arXiv:2601.04066v1 [stat.ME] 7 Jan 2026

巢式病例对照研究加权分析中纳入概率的估计

Tomeu L´opez-Nieto-Veitch*、RossellaDeSabbata†、Ryung Kim‡、Sven OveSamuelsen¶、NathalieC.Støer‖、Vivian Viallon**

2024年1月8日， 2026††‡‡

摘要

巢式病例对照（NCC）研究是流行病学中广泛采用的研究设计，用于探究暴露因素与疾病之间的关系。本文考察了巢式病例对照研究中的加权分析，重点关注两种主要的加权方法：Kaplan-Meier（KM）权重和广义加性模型（GAM）权重。我们考虑三个目标估计量：对数风险比、条件生存以及暴露因素之间的关联。尽管 KM 权重和 GAM 权重通常具有稳健性，但我们发现，在某些特定情形下，它们可能导致有偏估计。我们证明，当原始队列中有一定比例的个体实际上不符合巢式病例对照研究的选择条件时，尤其是在病例比例较小或匹配因素较多的情况下，KM 权重可能导致有偏估计。相较之下，如果匹配因素之间的交互作用会影响疾病风险，且在权重计算中未得到充分纳入，GAM 权重则可能产生有偏结果。利用有向无环图（DAG），我们构建了一个框架，以系统确定权重计算中应纳入哪些变量。我们表明，变量的最优集合取决于目标估计量，以及匹配因素、暴露因素与疾病风险之间的因果关系。我们利用合成数据和来自欧洲癌症与营养前瞻性调查（EPIC）研究的真实数据阐释了研究结果。此外，我们还将 GAM 权重的应用扩展到“非典型”巢式病例对照研究，即仅纳入部分病例的研究。我们的工作为在巢式病例对照研究中开展准确且稳健的加权分析提供了重要见解。

关键词： 巢式病例对照研究、加权分析、逆概率权重、纳入概率、流行病学、队列研究、EPIC

*博洛尼亚大学统计学系，意大利博洛尼亚。共同第一作者
†布里斯托大学经济学院，英国布里斯托。共同第一作者
‡阿尔伯特·爱因斯坦医学院流行病学与人口健康系，美国纽约州布朗克斯§奥斯陆大学数学系，挪威奥斯陆
¶挪威公共卫生研究所身体健康与老龄化系，挪威奥斯陆
‖挪威癌症登记处研究部，挪威公共卫生研究所，挪威奥斯陆
**国际癌症研究机构营养与代谢处，法国里昂。通讯作者。电子邮件：viallonv@iarc.who.int

††致谢：本研究获得欧盟地平线2020研究与创新计划的资助，资助依据为玛丽·斯克沃多夫斯卡-居里项目资助协议（ESSGN 101073237）。

‡‡国际癌症研究机构免责声明：凡作者被列为国际癌症研究机构/世界卫生组织工作人员，本文所表达的观点均由作者个人负责，并不一定代表国际癌症研究机构/世界卫生组织的决定、政策或观点。

1 引言

巢式病例对照研究（NCC）为完整队列研究提供了一种高效的替代方法，可用于考察暴露因素与疾病风险之间的关系 [1, 2, 3]。例如，在欧洲癌症与营养前瞻性调查（EPIC）研究中，研究人员开展了多项巢式病例对照研究，以评估包括代谢组学在内的各种分子变量与特定部位癌症风险之间的关系 [4, 5]。巢式病例对照研究通常纳入原始队列中在随访期间罹患目标疾病的所有参与者，称为 病例。对于每一例病例， $m \geq 1$ 对照 均从相应风险集中随机抽取；风险集指在病例发生时尚未罹患该疾病的参与者集合。为使对照与病例具有可比性并提高统计效率，通常会根据性别和招募时年龄等相关因素对两者进行匹配 [6]。

在嵌套病例对照研究中，通常使用条件Logistic回归模型评估暴露因素与疾病风险之间的关联；若底层模型为Cox比例风险模型，该模型可得出一致的风险比（HR）估计值 [7, 8]。另一种方法是加权分析，根据参与者被纳入NCC研究的估计概率为其赋予权重 [8]。与条件逻辑回归模型相比，在比例风险模型下，加权分析可以提高统计效率 [8, 9, 10, 11] ，并且在更一般的生存模型下，能够有效估计HR及相关量，包括生存函数 [8, 12]。加权分析还可以估计匹配因素的HR，并分析次要结局 [13]。此外，也可用于估计暴露因素之间的关联，但在这一方面已有一些顾虑 [11]。

已有多种方法用于估计纳入概率并推导相应的权重。Kaplan-Meier权重（KM 权重）最初是为未匹配情形提出的 [8], ，之后扩展至匹配情形 [10] ，以及“非典型”NCC研究；在这类研究中，病例是发生目标事件的受试者中的随机子集 [14, 15]。此外，还有基于模型的方法 [16, 17, 18, 11], ，例如使用广义加性模型（GAMs）推导GAM 权重。其他方法包括局部平均权重 [9], ，但研究表明，这种方法在匹配情形下可能不稳定 [11]。

少数研究比较了基于不同类型权重的加权分析的经验表现，例如 KM 权重和 GAM 权重。除涉及紧密匹配的情况外，通常认为这两类权重会产生相似的结果；在紧密匹配的情况下，KM 权重可能会导致比 GAM 权重更有偏倚的估计 [18]。在本研究中，我们通过强调一些具体情形来补充现有文献：在这些情形下，基于 KM 权重或 GAM 权重的分析都可能产生严重偏倚。首先，我们表明，当原始队列中的部分参与者实际上不符合被选入嵌套病例对照研究的资格时，基于 KM 权重的分析通常会产生偏倚。当病例比例较小或匹配因素数量较多时，通常会出现这种情况，正如第2.2节所示。

其次，当匹配因子之间的交互作用影响疾病风险时，基于 GAM 权重的分析可能会产生有偏估计，除非在估计

纳入概率。

这两个情景中的观测结果促使我们进一步探究计算权重时应考虑哪些变量。我们使用有向无环图（DAG）表示相关变量之间的因果关系，并证明，是否将匹配因子纳入权重计算，取决于目标估计量，以及该因子与暴露因素和疾病风险两者的关系。我们利用 EPIC研究的合成数据和真实数据展示了研究结果，重点关注三个目标估计量：作为暴露因素与疾病风险之间关联度量的对数风险比；暴露因素取某个给定值时的条件生存函数；以及作为暴露因素之间关联度量的线性回归模型参数。最后，我们的另一项贡献是将 GAM 权重扩展至非典型嵌套病例对照研究。

