配给簇——学校或医院——那么就有必要进行聚类校正（Ben-Michael 等人 2024b年）。

[5]请注意，此方差估计量针对的是样本内平均处理效应。若要推广到总体平均处理效应，我们可以在 $\hat{}\,V_0$, $\frac{1}{n_1^2}\sum_{Z_i=1}(\hat{}m_i - \hat{}\mu_0)^2$, 中增加一项，用以衡量对照条件下条件期望潜在结果的变异性。不过，与 $\hat{}\,V_1$ 相比，这一额外项通常较小，因此，关于样本的统计结论通常也适用于总体。

7 模拟

现在，我们通过一项旨在回答两个问题的模拟研究来比较这些方法：（1）带有随机截距的基于模型的逆概率加权与直接针对局部平衡的方法相比如何？（2）层级平衡权重与 Mundlak 平衡权重在实际应用中有何不同?

我们采用 Salditt and Nestler（2023年）的数据生成过程。我们从标准正态分布中抽取 10 个个体层面协变量；协变量 X_1、X_3、X_5、X_6、X_8和 X_9 随后被二分化：当其取值为 ≥ 0时设为 1，否则设为 0。我们还从标准正态分布中抽取一个簇层面变量 U_g。个体 e_{ig} 在簇 i 中的倾向得分g 由以下模型生成:

$$e^{\star}_{ig} = P(Z_{ig} = 1 \mid X_{1ig}, \ldots, X_{7ig}, U_g) = \frac{1}{1 + \exp[-f(X_{1ig}, \ldots, X_{7ig}, U_g)]} \tag{9}$$

$$e_{ig} = 0.8 \cdot e^{\star}_{ig} + 0.15 \tag{10}$$

在倾向得分模型中，f 表示以下函数形式:

$$\begin{aligned} f(X_{1ig}, \ldots, X_{7ig}, U_g) = \beta_0 &+ 0.8X_{1ig} - 0.25X_{2ig} + 0.6X_{3ig} - 0.4X_{4ig} - 0.8X_{5ig} - 0.5X_{6ig} + 0.7X_{7ig} \\ &- 0.25X^2_{2ig} - 0.4X^2_{4ig} + 0.7X^2_{7ig} \\ &+ 0.4X_{1ig}X_{3ig} - 0.175X_{2ig}X_{4ig} + 0.3X_{3ig}X_{5ig} \\ &- 0.28X_{4ig}X_{6ig} - 0.4X_{5ig}X_{7ig} + 0.4X_{1ig}X_{6ig} \\ &- 0.175X_{2ig}X_{3ig} + 0.3X_{3ig}X_{4ig} - 0.2X_{4ig}X_{5ig} - 0.4X_{5ig}X_{6ig} + \rho_U U_g \end{aligned} \tag{11}$$

随后，从伯努利分布中抽取观测到的处理分配：$Z_{ig} \sim$ 伯努利分布（e_{ig}）。最后，结果模型为

$$\begin{aligned} Y_{ig} = &-3.85 + Z_{ig} + 0.3X_{1ig} - 0.36X_{2ig} - 0.73X_{3ig} - 0.2X_{4ig} \\ &+ 0.71X_{8ig} - 0.19X_{9ig} + 0.26X_{10ig} + \alpha_U U_g + \varepsilon_{ig} \end{aligned} \tag{12}$$

其中 $\varepsilon_{ig} \sim \mathcal{N}(0,\ 2)$，真实处理效应为 $\tau = -0.4$。协变量发挥着不同的作用：X_1 至 X_4 是混杂因素，X_5 至 X_7 仅预测处理，X_8 至 X_{10} 仅预测结果。每个模拟数据集包含 100 个簇，每个簇约有 50 个个体。

该设计的关键特征是，U_g 充当未观测到的簇层面混杂因素。我们通过设置 $\rho_U \in \{0, 0.25, 0.5\}$ 来改变这种混杂的强度，其中 $\rho_U > 0$时，$\alpha_U = 0.5$。当 $\rho_U = 0$时，不存在簇层面混杂，因此无需使用考虑簇的 方法。当 $\rho_U > 0$时，基于簇的方法应优于忽略簇成员身份的方法。

我们比较四种方法:

- **平衡　　未进行簇调整的权重**。不调整簇的基线　　r

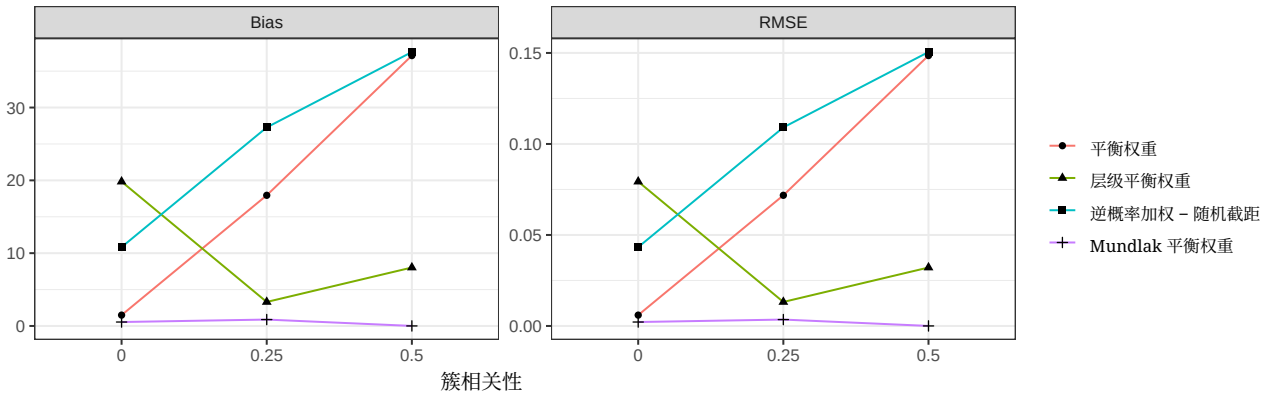

图1：在改变簇混杂程度的情况下，平衡权重、基于随机截距模型的逆概率加权、层级平衡权重和 Mundlak 权重的偏差与均方根误差。

簇。

- **带随机截距的基于模型的逆概率加权（RI-IPW）**。文献中的标准方法。如第4.1节所述，该方法仅针对全局平衡。在实践中，纳入随机系数并不可行。

- **层级平衡权重**。该估计量来自 Ben-Michael 等人 (2024a)，第 4.2 节对此进行了描述；该方法通过以簇成员身份为条件，同时针对全局平衡和局部平衡。

- **Mundlak 平衡权重**。该估计量来自第5.2节，其针对的是簇级充分统计量，而非簇成员身份。

我们使用 R 中的 `balancer` 程序包实现平衡权重（Ben-Michael et al. 2022），并通过以下指标衡量各方法的表现：(1) 标准化绝对偏差，$\left|\mathbb{E}[\widehat{\text{效应}} - \text{效应}]/\mathbb{E}[\text{效应}]\right|$；以及 (2) 均方根误差（RMSE），$\sqrt{\mathbb{E}[(\widehat{\text{效应}} - \text{效应})^2]}$。我们重复模拟 1,000 次。