本文其余部分安排如下。第 2 节介绍嵌套病例对照研究的框架，说明典型和非典型嵌套病例对照研究中 KM 权重和 GAM 权重的计算方法，并介绍 EPIC ENDO 研究，该研究为本文提供了研究动机和示例。第 3节使用合成数据，展示上述 KM 权重和 GAM 权重的局限性，以及在计算权重时忽略部分匹配因素的可能性。第 4节利用 EPIC ENDO 研究，比较基于不同类型权重以及计算权重时采用不同匹配因素集合的加权分析。第5节最后作总结。附录提供了更多技术细节。补充材料提供了模拟研究的更多结果，以展示本文提出的适用于非典型嵌套病例对照研究的 GAM 权重的性能；此外，还进一步讨论了如何利用加权分析估计暴露因素之间的关联（或者更一般地说，暴露分布的任意函数）。

2 嵌套病例对照研究的加权分析及示例

2.1 嵌套病例对照研究的加权分析

2.1.1 嵌套病例对照研究

设 Y 表示发生关注事件的时间， C 表示删失时间， L 表示左截断时间。队列中每位受试者 $1 \le j \le n$（队列规模为 $n \ge 1$）均从 L_j随访至 $T_j = \min(Y_j, C_j)$，并将事件指示变量定义为 $D_j = \mathbb{I}(T_j = Y_j)$，其中 $\mathbb{I}(\cdot)$为指示函数。队列中 $D_j = 1$的参与者是在随访期间发生关注事件者，称为病例；而 $D_j = 0$ 的参与者则称为非病例。我们重点研究匹配巢式病例对照研究，并令$\mathbf{M}_j$ 表示受试者 j的 q维匹配因子向量，其中 $q \ge 1$。

令 $K = \sum_{j=1}^{n} D_j$ 表示队列中的事件总数。对于每个事件 $1 \le i \le K$，将对应的风险集记为 $\mathrm{R}_i = \{ j : 1 \le j \le n,\ L_j \le T_i \le T_j \}$。我们将有效风险集 $\mathrm{R}_i^{\mathbf{M}}$定义为 R_i 中匹配因素与 $\mathbf{M}_i$足够接近的受试者子集，可从中为事件 i 抽取 m 匹配对照：一种常见方法是卡钳匹配，此时 $\mathrm{R}_i^{\mathbf{M}} = \mathrm{R}_i \cap \mathrm{M}_i^{\mathbf{M}}$，其中 $\mathrm{M}_i^{\mathbf{M}} = \{ j : 1 \le j \le n : |\mathbf{M}_j - \mathbf{M}_i| \le \varepsilon \}$，容差向量为 $\varepsilon \in \mathbb{R}^q$，且 $\mathbf{m} \le \varepsilon$表示

m 逐分量小于或等于 ε。为便于说明，我们还考虑最近邻匹配，即选择 R_i 中匹配因素与 $\mathbf{M}_i$最接近的 $m = 1$ 名受试者。令 S_j 表示抽样指示变量，使得若受试者 j 在嵌套病例对照研究中被选中，则 $S_j = 1$，否则为 0。进一步令$S_{1,j} = S_j D_j$，这样 $S_{1,j} = 1$ 表示受试者 j在 NCC研究中被选为病例。

嵌套病例对照研究是一种特殊情形：研究中的选择并非完全随机，而是取决于变量 D，T，以及 **M**。然而，以下选择的条件可忽略性条件成立：$S \perp\!\!\!\perp (\mathbf{X}, Y) | (D, T, \mathbf{M})$。因此，对嵌套病例对照研究进行加权分析所得估计的一致性，是以下一般结果的推论，其证明见附录。用 $\mathbb{P}(S = 1|D, T, \mathbf{M})$ 表示个体被纳入嵌套病例对照研究的概率，该概率以其事件状态 D、删失时间 T 和匹配因素 **M** 为条件。假设 $\mathbb{P}(S = 1|D, T, \mathbf{M}) > 0$几乎必然成立，则有

$$\mathbb{E}\left[\frac{S}{\mathbb{P}(S = 1|D, T, \mathbf{M})}\phi(\mathbf{X}, Y)\right] = \mathbb{E}[\phi(\mathbf{X}, Y)], \tag{1}$$

对于任意可积函数 ϕ。计算权重 $1/\mathbb{P}(S = 1|D, T, \mathbf{M})$ 的两种常见策略是 KM 权重和 GAM 权重，以下两段将对此进行介绍。

2.1.2 逆概率权重

KM 权重 首先考虑典型 NCC 的情形，其中所有事件都被选为病例。令 n_i 表示 $\mathcal{R}_i^{\mathbf{M}}$的大小。若受试者 j 未发生事件且属于 $\mathcal{R}_i^{\mathbf{M}}$，则其被选为病例 i 的匹配对照的概率为 $m/(n_i - 1)$。由于针对每个索引病例独立进行匹配对照的随机选择，[8],因此可按以下方式计算受试者 j 的纳入概率：

$$p_{KM,j} = \begin{cases} 1 & \text{if } D_j = 1 \\ 1 - \prod\limits_{i: j \in \mathcal{R}_i^{\mathbf{M}}} \left(1 - \dfrac{m}{n_i - 1}\right) & \text{if } D_j = 0, \end{cases} \tag{2}$$

这是 Kim (2013) [10] 提出的匹配情形扩展。据此，可将 KM 权重定义为 $1/p_{KM,\,j}$。

KM 权重被扩展到非典型嵌套病例对照研究的情形，其中只有比例为 $\pi_1 \le 1$ 的事件被选为病例 [14, 15]。例如，Zhou 等人（2022） [14] 根据以下纳入概率估计值引入了 Horvitz–Thompson (HT) 权重，

$$p_{HT,j} = \begin{cases} \pi_1 + (1 - \pi_1)\left[1 - \prod\limits_{i: j \in \mathcal{R}_i^{\mathbf{M}}} \left(1 - \dfrac{m}{n_i - 1} D_i S_{1,i}\right)\right] & \text{if } D_j = 1 \\ 1 - \prod\limits_{i: j \in \mathcal{R}_i^{\mathbf{M}}} \left(1 - \dfrac{m}{n_i - 1} D_i S_{1,i}\right) & \text{if } D_j = 0. \end{cases} \tag{3}$$

在 $\pi_1 = 1$, $p_{HT,\,j} = p_{KM,\,j}$的特殊情况下，HT 权重会简化为 KM 权重。为

简便起见，下文将 HT 权重和 KM 权重统称为 KM 权重。

GAM 权重 作为 KM 权重的一种替代方法，Støer 和 Samuelsen（2013，2016） [18, 11], 提出了基于模型的方法来估计纳入概率。针对典型 NCC，非病例的抽样指示变量 S 的条件期望可以建模为变量 T、 L 和$\mathbf{M}$的函数，采用带 logit 链接函数的广义线性模型或加性模型，并使用队列中可获得的非病例数据。具体而言，Støer 和 Samuelsen（2016） [11] 建议采用以下广义加性模型设定来表示巢式病例对照研究中的条件纳入概率：

$$\mathbb{P}(S=1|D=0,L,T,\mathbf{M}) = \frac{1}{1+\exp\{-[\alpha + f(L_j) + g(T_j) + h(\mathbf{M}_j)]\}},$$