图 1 展示了结果。当不存在簇级混杂（$\rho_U = 0$）时，层级平衡权重存在一定偏差，且均方根误差最高；RI-IPW 也存在偏差，均方根误差次高。因此，当簇并不重要时，考虑簇会带来一定代价。相比之下，在这种情形下，标准平衡权重和 Mundlak 权重均无偏——标准平衡权重无偏并不意外，但 Mundlak 权重无偏则值得注意。

一旦引入簇层面混杂，标准平衡权重的偏差就会大幅增加，这与预期一致。更令人意外的是，RI-IPW 的偏差也会增加，且实际上超过标准平衡权重的偏差；即使在混杂程度最高时，这种情况仍然存在。在中等程度的混杂下，层级平衡权重与 Mundlak 权重的表现几乎相同。在混杂程度最高时，层级平衡权重的偏差会增加，而 Mundlak 权重仍几乎无偏。

这些结果带来三点启示。第一，RI-IPW 在所有条件下的表现都很差。当簇无关时，出现一些偏差或许在意料之中；但存在簇层面混杂时偏差仍然很高，这一点尤为突出：RI-IPW 的表现甚至不如完全忽略簇的方法。仅关注全局平衡，几乎无法减少簇层面混杂造成的偏差。第二，层级平衡权重能够有效处理簇层面混杂因素，但当簇无关时则要付出代价。第三，Mundlak 权重在所有情境下表现良好：无论是否存在簇层面混杂，它们都几乎无偏，避免了其他基于簇的方法在簇无关时所付出的代价。

8 应用

现在，我们介绍两个簇结构迥异的应用研究结果：一项教育研究考察学前教育项目与考试成绩，另一项卫生服务研究考察急诊普通外科手术。在这两种情况下，个体都嵌套于簇中，研究目标是对簇进行条件化，以捕捉未测量情境。两项应用的簇大小各不相同：教育应用包含许多簇，但每个簇中的个体较少；卫生服务应用也包含许多簇，且簇大小差异很大，其中一些簇相当大。

对于这两项应用，我们比较四种对观测到的混杂因素进行调整的方法，如第7节所述：（1）未进行簇调整的平衡权重；（2）带随机截距的基于模型的逆概率加权；（3）层级平衡权重；以及（4）Mundlak平衡权重。这些比较使我们能够评估簇调整相对于仅进行个体层面平衡的效果，以及相较于 RI-IPW 增加局部平衡约束的价值；RI-IPW 仅强制实现全局平衡，在应用研究中得到广泛使用。我们还将层级平衡权重与 Mundlak 平衡权重之间的一致性作为非正式依据，用于检验较强假设3的合理性：由于 Mundlak 方法要求满足这一额外的指数族结构，而层级方法则不要求，因此这些估计值若存在显著差异，可能表明所选的充分统计量未能充分捕捉簇级混杂。

我们在正文中报告部分结果。对于 Mundlak 方法，我们采用两个版本：全局平衡（*GB*），即对簇级协变量施加全局平衡；以及平均值为一（*AvTO*），即约束每个簇内权重的均值为一。由于我们发现这两种方法之间差异有限，因此在正文中报告 GB 估计值，并将 AvTO 结果放在补充材料中。我们还使用结果模型进行偏差校正后估计了处理效应，但结果仅有细微差异；这些结果也见于补充材料。接下来，我们介绍基函数的设定。首先，$\phi(x)$由个体层面协变量构成，并针对与结果存在非线性关联的连续协变量添加样条拟合。其次，Mundlak 权重所用的基函数更为复杂。这里，我们以 $\phi(x)$、$z\phi(x)$ 和 $\psi(s,\ x)$ 为条件，其中 $\phi(x)$ 是个体层面协变量集合，$z\phi(x)$ 是 $\phi(x)$ 与每个簇中处理组和对照组比例之间的一阶交互项集合，而 $\psi(s,\ x)$ 是 $\phi(x)$ 与 S_{G_i} 之间的一阶交互项集合，其中 S_{G_i} 是个体层面协变量均值的向量

个体层面协变量的均值。

我们使用`balancer` R 库（Ben-Michael et al. 2022）来实现除基于模型的逆概率权重之外的所有方法。我们采用数据驱动方法选择所有超参数值，该方法基于将 Y 对 X 进行回归，且仅限于对照组（$Z = 0$）。该超参数是此回归的残差方差（Ben-Michaelet al. 2022）。对于每个应用，我们估计处理组平均处理效应及相应的95%置信区间。我们还报告$\hat{}\mu_0$ 的有效样本量（ESS）（因为我们的目标是处理组平均处理效应）：平衡更多特征的方法往往有效样本量更低，因而方差更高。

平衡度量。 我们报告若干全局与局部平衡度量。首先，我们使用一种常用度量，即协变量 j的标准化均值差SMD$_j$，以及聚类特定版本SMD$_{g,\ j}$:

$$\mathrm{SMD}_j = \frac{\bar{\phi}_j^{(1)} - \bar{\phi}_j^{(0)}}{s_j} \qquad \mathrm{SMD}_{g,j} = \frac{\bar{\phi}_{g,j}^{(1)} - \bar{\phi}_{g,j}^{(0)}}{s_{g,j}}$$

其中，$\bar{\phi}_j^{(1)}$ 为处理组的均值；$\bar{\phi}_j^{(0)}$ 为对照组的加权均值；s_j为合并标准差 $s_j = \sqrt{\frac{1}{2}\left[\left(s_j^{(1)}\right)^2 + \left(s_j^{(0)}\right)^2\right]}$；$g$ 表示聚类特定版本。我们还计算协变量在全局与局部上的 L_2 不平衡度量:

$$L_2^{\text{global}} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(\mathrm{SMD}_j)^2}, \qquad L_2^{\text{local}} = \sqrt{\frac{1}{PK}\sum_{j=1}^{p}\sum_{g=1}^{K}\mathrm{SMD}_{g,j}^2}$$

我们可以通过计算加权前后该 L_2 指标的百分比降幅，比较各方法降低不平衡程度的效果：我们将其称为偏差百分比降低（PBR）。需要注意的是，我们会计算多个 L_2 值，因为不同方法平衡的是协变量的不同特征，因此结果并不总是可以直接比较。

8.1 教育：小型聚类

第一个应用关注的是学前教育项目对幼儿园后数学测试成绩的影响，相关研究见Dong等人（2020年）和Lee等人（2021）。在这里，学生嵌套于学校之中，我们希望在估计处理效应时考虑学校情境。换言之，分析的目标是利用学校来考虑影响考试成绩的“未测量情境”因素：学校归属应能反映学习过程中的许多方面，而这些方面应与考试成绩相关。