其中， α_0 为截距项， f、 g 为一些平滑函数，而$h(\mathbf{m}) = \sum_{q=1}^{Q} h^{(q)}(m_q)$，其中函数$h^{(1)}$、.....、$h^{(Q)}$用于对各匹配因子建模。类别匹配变量可以作为分类因子纳入，而卡钳匹配变量可以作为平滑函数纳入。GAM 权重定义为 $1/p_{GAM,\ j}$其中

$$p_{GAM,j} = \begin{cases} 1 & \text{if } D_j = 1 \\ \dfrac{1}{1+\exp\{-[\widehat{\alpha} + \widehat{f}(L_j) + \widehat{g}(T_j) + \widehat{h}(\mathbf{M}_j)]\}} & \text{if } D_j = 0 \end{cases} \tag{4}$$

与 $p_{KM,\ j}$ 不同，后者可视为非参数估计，且等于参与者j在嵌套病例对照研究中被实际选中的概率；$p_{GAM,\ j}$则可视为该概率的平滑估计。

对于非典型嵌套病例对照研究，为了考虑可能不同的入选机制，我们提出分别为事件（$\nu = 1$）和非病例（$\nu = 0$）采用不同的广义加性模型设定，形式如下

$$\mathbb{P}(S=1|D=\nu,L,T,\mathbf{M}) = \frac{1}{1+\exp\{-[\alpha_\nu + f_\nu(L_j) + g_\nu(T_j) + h_\nu(\mathbf{M}_j)]\}}, \quad \nu \in \{0,1\}$$

接着，我们将非典型嵌套病例对照研究的 GAM 权重定义为 $1/\tilde{p}_{GAM,\ j}$，其中

$$\tilde{p}_{GAM,j} = \frac{1}{1+\exp\{-[\widehat{\alpha}_{D_j} + \widehat{f}_{D_j}(L_j) + \widehat{g}_{D_j}(T_j) + \widehat{h}_{D_j}(\mathbf{M}_j)]\}} \tag{5}$$

这些权重分别根据样本中的事件（$D_j = 1$）和非病例（$D_j = 0$）进行估计。补充材料中，我们基于这些权重对非典型嵌套病例对照研究的加权分析进行了实证评估。

2.2 示例说明：EPIC 中一项子宫内膜癌巢式病例对照研究

欧洲癌症与营养前瞻性调查（EPIC）研究是一项大型前瞻性研究，研究对象为1992年至2000年在10个欧洲国家招募的50多万名男性和女性 [4]。该研究最初旨在调查饮食与癌症风险之间的关系。新发癌症病例通过

多种方法识别，包括与健康保险记录和癌症登记系统进行关联。整个队列的相关社会人口学、人体测量学和生活方式变量信息均在招募时收集。约有386,000名参与者在招募时还提供了血液样本，从而可以在EPIC研究内开展的多个嵌套病例对照研究中测量分子数据，包括临床生物标志物、全基因组关联研究和代谢组学数据 [5]。作为示例，我们考察一项聚焦于靶向代谢组学数据的嵌套病例对照研究（使用商业试剂盒 AbsoluteIDQ p180；Biocrates Life Science AG，奥地利因斯布鲁克），这些数据在853名罹患子宫内膜癌的EPIC参与者及853名匹配对照中进行了测量 [19]。这项研究下文称为EPIC-ENDO，聚焦于未接受子宫切除术的女性，并排除了来自希腊、瑞典和丹麦的参与者。共有141,411名女性符合这些条件，构成了抽取嵌套病例对照研究样本的原始队列。在原始队列的141,411名女性中，873名女性在随访期间罹患子宫内膜癌，其中随机选取了853名病例。每个病例均通过风险集抽样随机选取一名对照，并按招募中心、绝经状态（绝经前/围绝经期/绝经后）、招募时年龄（± 6 个月）、采血日期（± 1 月）、采血时间（± 1时）、空腹状态（<3时/3–6时/ >6时）进行匹配；对于绝经前女性，还按月经周期阶段（卵泡期/排卵期/黄体期）进行匹配

EPIC-ENDO 是一个非典型的嵌套病例对照研究，其 $\pi_1 = 853/873 = 0$。977。更重要的是，该研究的事件数较少，而匹配因素较多，因此很大比例的参与者被选中的概率为零。具体而言，原始队列中的 137,979 名参与者既未罹患子宫内膜癌，也不属于 EPIC-ENDO 选取的 853 个病例中的任何一个病例所对应的有效风险集。因此， ~ 97。原始队列中有 6% 实际上不符合 EPIC-ENDO 的选择条件。图1 显示了子宫内膜癌绝对生存概率的估计值， $\mathbb{P}(Y \geq t)$这些估计值分别来自原始队列的未加权分析（实线）、3,432 名符合条件参与者的未加权分析（长虚线），以及对 EPIC-ENDO 中选取的 1,706 名参与者进行的加权分析；后者分别采用 GAM 权重（点线）或 KM 权重（点划线）。实线与点线曲线无法区分，表明基于 GAM 权重的分析是准确的。相比之下，点划线与实线明显偏离，说明基于 KM 权重的加权分析会产生严重偏倚的绝对风险估计。另一方面，点划线与长虚线曲线非常相似，表明基于 KM 权重的加权分析再现了符合条件亚群中观察到的结果。

我们应当记住，加权分析中使用的逆概率权重旨在构建一个能够反映原始队列的伪总体。然而，不符合选择条件的参与者与任何病例或匹配对照都不具备完全相同的匹配特征。从 Kaplan-Meier（KM）权重所依据的非参数方法来看，不符合条件的参与者无法由任何匹配对照代表。因此，所得伪总体排除了不符合条件的个体，与原始队列存在差异。 consequently, KM加权分析只能估计代表符合条件人群的绝对风险。这一点可由图1中点划线和长虚线的高度相似性得到说明。验证这一点的一个简单方法是计算