数据来自幼儿园早期儿童纵向研究（ECLS-K），这是一项具有全国代表性的儿童队列研究。在该数据中，共有 9,862 名学生，其中 6,993 名学生参加了学前教育项目，2,869 名学生未参加。按照 Lee等人 (2021) 的做法，我们关注未参加学前教育项目的影响；也就是说，处理组平均处理效应（ATT）是未参加学前教育对未参加学前教育的学生产生的影响。共有 937 所学校作为聚类指标。学生层级基线协变量包括性别、种族、年龄（幼儿园入学时）、

表2：有　效样本量：班级规模应用中五种不同估计量的有效样本量

估计量	有效样本量
未调整	6,993
未进行聚类调整的平衡权重	3,172
含聚类随机截距的基于模型的逆概率加权	4,662
层级平衡权重	1,133
带 GB 约束的 Mundlak 平衡权重	595

身高、体重、家庭类型、家庭社会经济地位、收入、父母教育水平，以及表示学生是否与亲生母亲共同生活的指标。我们关注的结果变量是平均数学测验成绩。这里，簇大小较小，每所学校有 1 至 22 名学生。有 131 个簇不包含任何处理组学生，另有 27 所学校不包含任何对照组学生。我们考察的不同估计策略之间的一个关键区别在于，它们能否纳入所有学生均接受处理或均未接受处理的学校。因此，一个相关问题是，剔除这些簇并改变估计目标是否会导致结果出现实质性差异。剔除没有处理组学生或对照组学生的学校后，剩下 8,703 名学生，分布在 779 所学校中。

表2 报告了每种估计量对应的 $\hat{}\mu_0$ 有效样本量。加权前，对照组学生共有 6,993 名。标准平衡权重和带聚类随机截距的逆概率加权（IPW）的有效样本量较大，分别为 3,172 和 4,662。与之相反，分层平衡权重和 Mundlak 平衡权重得到的有效样本量要小得多，分别为 1,133 和 595。总体而言，仅针对全局平衡的方法，其有效样本量显著高于同时针对局部平衡的方法。这并不意外，因为要平衡的特征集大得多，通常需要变异性更大的权重，从而降低有效样本量。不过请注意，Mundlak 方法平衡的特征少于分层平衡方法（ $\psi(x, s)$ 的维度为 1200，而局部平衡特征的数量为 $d \times K = 18{,}696$）。尽管如此，Mundlak 方法的有效样本量仍然较低。这主要是因为两种方法所用超参数 λ 的尺度和选择不同。

图2 展示了三种聚类调整方法在学校层级、学生层级以及学生 $\times$ 学校交互项所定义协变量上的协变量不平衡。正如预期，随机截距逆概率加权无法平衡学校层面协变量，并且还允许存在相当大的局部不平衡。两种专为簇设计的平衡权重方法在很大程度上减少了这些不平衡，同时也改善了学校层面的协变量平衡。汇总的 L_2 不平衡度量反映了这些定性比较。在完整数据集中（移除任何簇之前），未调整的全局失衡指标为$L_2^{\text{global}} = 0.33$；RI-IPW将其降至$L_2^{\text{global}} = 0.14$；标准平衡权重与层级平衡权重均将其进一步降至 $L_2^{\text{global}} = 0.001$。因此，平衡权重方法实现的全局平衡远优于RI-IPW。相应的未调整的局部失衡指标为 $L_2^{\text{local}} = .26$；层级平衡权重将其降至 $L_2^{\text{local}} = .05$，降幅达80%，而RI-IPW完全没有有意义地改变该指标。因此，相合

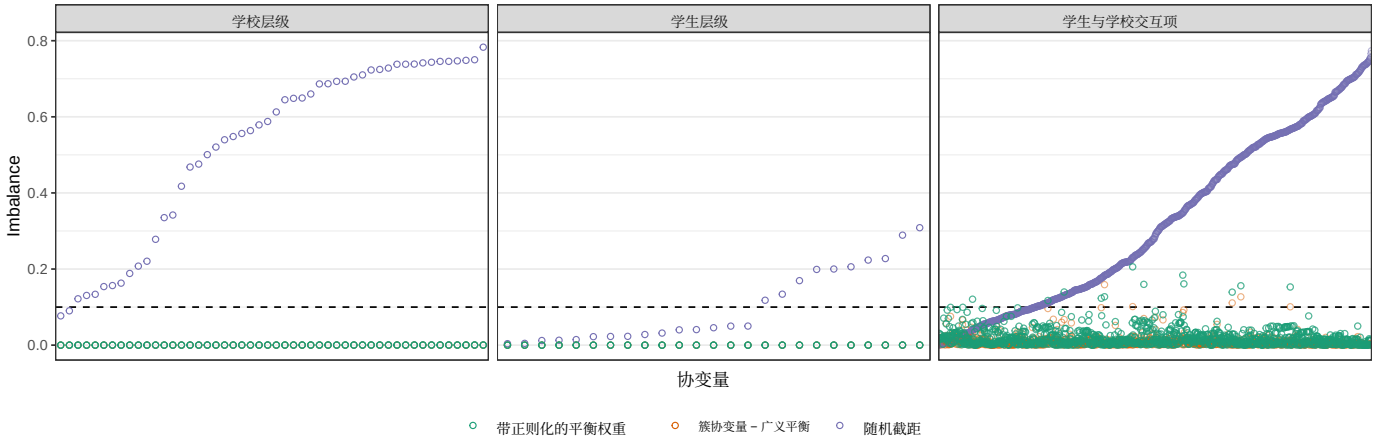

图2：教育应用中三种簇调整方法的平衡度量。

与图4中的结果，我们看到平衡权重方法相对于RI-IPW在协变量平衡方面取得了显著改善。

图3 总结了四种估计方法的结果。未经任何调整时，未参加学前教育项目的学生在幼儿园的数学成绩比参加学前教育项目的同龄学生低 0.5 个标准差，差异相当显著。仅对学生层面的协变量进行调整后，这一差异降至 0.12 个标准差，且置信区间不包含零。通过 RI-IPW 对簇成员身份进行调整后，处理效应为 0.25 个标准差。这里的处理效应似乎相当大，可能反映了 RI-IPW 调整后仍存在的显著不平衡。该估计值与Lee等人(2021）的结果几乎相同。他们采用了不同的聚类协变量处理方法，但同样未纳入任何局部平衡约束。层级平衡权重给出的点估计约为 0.07 个标准差，置信区间也不包含零。最后，Mundlak 平衡权重的估计值同样为 0.07 个标准差，但置信区间包含零，这与表2中报告的有效样本量（ESS）要低得多的情况一致。值得注意的是，在这些方法中，剔除较小的簇似乎影响甚微。

因此，聚类调整方法在实际应用中存在显著差异。RI-IPW 得出的估计值远大于层级平衡权重或 Mundlak 平衡权重的估计值。鉴于图2中平衡权重估计值相近且不平衡程度显著，这一较大的估计值很可能反映了关键协变量未能实现平衡。与此同时，层级平衡权重和 Mundlak 平衡权重得出的估计值几乎完全相同，且远小于其他估计值。这凸显了纳入局部平衡约束或更充分地对聚类协变量进行条件化的重要性，尽管这样做会导致有效样本量较小。层级估计值与 Mundlak 估计值相符，令人安心：请回想，层级平衡权重只要求假设2，而 Mundlak 权重则要求更强的假设3。这些估计值相近，说明本例中 Mundlak 方法所依据的指数族假设并未受到严重违背——尽管这并不能保证该假设完全成立。