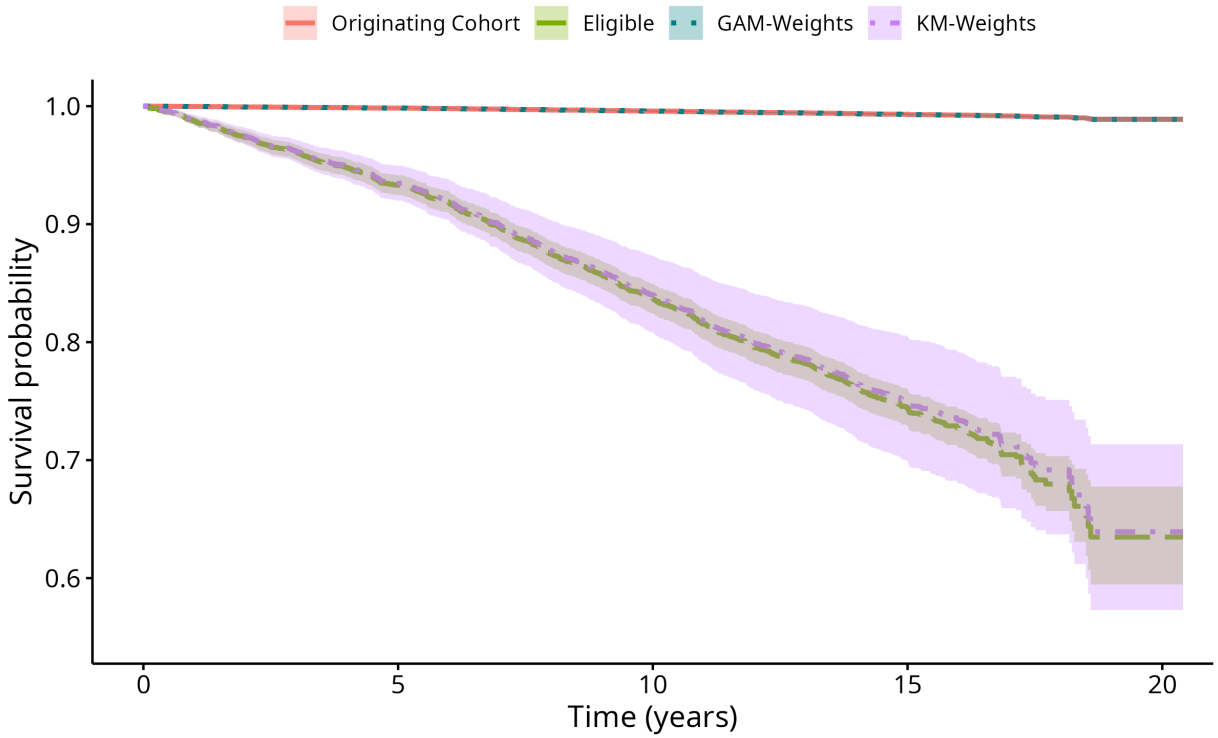

图1： $\mathbb{P}(Y \geq t)$子宫内膜癌绝对生存概率的Kaplan-Meier估计：对由141,411名EPIC参与者组成的原始队列进行未加权分析所得结果（实线）、限制于3,432名符合条件的参与者进行未加权分析所得结果（长虚线），以及对EPIC-ENDO中选取的1,706名参与者进行加权分析所得结果，其中分别使用GAM 权重（点线）或KM 权重（点划线）。在此示例中，我们将研究时间作为时间尺度。

整个NCC研究中KM 权重的总和，此处为3，420，与符合条件的子总体规模相近。相比之下，GAM 权重的总和为137，504，接近原始队列的规模。从用于计算GAM 权重的平滑方法来看，在NCC研究中选中的某些参与者可以作为那些不符合选择条件者的代表，从而使所得伪总体与原始队列一致。KM 权重与GAM 权重之间的这一差异，在采用最近邻匹配这一简单设定的补充图1中得到了体现：KM 权重方法会导致纳入概率出现突变，而GAM 权重方法则产生平滑变化的概率，从而捕捉个体间相似性的渐进变化。

这个说明性示例突出了支撑逆概率权重计算的建模的重要性。它还提出了以下问题，我们将在本文其余部分逐一讨论。首先，很容易认识到，当符合条件的子总体不同于原始队列时，与原始队列相比，该子总体中过度代表了病例，从而导致基于 KM 权重的加权分析对绝对风险产生有偏估计。在第3节中，我们将描述这样的情形：KM 权重也会导致对暴露因素与疾病风险之间关联以及暴露因素之间关联的有偏评估，而 GAM 权重则会产生无偏估计。第二个自然问题是，是否存在 KM 权重不会导致有偏结果、而 GAM 权重会导致有偏结果的情形。

我们将证明，当匹配因素彼此在疾病风险上存在交互作用时，基于方程 (4) 或 (5) 的 GAM 权重通常会导致有偏估计。第三个更一般的问题涉及在估计纳入概率时应考虑哪些变量。利用与证明方程 (1) 的论证相似的论证可知，对于满足 $(X, Y) \perp\!\!\!\perp S|(D, T, M_0)$ 且 $\mathbb{P}(S = 1|D, T, M_0) > 0$ 几乎处处成立的任意变量集 M_0，使用基于 $\mathbb{P}(S = 1|D, T, M_0)$ 的权重都会得到一致估计。具体而言，我们将证明，尽管嵌套病例对照研究中的选择实际上取决于 D、T 以及所有匹配因素 M，但在特定条件下，可以将某些匹配因素排除在权重计算之外。

3 合成数据示例

我们设计了一项全面的模拟研究，以回答上述主要问题，并评估使用不同版本的 GAM 权重和 KM 权重进行加权分析的表现。本模拟研究的总体设计见第 3.1 节。在第 3.2 节中，我们通过最近邻匹配的情形说明基于 KM 权重的加权分析的局限性；在该情形下，原始队列中的部分参与者被纳入的概率为零。在第3.3节中，我们考察匹配变量之间的交互作用会影响事件风险的情形，说明忽略此类交互作用的基于 GAM 权重的分析存在局限性。在第3.4节中，我们呈现模拟结果，以确定在何种情形下可在 GAM 权重和 KM 权重的计算中省略匹配变量。此外，我们还通过探讨非典型嵌套病例对照研究设计中加权分析的表现及其在估计暴露因素之间关联方面的局限性，对这些分析进行了补充。这些额外情形见补充材料。

3.1 模拟研究中考虑的一般框架

3.1.1 数据生成

原始队列：首先，我们根据图 2 所示的生成模型模拟一个包含 $n = 100,000$ 名参与者的队列。具体而言，$M = (M_1, M_2)$ 表示用于选择嵌套病例对照研究的匹配因子向量，其中 M_1 and M_2 分别是均匀变量 U [0, 1]和期望值为 1 2 的二元变量 {0, 1} 经中心化和标准化后的版本。所有情景均使用变量 M_1进行匹配，而仅在第 3.3 节中也使用变量 M_2 进行匹配。暴露 X_a 被建模为一个 3 级分类变量，用以模拟次要等位基因频率（MAF）等于 MAF $=0.25 + \rho_{MX_a}\mathbb{I}(M_1 \leq -1)$ 的单变量 M 病例，以及 MAF $= 0.25 + \rho_{MX_a}\mathbb{I}(M_1 \leq 0)\mathbb{I}(M_2 = 0)$ 的双变量情形下的单核苷酸多态性。换言之，$X_a \in \{0, 1, 2\}$ 的取值概率为 $\mathbb{P}(X_a = 2) = \text{MAF}^2$ 和 $\mathbb{P}(X_a = 1) = 2 \times \text{MAF} \times (1 - \text{MAF})$。暴露 X_b 服从方差为 σ^2、均值为$X_a(\beta + \gamma M_1)$ 的高斯分布生成，其中选择 β、γ和 σ^2，使变量 X_a和M 对 X_b的方差解释比例为 0.1。将 X_a 和 X_b中心化并标准化后，根据 Cox 比例风险模型生成生存时间 Y，

风险模型，其中风险率设为

$$\lambda(t; X_a, X_b, M_1, M_2) = \lambda_0(t)\exp(\alpha_a X_a + \alpha_b X_b + \alpha_{M_1} M_1 + \alpha_{M_2} M_2 + \alpha_{M_1 M_2} M_1 M_2 + \alpha_{M \cdot X_a} M_1 M_2 X_a) \quad (6)$$

参数 α_a、 α_b、 α_{M_1}、 α_{M_2}、 $\alpha_{M_1 M_2}$ 和 $\alpha_{M \cdot X_a}$ 取特定值，基线风险率为 $\lambda_0(t)$ $=3t^2/70^3$, ，对应于形状参数为 $k = 3$ 、尺度参数为 $\lambda = 70$的威布尔分布的风险率。在这一生成机制下，我们可以纳入与暴露和结局存在不同关联的匹配因子，从而评估不同情景下的方法性能和偏倚。根据 ρ_{MX_a}、 γ_{MX_b}、 α_{M_1}、 α_{M_2}、 $\alpha_{M_1 M_2}$ 和 $\alpha_{M \cdot X_a}$的取值， M_1 和 M_2 可能同时影响暴露 X 和结局 Y （记为 W），仅影响结局 Y （记为 V），或两者均不影响（记为 Z），如图8所示。在第3.2节和第3.4节中， $\alpha_{M_1 M_2}$ 和 $\alpha_{M \cdot X_a}$ 均设为 0，且不使用 M_2 进行匹配，因此可以忽略此变量。

删失时间 C 是在特定参数值（u_0， u_1）下按照 $\mathrm{U}[u_0, u_1]$ 分布生成的。结果展示了 $u_0 = 20$ 和 $u_1 = 50$的情形；这些参数取其他值时，对数风险比和条件生存函数的估计结果也类似，因此不再列出[1]。最后，观测时间和事件指示变量分别定义为 $T = \min(Y, C)$ 和 $D = \mathbb{I}(Y = T)$。为简便起见，我们不考虑延迟入组。最终，合成队列包含 $n = 100,000$ 个变量向量（M_1， M_2， X_a， X_b， T， D）的观测值。

嵌套病例对照研究：嵌套病例对照研究从合成队列中按如下方式抽取。事件（$D = 1$）以概率 $\pi_1 = 1$ 被选为病例（典型 NCC），而在非典型嵌套病例对照研究中则以概率 $\pi_1 = 0.5$ 被选中（见补充材料中的第3节）。对于每个选中的病例 i，从风险集 R_i 中随机选取 $m = 1$ 个对照：第3.2节和第3.4节中根据 M_1 进行匹配，第3.3节中则根据（M_1， M_2）进行匹配。第3.2节、第3.3节和第3.4节中，对于 M_1 均采用标准卡钳匹配，容差设为 $\varepsilon = \sqrt{3}/5$，因此对于任意病例 i，集合 $\mathrm{R}_i \cap \{l : |M_{1,l} - M_{1,i}| \le \varepsilon\}$ 的基数平均为风险集 R_i 基数的 20%（因为 M_1 对应于经过中心化和标准化的 $\mathrm{U}[0, 1]$ 变量）。在第3.3节中，我们还对分类匹配因子 M_2 进行精确匹配，因此病例 i 的有效风险集为 $\mathrm{R}_i^{M_1\, M_2} = \{l \in \mathrm{R}_i : |M_{1,l} - M_{1,i}| \le \varepsilon\} \cap \{l \in \mathrm{R}_i : M_{2,l} = M_{2,i}\}$。为说明当部分受试者的纳入概率为零时加权分析的表现，我们还在第3.2节中考虑最近邻（NN）匹配方案，即从˜有效风险集 $\mathrm{S}_i = \mathrm{argmin}_{l \in R_i} |M_{1,l} - M_{1,i}|$ 中选取匹配对照。由于 M_1 是连续变量，在最近邻（NN）匹配方案下，有效风险集 S_i 缩减为单个受试者；对于典型 NCC（$\pi_1 = 1$），所有未被选入嵌套病例对照研究的受试者，其纳入概率均为零。

[1]有关 u_0 和 u_1 ——更一般地说，删失率——对 X_a 与 X_b之间关联估计的影响，详情请参阅补充材料中的第 4.2 节。

3.1.2 加权分析的权重

我们使用 KM 权重和 GAM 权重对嵌套病例对照研究进行加权分析；对于典型设计（$\pi_1 = 1$），其定义见方程（2）和（4）；对于非典型设计（$\pi_1 < 1$），其定义见方程（3）和（5）。在第 3.3 节中，我们还考虑一种 GAM 权重的变体，该变体包含匹配变量之间的交互作用，称为 “含 *Inter M* 的 *GAM* 权重”[2]。

此外，在第 3.4 节中，我们还考虑了忽略匹配因素 $\mathbf{M}$ 的 GAM 权重和 KM 权重。前者由 $P(S = 1 \mid D, T)$的估计值推导得出，而非 $P(S = 1 \mid D, T, M)$，称为“不含 M 的 *GAM* 权重”。后者对应于 KM 权重的一个版本，其中方程（2）和（3）中的 $\mathcal{R}_i^{\mathbf{M}}$ 被 $\mathcal{R}_i$ 替代，在图中标记为“不含 M 的 *KM* 权重”。

3.1.3 评估标准

我们关注三个目标估计量：(*i*)， X_b的对数风险比 α_b ；(*ii*)，随访 u_1结束时的条件生存概率$\mathbb{P}(Y \geq u_1 | X_a = 0)$；以及(*iii*)，在线性回归 X_b 对X_a 的回归中，参数 θ 作为 X_a 与 X_b之间边际关联的度量。在加权分析中，这些估计目标分别通过以下方法估计：(*i*)，使用针对 X_a 和 M[3]调整的 Cox 比例风险模型；(*ii*)，使用在具有 $X_a = 0$的受试者子组中计算的 Kaplan-Meier 估计量；以及(*iii*)，使用（未调整的）线性回归模型回归 X_b 对 X_a。