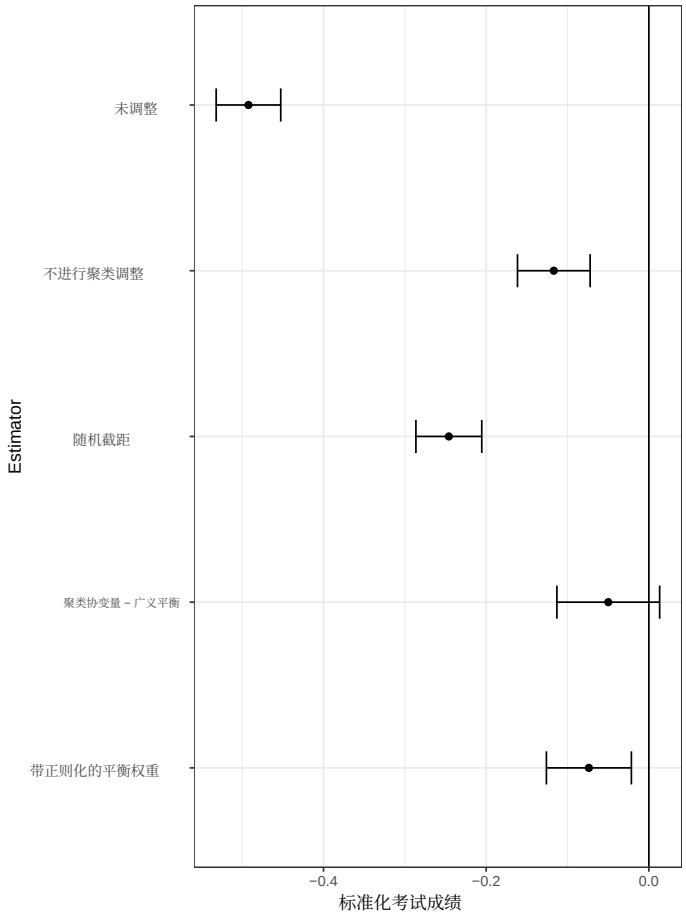

图3：四种不同协变量调整策略下，未参加学前教育项目对考试成绩影响的估计值。

8.2 卫生服务研究：大型簇

急诊普通外科（EGS）是指损伤发生在体内的医疗急症（例如阑尾破裂）；目前共有 51 种疾病被归入这一类别。EGS 疾病每年在美国涉及超过 800,000 台手术，患者估计达 300 万至 400 万（Shafi 等，2013；Gale 等，2014；Havens 等，2015；Ogola 等，2015；Scott 等，2016）。大量研究聚焦于 EGS 疾病手术与非手术治疗的比较有效性，并发现手术治疗的有效性会因 EGS 疾病而异（Hutchings 等，2022年；Kaufman 等，2023年；Moler-Zapata 等，2022年；Grieve 等，2023年）。本文重点讨论手术处理是否能使患有肝胰胆道（HPB）疾病的患者获益；这类疾病会影响肝脏、胰腺和胆道系统。常见的 HPB 疾病包括胆结石、胆囊炎（胆囊的炎症）和胰腺炎（胰腺的炎症）。

我们使用的数据集将美国医学会（AMA）医师主档与纽约州、佛罗里达州和宾夕法尼亚州2012—2013年的全支付方医院出院索赔数据合并。研究总体包括因肝胰胆疾病而接受急诊或紧急住院治疗的所有患者。数据包含少量人口统计学指标，例如种族和族裔、性别、年龄及保险类型等指标。数据集还包含基线患者衰弱程度的指标，以重度脓毒症或脓毒性休克以及既存残疾为表征。此外，数据还包含基于 Elixhauser 指数（Elixhauser 等，1998年）的31种合并症指标，这些指标反映了可能使肝胰胆疾病处理复杂化的既往病症。最后，我们设置了三个指标，分别对应用于治疗肝胰胆病例的三种手术类型。我们也纳入这些指标变量，以控制手术类型差异。主要结果是一个二元指标，用于表示接受任一种处理后是否发生不良事件，即死亡或住院时间延长。

在这一设定中，医院是相关的簇，因为我们担心存在因医院而异的关键未测量照护方面。换言之，我们设想理想的比较是在同一家样本医院内接受手术的患者与未接受手术的患者之间进行，从而保持医院层面的系统性因素不变。在我们的数据中，有143,415名患者嵌套于540家医院之内。各医院内的患者总体规模差异很大，从仅有1名HPB患者到每家医院1,556名患者不等，均值为265名患者。有10家医院没有治疗组患者，另有18家医院的每名患者均接受了治疗。剔除这28家医院会移除176名患者，占全部患者的0.1%。

表 5 给出了估计 $\hat{}\mu_0$的有效样本量。加权前有36,786名对照组患者。未进行聚类调整的平衡权重使有效样本量减半（19,348）；纳入随机医院截距后，大部分有效样本量得以恢复（31,124）。因此，仅以全局平衡为目标对有效样本量的影响不大。相比之下，额外以局部平衡为目标会大幅降低有效样本量，正如上文的教育示例所示：分层权重与Mundlak 权重的有效样本量分别为4,775和6,623。

图 4 展示了三种簇调整方法在医院层级、患者层级及患者

表3: 手术应用中五种不同估计量的有效样本量 n

估计量	有效样本量
未调整	36,786
未进行聚类调整的平衡权重	19,348
含聚类随机截距的基于模型的逆概率加权	31,124
层级平衡权重	4,775
带有 GB 约束的 Mundlak 平衡权重	6,623

层级 × 医院交互项协变量上的不平衡。与教育应用中的情况相似，RI-IPW 无法平衡医院层级协变量，也会造成显著的局部不平衡。专为簇设计的两种平衡权重方法完全消除了这些不平衡，同时也改善了医院层级平衡。

与教育示例一样，汇总后的 L_2 不平衡度量反映了这些定性比较。在完整数据集（移除任何医院之前）中，未调整的全局失衡指标为 $L_2^{\text{global}} = 0.22$；RI-IPW 将其降至 $L_2^{\text{global}} = 0.15$；标准平衡权重和层级平衡权重则再次将其降至 $L_2^{\text{global}} = 0.001$。因此，与 RI-IPW 相比，平衡权重方法实现的全局平衡效果要好得多。相应的未调整局部失衡指标为$L_2^{\text{local}} = .21$；层级平衡权重将其降至 $L_2^{\text{local}} = .02$，降幅达 89%，而 RI-IPW 完全没有实质性改变这一指标。同样，与 RI-IPW 相比，平衡权重方法在协变量平衡方面取得了显著改善。