对于每个目标估计量，我们将未加权完整队列的估计值视为金标准。我们通过计算该参考估计值与嵌套病例对照研究（NCC）样本中相应估计值之间的差异来评估性能。

在每种模拟设定中，我们根据第3.1.1节介绍的特定参数组合，生成 100 项队列研究。从每个队列中抽取一个嵌套病例对照研究样本，并计算完整队列和加权嵌套病例对照研究的估计值。随后，我们根据 100 次模拟中两组估计值之差的观测分布评估性能。此外，我们还报告相对于完整队列估计值的均值、标准差和统计差异，详见补充材料2。

3.2 最近邻匹配与卡钳匹配

我们首先通过最近邻匹配的情形，说明加权分析的估计准确性，以展示当原始队列中的某些受试者的纳入概率为零时，KM 权重可能存在的局限性。我们也考察卡钳匹配的情形，以作比较。我们考虑典型 NCC 的标准设置（$\pi_1 = 1$），并生成仅含一个匹配因子的合成队列 M_1，其中参数 $\alpha_{M_1 M_2}$ 和 $\alpha_{M \cdot X_a}$ 均设为 0。这里，我们给出参数取值为 $\rho_{M X_a} =$

[2]连续匹配变量 M_1 与二元匹配变量 M_2之间的交互作用被建模为一个平滑项 M_1，该项随 M_2的水平而变化。[3]作为比较，我们还使用条件逻辑回归模型 X_a 估计 X_a的对数风险比系数 α_b；研究表明，即使在封闭匹配下，该估计也无偏（例如 [18]）

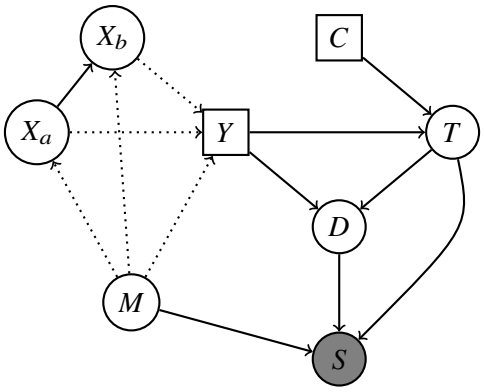

图 2： 我们模拟研究中考虑的数据生成机制。实线箭头表示在模拟研究所考虑的所有情形中始终存在的关系。相反，虚线箭头表示是否存在取决于模型中特定参数取值的关系。右删失变量以方框表示，非删失变量以圆圈表示。变量 S 代表嵌套病例对照研究中的选择，以阴影标出。

0.2，$\gamma_{MX_b} = 0$，且 $\alpha_{M_1} = \alpha_a = \alpha_b =$ 为 log(2) 时的结果。参数取其他值时结果相似，因此不再列出。

图 3 展示了对三个感兴趣估计目标进行估计的结果。采用标准 GAM 权重的加权分析，在两种匹配情形下对所有估计目标均能得出无偏估计。相反，尽管基于卡钳匹配的分析得出的估计无偏，但在最近邻匹配情形下，基于 KM 权重的分析对所有估计目标得出的估计均有偏。这些结果与第 2.2 节的观察结果一致，并证实当队列中的部分受试者纳入概率为零时，KM 权重通常并不适用。再次可以预料，使用 KM 权重进行最近邻匹配时，$\mathbb{P}(Y \geq u_1 | X_a = 0)$ 的估计会有偏，因为所有病例都会被选中，只有不会患上目标疾病的受试者才可能具有为零的纳入概率。因此，符合条件的受试者亚群包含所有事件，但只包含部分非事件；与完整队列相比，可选总体中的 $\mathbb{P}(Y \geq u_1 | X_a = 0)$ 更大。此外，在此处考虑的最近邻匹配情形中，所有被选入嵌套病例对照研究的个体的 KM 权重均为 1，因此，基于 KM 权重对嵌套病例对照研究进行的加权分析，等同于对该研究进行未加权分析；各目标估计量估计中观察到的偏倚则直接源于碰撞偏倚。

3.3 匹配因素与疾病风险之间的交互作用

现在，我们考虑两个匹配因素 M_1 和 M_2，它们彼此之间存在交互作用（M_1M_2），且均与暴露X_a存在交互作用（$M_1M_2X_a$），并共同影响疾病风险，以此说明在存在影响疾病风险的交互作用时，GAM 权重可能存在的局限性。与上一节一样，我们给出参数设定为 $\{\rho_{MX_a} = 0。2，\ \gamma_{MX_b} = 0，\ \alpha_a = \log(2)，\ \alpha_b = \log(2)，\ \alpha_{M_1} = \alpha_{M_2} = 0\}$ 时的结果。参数取其他值时，结果也类似，因此不再列出。为全面考察交互作用的影响，我们考虑 $\alpha_{M1M2} \in \{0，\log(2)\}$ 和 $\alpha_{M \cdot X_a} \in \{0，\log(2)\}$ 的所有组合。

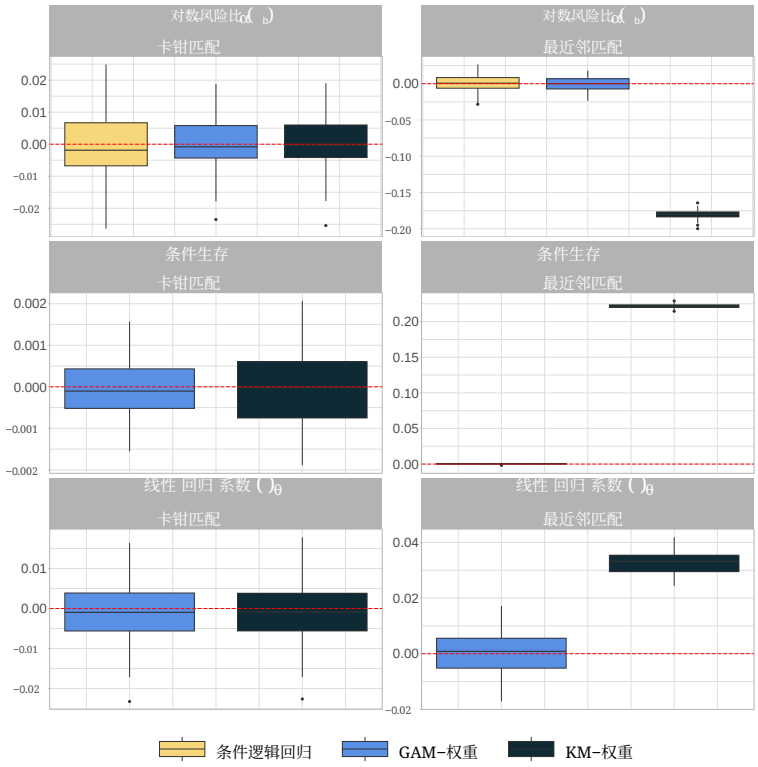

图3： 模拟研究结果：展示在队列中部分受试者的纳入概率为零时，基于 KM 权重的分析所产生的偏倚。每个箱线图表示在 100 次重复中，某一种特定的嵌套病例对照研究分析所得估计值与完整队列所得估计值之差的观测分布。