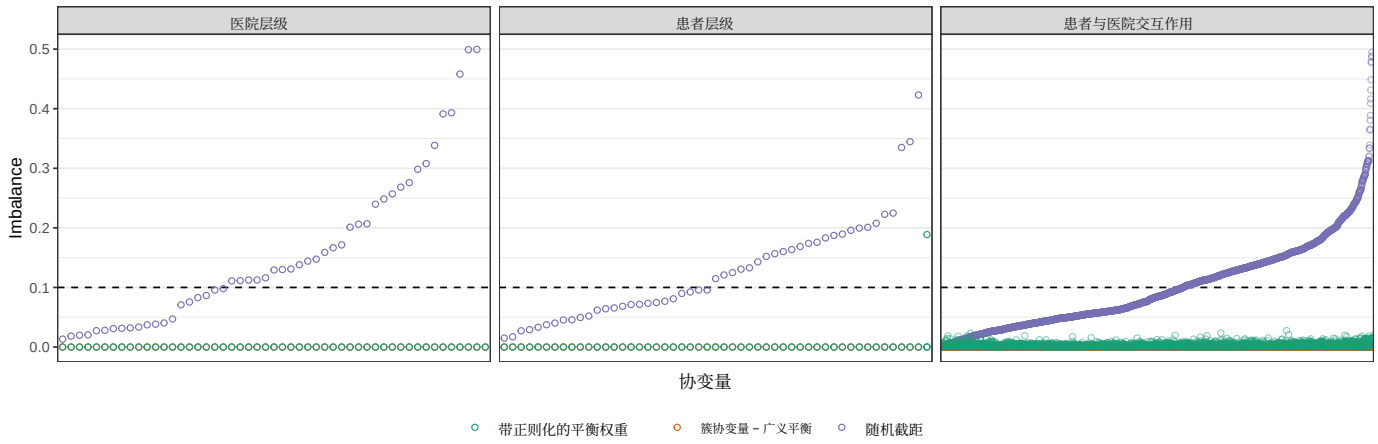

图 4：外科应用中三种簇调整方法的平衡度量。

图5 展示了手术效果的点估计和95%置信区间。未调整的估计值表明，手术治疗可使不良事件的风险降低近9%，尽管如果选择接受手术的患者本身更健康，这一估计很可能存在偏差。事实上，仅对患者协变量进行调整就会使效应方向逆转，表明手术治疗实际上会使不良事件的风险增加约3%，且置信区间完全高于零。接下来，RI- IPW得到的估计值与未调整的估计值相当接近。然而，正如我们观察到的

上文所述，RI-IPW未能实现良好的协变量平衡，因此这一估计值很可能反映了显著偏差。相比之下，层级平衡权重估计的效果基本为零，而且比包含聚类调整的平衡权重估计更接近未包含聚类调整的估计值。最后，在Mundlak估计量中对医院层级充分统计量进行调整，同样得到一个较小且无统计学显著性的估计值。

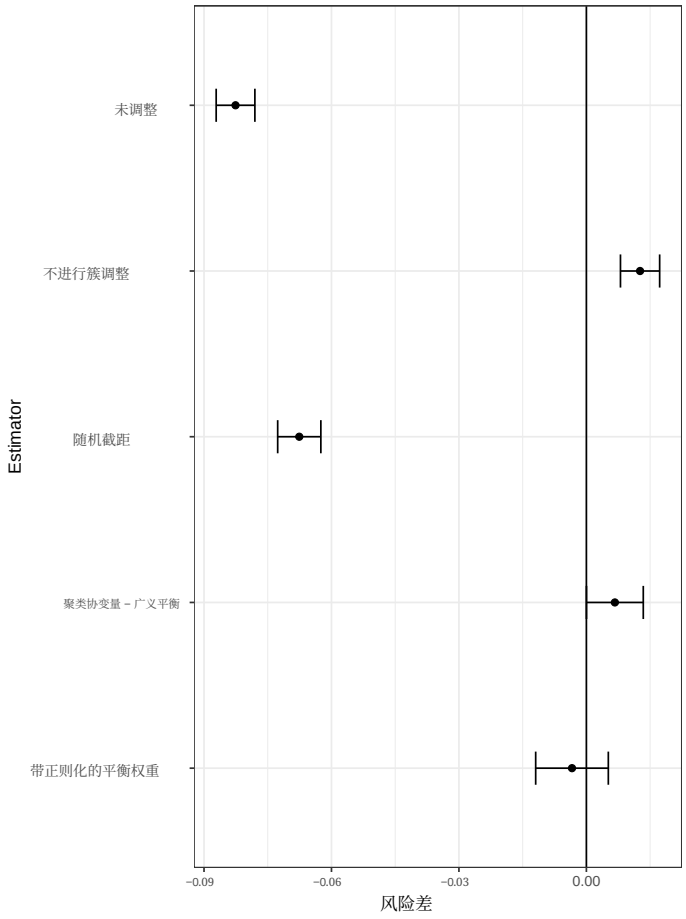

图 5：五种不同协变量调整策略下，手术对不良事件的效果估计值。

比较这些估计值，RI-IPW 很可能给出较差的因果效应估计，存在较大的协变量不平衡，且其估计值相对于其他方法明显为离群值；这与模拟证据和教育应用均相合。接下来，两种聚类调整方法产生了相似的估计值，这表明考虑医院环境很可能是设计中的重要组成部分。我们发现，一旦考虑照护环境，手术治疗对肝胰胆疾病似乎几乎没有益处。相比之下，如果仅平衡患者协变量，我们会发现手术似乎对患者有害。

与教育应用中的情况一样，层级平衡权重与 Mundlak 平衡权重之间的一致性，为假设3提供了一些保证。然而，研究者应注意，这种一致性并不能验证更强的假设；它只是表明，充分统计量

能够捕捉足够的簇级变异，从而得出相似的估计值。在层级方法可行的情形下——例如本研究中，簇大小较大，因此层级方法可行——研究者或许会倾向于使用层级平衡权重，正是因为它们避免了额外的指数族假设。

9 讨论

总的来说，我们的结果表明，研究人员通常应优先选择平衡权重方法，而不是采用层级倾向评分的传统基于模型的逆概率加权。处理分配在簇内存在差异时，RI-IPW 尤其成问题。在可行的情况下，研究人员还应优先考虑同时实现全局与局部平衡的方法，例如层级平衡权重和 Mundlak 平衡权重。当存在簇级协变量且簇大小较小时，后者尤其具有发展前景。然而，研究人员应认识到，Mundlak 方法要求更强的假设3，该假设认为簇级充分统计量能够完全捕捉簇级混杂。当簇大小足够大、可以直接根据簇成员身份进行条件化时，在较弱的假设 2 下，层级平衡权重可能更为合适，因为它们避免了额外的指数族结构。

总的来说，我们建议比较多种方法下的估计值（例如，不进行簇调整、层级平衡权重和 Mundlak 平衡权重）。如果簇调整很重要，那么采用旨在实现局部平衡的方法所得的估计值应彼此相近，并且可能不同于忽略簇或仅旨在实现全局平衡的方法所得的估计值。层级平衡权重与 Mundlak 平衡权重之间的一致性，可以在一定程度上让人放心，认为假设 3并未受到严重违背；如果两者不一致，则可能表明充分统计量未能完全捕捉簇级混杂，或者层级方法未能实现良好的局部平衡。