在图 4中，我们展示了：只要匹配因素之间的交互作用影响生存时间，标准 GAM 权重就会对所有三个关注的估计目标产生显著有偏估计。值得注意的是，即使匹配因素之间的交互作用不影响生存时间，只要匹配因素与单核苷酸多态性之间存在交互作用（$\alpha_{M\cdot X_a} = 0$。69），标准 GAM 权重对条件生存的估计仍存在严重偏倚。相比之下，KM 权重以及纳入交互作用的 GAM 权重在所有设定下均能产生无偏估计。

3.4 何时可以在权重计算中排除匹配因素

现在我们将注意力转向这样一些情形：在计算权重时可以忽略匹配因素。我们再次考虑典型嵌套病例对照研究的标准设定（$\pi_1 = 1$），并生成仅包含一个匹配因子 M 的合成队列（$= M_1$），其中参数 $\alpha_{M_1M_2}$ 和 $\alpha_{M\cdot X_a}$ 均设为 0。我们考虑三种设定，分别为 $\{\rho_{MX_a} = 0.2, \alpha_{M_1} = \log(2)\}$、$\{\rho_{MX_a} = 0, \alpha_{M_1} = \log(2)\}$ 和 $\{\rho_{MX_a} = 0, \alpha_{M_1} = 0\}$；它们分别对应图8中 M 的作用类似于 W（即同时影响 X 和 $\mathcal{Y}$）、V（即仅影响 Y）以及 Z（即既不影响 X 也不影响 $\mathcal{Y}$）的情形。图5展示了参数取值为 $\gamma_{MX_b} = 0$ 和 $\alpha_a = \alpha_b = \log(2)$ 时的结果。参数取其他值时结果类似，因此不再赘述。

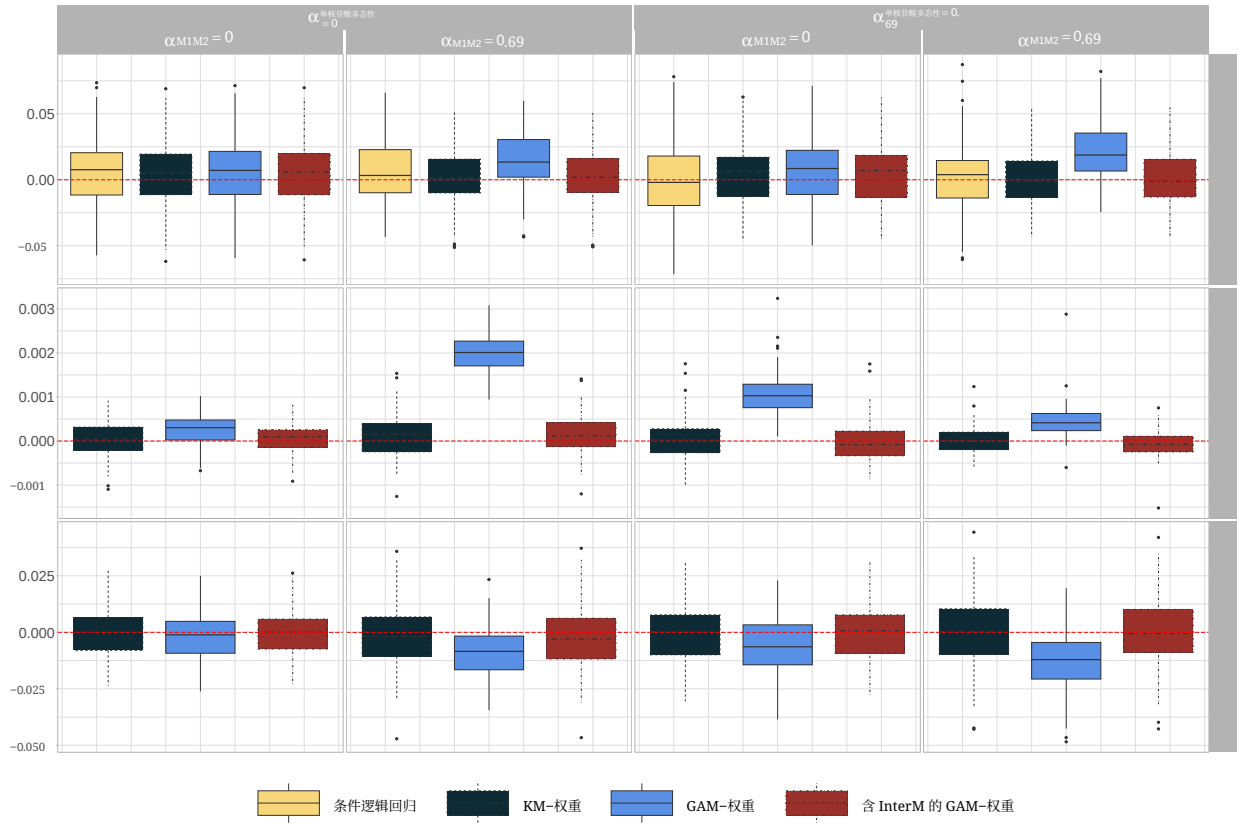

图 4： 纳入匹配因素对疾病风险的交互作用的模拟研究结果。每个箱线图展示了基于 100 次模拟重复，从各种 NCC 分析得出的估计值与相应完整队列估计值之间差异的观测分布。

结果与附录 A 中给出的形式推导一致。首先看 GAM 权重：基于标准 GAM 权重（即计算时纳入 M）进行的加权分析，在所有考察的设定下，对所有估计目标均得到无偏估计。计算 GAM 权重时忽略 M，则当 M ~ W 时，对所有估计目标得到的估计均有偏；当 M ~ Z 时，估计则均无偏。当 M ~ V 时，对数风险比和条件生存函数的估计有偏，而 θ（即 X_a 与 X_b之间的关联）的估计无偏，且经验方差小于使用标准 GAM 权重时所得的估计。基于 KM 权重的分析也观察到相同模式：忽略 M 会导致 M ~ W 时估计有偏、M ~ Z 时估计无偏；当 M ~ V 时，对数风险比和条件生存函数的估计有偏，而 θ 的估计无偏。当 M ~ V 时，使用忽略 M 的 KM 权重得到的 θ 估计，其经验方差小于使用标准 KM 权重所得的估计。