未来研究有几个重要方向。首先，本文聚焦于加权方法，仅简要讨论了基于结果模型的偏差校正。理解加权与结果建模在这一情境中的相互作用十分重要。此外，我们聚焦于具有聚类数据结构的观察性研究。Arkhangelsky 和 Imbens(2024年)则将这一相同的设定与面板数据情境联系起来；同样，将我们的框架拓展至面板数据估计也是一个有前景的方向。另一项拓展是从研究内平均处理效应（如处理组平均处理效应）进一步延伸至异质性处理效应和可迁移性。最后，我们在此仅关注两个层级。尽管这是迄今最常见的情形，但多层级情形（例如学生嵌套在教室中、教室再嵌套在学校中）会带来许多额外的复杂问题。

A 更多实证结果

此处，我们呈现正文中省略的补充实证结果。我们补充了两组结果。首先，报告 AvTO Mundlak加权方法的结果。其次，报告纳入结果模型增广项的处理效应估计。两种情况下，结果均与正文所报告的结果一致。

A.1 教育：小型聚类

表4：有　　效样本量：教育应用中五种不同估计量的样本量

估计量	有效样本量
未调整	6,993
未进行聚类调整的平衡权重	3,173
含聚类随机截距的基于模型的逆概率加权	4,662
层级平衡权重	1,133
带 GB 约束的 Mundlak平衡权重	595
带 AvTO 约束的 Mundlak平衡权重	609

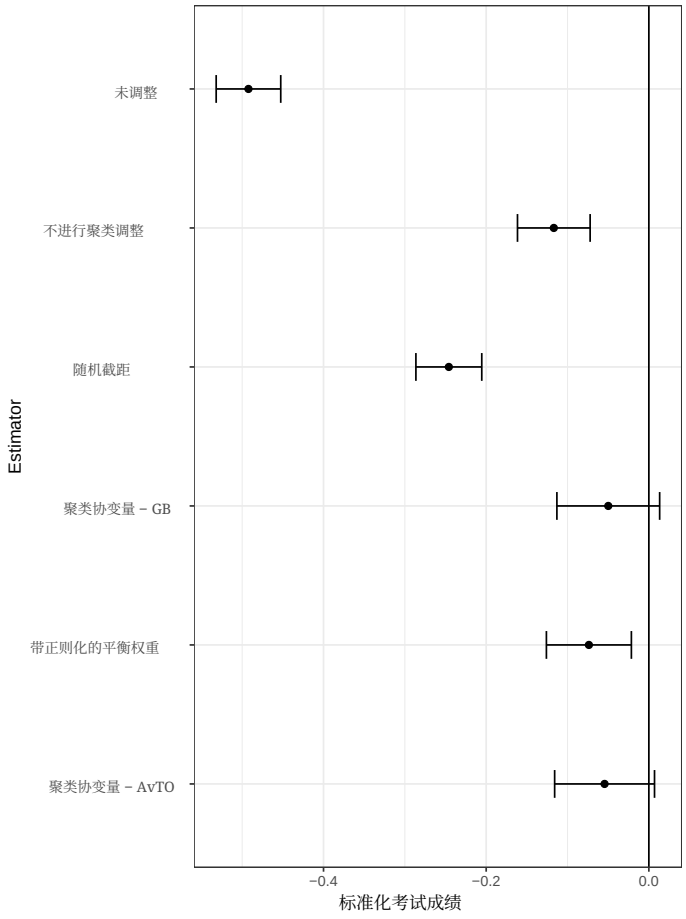

图6：五种不同协变量调整策略对未参加学前教育项目对考试成绩影响的估计值。

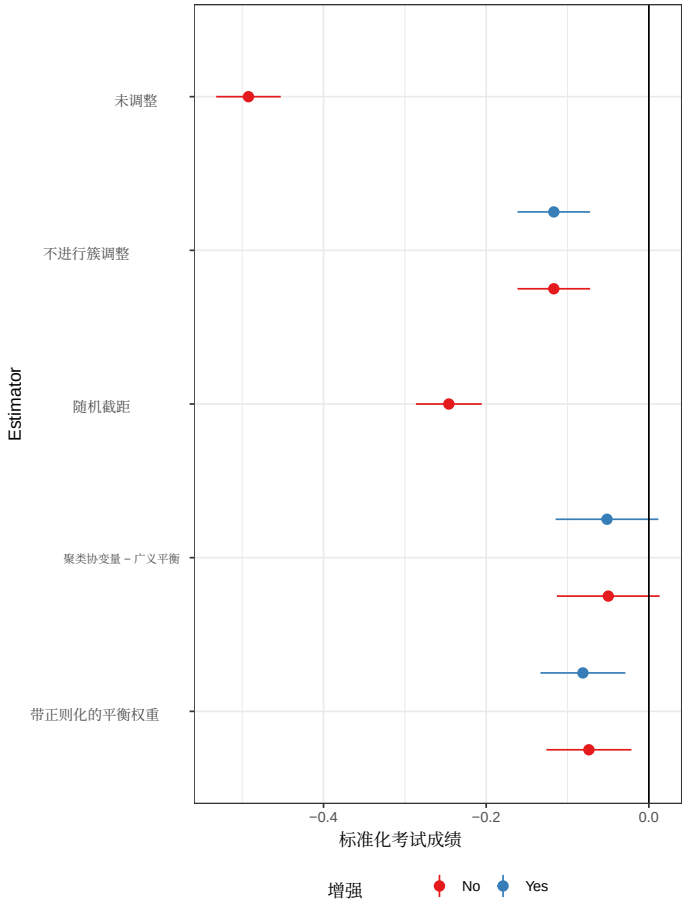

图7：在五种不同的协变量调整策略下，采用和不采用结果模型增强/偏差校正时，不参加学前教育项目对考试成绩影响的估计值。

A.2 手术

表5：外科应用中五种不同的估计量的有效样本量

估计量	有效样本量
未调整	36,786
未进行聚类调整的平衡权重	19,348
含聚类随机截距的基于模型的逆概率加权	31,124
层级平衡权重	4,775
带有 GB 约束的 Mundlak平衡权重	6,623
带有 AvTO 约束的 Mundlak平衡权重	5,677

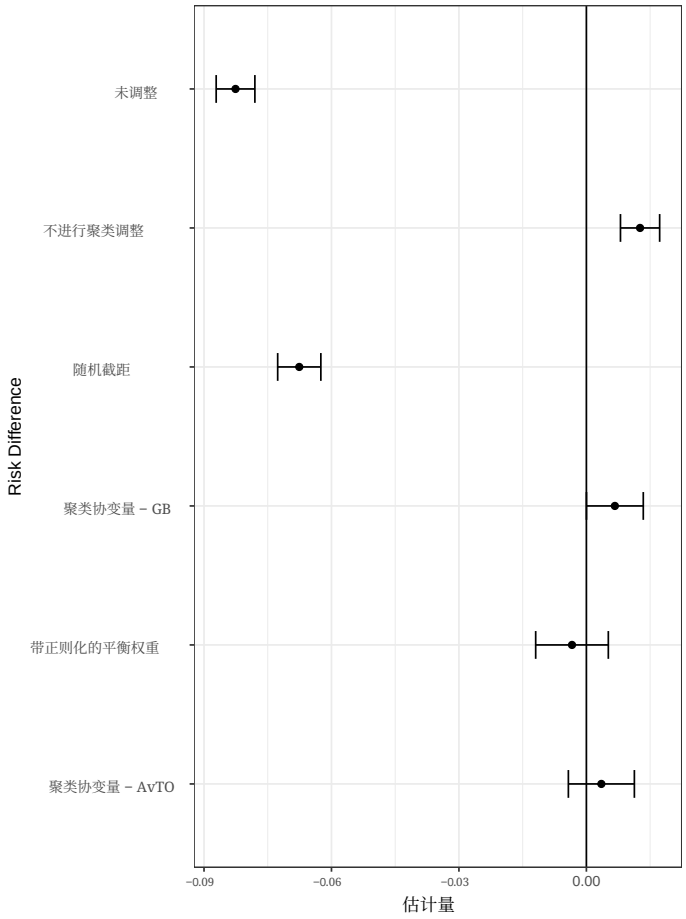

图 8：采用五种不同协变量调整策略时，手术对不良事件影响的估计值。

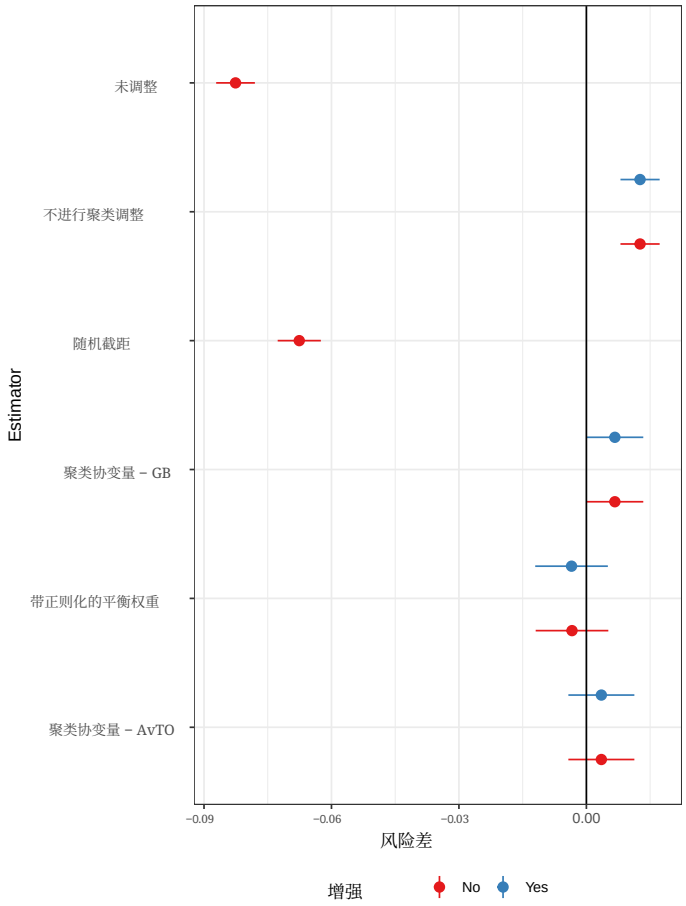

图 9：采用五种不同协变量调整策略，并分别使用和不使用结果模型增强/偏差校正时，手术对不良事件的影响估计值。

参考文献

Abadie, A., Athey, S., Imbens, G. W., 和 Wooldridge, J. M.（2023），《何时应针对聚类调整标准误？》《经济学季刊》，138，1–35。Arkhangelsky, D. 和 Imbens, G. W.（2024），《固定效应与广义 Mundlak 估计量》，《经济研究评论》，91，2545–2571。Arpino, B. 和 Mealli, F.（2011），《多层观察性研究中的倾向得分设定》，《计算统计与数据分析》，55，1770–1780。Ben-Michael, E., Feller, A., Hirshberg, D. A., 和 Zubizarreta, J. R.（2021），《因果推断中的平衡之道》，电子预印本。Ben-Michael, E., Feller, A., Kelz, R., 和 Keele, L.（2022），《急诊普通外科中的种族差异测量》，未发表手稿。—（2024a），《估计急诊普通外科中的种族差异》，《皇家统计学会杂志 *A* 辑：社会统计学》，qnae095。Ben-Michael, E., Feller, A., 和 Rothstein, J.（2023），《推荐信对大学录取的影响各异：观察性研究中亚组效应的近似平衡权重》，《应用统计年刊》，17，2843–2864。Ben-Michael, E., Hirschberg, D., Feller, A., 和 Zubizarreta, J.（2020），《因果推断中的平衡之道》，。Ben-Michael, E. 和 Keele, L.（2023），《重叠性较差时，使用平衡权重以针对处理组平均处理效应》，《流行病学》，34，637–644。Ben-Michael, E., Page, L., 和 Keele, L.（2024b），《聚类观察性研究设计的近似平衡权重》，《医学统计》。Cafri, G., Wang, W., Chan, P. H., 和 Austin, P. C.（2019），《聚类观察性数据因果推断方法的综述与实证比较，并将其应用于医疗器械有效性的评估》，医学研究统计方法，28，3142–3162。Chang, T.-H., Nguyen, T. Q., Lee, Y., Jackson, J. W., 和 Stuart, E. A.（2022），《观察性研究中聚类数据的灵活倾向得分估计策略》，《医学统计》，41，5016–5032。Chang, T.-H. 和 Stuart, E. A.（2022），《聚类数据观察性研究中的倾向得分方法：综述》，《医学统计》，41，3612–3626。Chattopadhyay, A., Christopher H. Hase, 和 Zubizarreta, J. R.（2020），《实践中加权的平衡方法与建模方法之比较》，《医学统计》，即将刊出。Dong, N., Stuart, E. A., Lenis, D., 和 Quynh Nguyen, T.（2020），《使用调查数据的倾向得分分析估计总体平均处理效应：比较不同方法的案例研究》，《评估评论》，44，84–108。Elixhauser, A., Steiner, C., Harris, D. R., 和 Coffey, R. M.（1998），《用于行政数据的合并症测量指标》，《医疗保健》，36，8–27。

Fuentes, A., L¨udtke, O., 和 Robitzsch, A.（2022），《多层数据的因果推断：不同倾向评分加权方法的比较》，多元行为研究，57，916–939。

Gale, S. C., Shafi, S., Dombrovskiy, V. Y., Arumugam, D. 和 Crystal, J. S.（2014），“美国急诊普通外科的公共卫生负担：基于全国住院患者样本的十年分析——2001年至2010年”，创伤与急性护理外科杂志，77，202–208。

Gao, K. 和 Owen, A. B.（2020），“超大型线性混合效应模型的估计与推断”，*Statistica Sinica*，30，1741–1771。

Gelman, A.、Jakulin, A.、Pittau, M. G. 和 Su, Y.-S.（2008），“适用于逻辑回归及其他回归模型的弱信息默认先验分布”，应用统计年刊，2，1360–1383。

Grieve, R.、Hutchings, A.、Moler Zapata, S.、O’Neill, S.、Lugo-Palacios, D. G.、Silverwood, R.、Cromwell, D.、Kircheis, T.、Silver, E.、Snowdon, C. 等（2023），“常见急性胃肠道疾病成人急诊住院患者急诊手术的临床有效性与成本效益：ESORT研究”，健康与社会照护服务研究，11，1–132。

Hainmueller, J.（2011年），“因果效应的熵平衡：一种在观察性研究中生成平衡样本的多变量再加权方法”，政治分析，20，25–46。

Havens, J. M., Peetz, A. B., Do, W.S., Cooper, Z., Kelly, E., Askari, R., Reznor, G. 和 Salim, A.（2015），“急诊普通外科的超额发病与死亡”，创伤与急性护理外科杂志，78，306–311。

He, Z.（2018），“因果推断中聚类数据的逆条件概率加权”，*arXiv* 预印本 *arXiv:1808.01647*。

Hirshberg, D. 和 Wager, S.（2021），“增广极小极大线性估计”，统计学年刊，49，3206–3227。

Hirshberg, D. A., Maleki, A. 和 Zubizarreta, J.（2019），“重定向均值的极小极大线性估计”，arXiv 预印本 arXiv:1901.10296。

Hudgens, M. G. 和 Halloran, M. E.（2008），《针对存在干扰情形的因果推断》，美国统计学会杂志，103，832–842。Hutchings, A.、O’Neill, S.、Lugo-Palacios, D.、Moler Zapata, S.、Silverwood, R.、Cromwell, D.、Keele, L.、Bellingan, G.、Moonesinghe, S.、Smart, N. 等（2022），《五种常见急性疾病的急诊手术疗效：基于全国常规数据库的工具变量分析》，*Anaesthesia*。Jin, Y. 和 Zubizarreta, J.（2025），《用于数据知情设计和观察性研究高效分析的交叉平衡》，。Kaufman, E. J.、Keele, L. J.、Wirtalla, C. J.、Rosen, C. B.、Roberts, S. E.、Mavroudis, C. L.、Reilly, P. M.、Holena, D. N.、McHugh, M. D.、Small, D. 等（2023），《急诊普通外科疾病的手术与非手术结局：采用新型工具变量的观察性研究》，外科学年鉴，278，72–78。

Kim, G.-S.、Paik, M. C. 和 Kim, H.（2017），《簇特异性非忽略分配机制下基于观察性数据的因果推断》，《计算统计与数据分析》，113，88–99。Kim, J.-S. 和 Steiner, P. M.（2015），《用于估计因果效应的多层倾向评分方法：一种潜在类别建模策略》，载于定量心理学研究：心理测量学会第*79*届年会，麦迪逊，威斯康星州，*2014*，Springer，第293–306页。Lancaster, T.（2000），《自1948年以来的附带参数问题》，计量经济学杂志，95，391–413。Lee, Y.、Nguyen, T. Q. 和 Stuart, E. A.（2021），《用于多层数据平均处理效应估计的部分汇聚倾向得分模型》，《皇家统计学会杂志 *A* 辑：社会统计学》，184，1578–1598。Li, F.、Morgan, K. L. 和 Zaslavsky, A. M.（2018），《通过倾向评分加权平衡协变量》，美国统计学会杂志，113，390–400。Li, F.、Zaslavsky, A. M. 和 Landrum, M. B.（2013），《多层数据的倾向评分加权》，《医学统计》，32，3373–3387。Moler-Zapata, S.、Grieve, R.、Lugo-Palacios, D.、Hutchings, A.、Silverwood, R.、Keele, L.、Kircheis, T.、Cromwell, D.、Smart, N.、Hinchliffe, R. 等（2022），《利用电子健康记录时用于处理混杂与异质性的局部工具变量方法：在急诊手术中的应用》，医疗决策，0272989X221100799。Mundlak, Y.（1978），《时间序列与截面数据的汇聚》，《计量经济学》：计量经济学会杂志，69–85。Neyman, J. 和 Scott, E. L.（1948），《基于部分相合观察值的相合估计》，《计量经济学》：计量经济学会杂志，1–32。Ogola, G. O.、Gale, S. C.、Haider, A. 和 Shafi, S.（2015），《急诊普通外科的经济负担：2010年至2060年的全国估计》，创伤与急性护理外科杂志，79，444–448。Rosenbaum, P. R.（1987），《基于模型的直接调整》，美国统计学会杂志，82，387–394。Rosenbaum, P. R. 和 Rubin, D. B.（1983），《倾向评分在观察性研究因果效应分析中的核心作用》，*Biometrika*，70，41–55。Rubin, D. B.（1986），《哪些“如果”能得到因果答案》，美国统计学会杂志，81，961–962。Salditt, M. 和 Nestler, S.（2023），《多层观察性研究中的参数与非参数倾向评分估计》，《医学统计》，42，4147–4176。Scott, J. W.、Olufajo, O. A.、Brat, G. A.、Rose, J. A.、Zogg, C. K.、Haider, A. H.、Salim, A. 和 Havens, J. M.（2016），《利用全国负担界定手术急诊普通外科》，《*JAMA*外科》，151，e160480–e160480。

Shafi, S., Aboutanos, M. B., Agarwal Jr, S., Brown, C. V., Crandall, M., Feliciano, D. V., Guillamondegui, O., Haider, A., Inaba, K., Osler, T. M., 等（2013），“急诊普通外科：定义与疾病负担估算”，创伤与急性护理外科杂志，74，1092–1097。Shen, A. A., Ben-Michael, E., Feller, A., Keele, L., and Murray, J. (2025), “森林核平衡权重：用于因果推断的结果导向特征”，arXiv:2512.11751。Suk, Y., Kang, H., and Kim, J.-S. (2021), “利用随机森林对聚类观察性数据进行因果推断”，多元行为研究，56，829–852。Thoemmes, F. J. and West, S. G. (2011), “倾向得分在含聚类数据的非随机化设计中的应用”，多元行为研究，46，514–543。Wang, Y. and Zubizarreta, J. R. (2020), “最小离散近似平衡权重：渐近性质与实践考量”，*Biometrika*，107，93–105。Yang, S. (2018), “利用倾向评分加权对含聚类数据进行因果推断”，因果推断杂志，6，20170027。Zhao, Q. (2019), “通过定制损失函数实现协变量平衡倾向评分”，统计年鉴，47，965–993。Zhao, Q. and Percival, D. (2016), “熵平衡具有双重稳健性”，因果推断杂志。Zhou, Y., Matsouaka, R. A., and Thomas, L. (2020), “有限重叠与模型误设情况下的倾向评分加权”，医学研究统计方法，29，3721–3756。Zubizarreta, J. R. (2015), “用于不完整结果数据估计的稳定协变量平衡权重”，美国统计学会杂志，110，910–922。