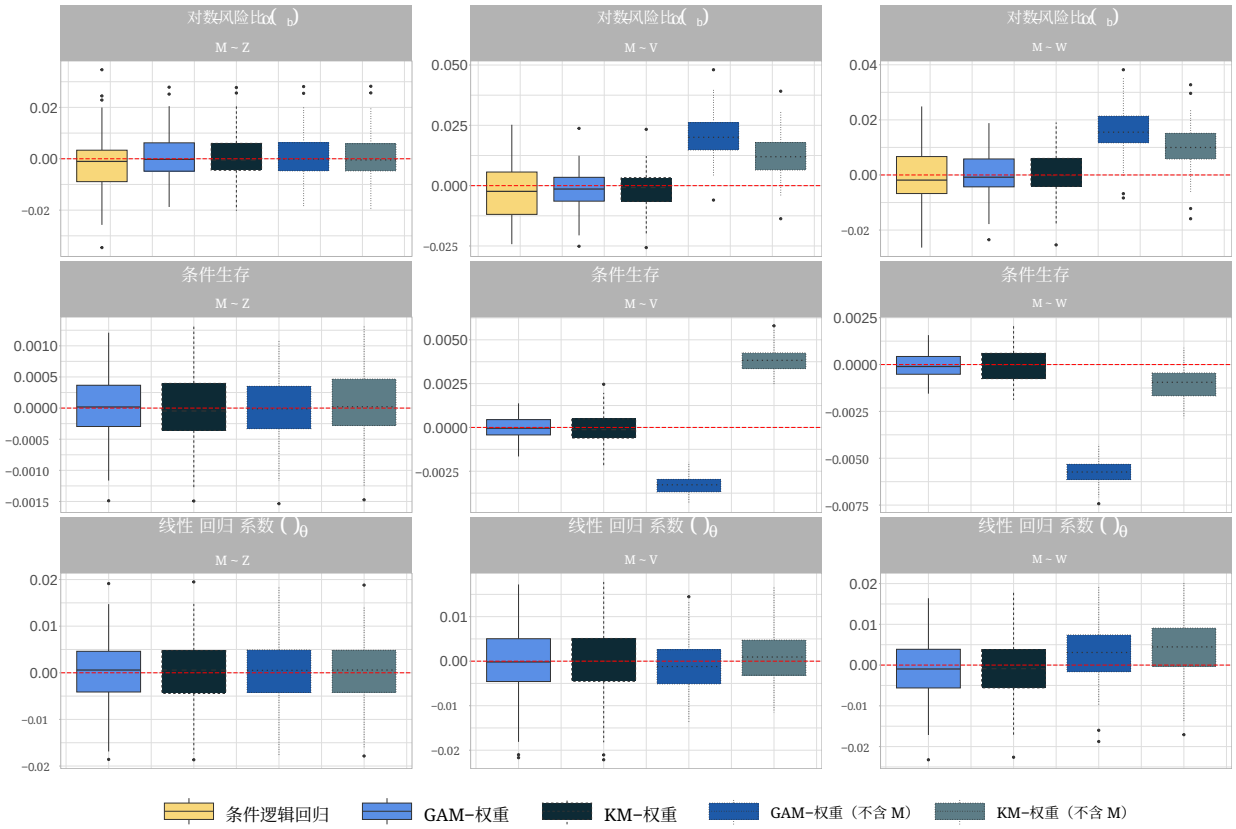

图5： 模拟研究的结果表明，在计算 KM 权重和 GAM 权重时，可以忽略某些匹配因素。每个箱线图表示在 100 次重复中，针对嵌套病例对照研究采用某一种特定分析方法得到的估计值与基于完整队列得到的估计值之差的观测分布。

4 EPIC ENDO 研究示例

现在，我们进一步展示 EPIC-ENDO 研究加权分析的结果。首先，我们提出，可将 EPIC 中使用的匹配因素分为两类：(*i*) 因素 **W**，例如招募中心和绝经状态，可能同时影响代谢物水平**X** 和结局 Y；以及 (*ii*)“技术性”因素**Z**，例如空腹状态，只影响测量过程，且假定既不影响**X**，也不影响 Y。

图6 展示了 EPIC-ENDO 中对应这一结构的有向无环图。变量**X** 表示代谢物的“潜在”或具有生物学相关性的血液水平，这些水平是˜结局 Y的潜在因果决定因素。相反，观测变量**X** 表示实际测得的代谢物水平，这些水平可能受到**Z**中的技术因素影响。由于慢性病风险更可能与˜持续的代谢特征相关，而非短期波动，**X** 不太可能是 Y的原因，而**X** 则可能是。接下来，我们认为**Z**中的变量不会影响潜在暴露**X**，也不太可能影响结局 Y。例如，虽然空腹状态可能会暂时改变代谢物的测量浓度——例如，导致非空腹个体的代谢物水平较高——但它不太可能影响长期

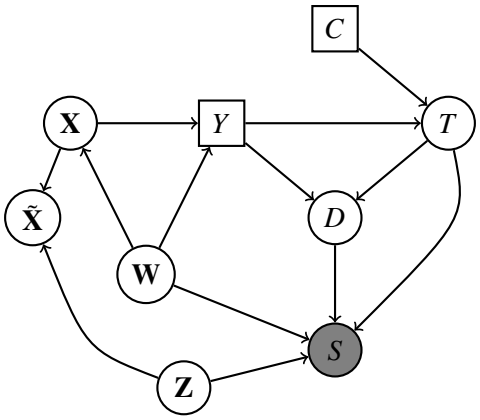

图6： EPIC队列中数据生成机制的示意图。**X**中的变量对应代谢物真实的“潜在”血液水平，这些水平可能是结局的原因 Y ~相反，**X** 中的变量对应测得的代谢物血液水平。这些测量值会受到“技术变量”的影响，例如空腹状态和采血时间。但这些技术变量不会影响真实的“潜在”血液水平**X**。EPIC中使用的匹配因素可分为两类：**W**，例如招募中心和绝经状态，这些因素预计会同时影响**X** 和 Y，以及纯粹的“技术”因素**Z**，这些因素既不影响**X**，也不影响结局 Y。右删失变量以方框表示，非删失变量以圆圈表示。变量 S代表嵌套病例对照研究中的选择，以阴影标出。

代谢物**X**的平均血液水平或具有生物学意义的血液水平。

因此，我们考虑四类权重：方程（3）和（5）中定义的 KM 权重和 GAM 权重，以及排除技术匹配因素（具体而言，即采血时间和日期、空腹状态及月经周期阶段）的变体。在下文的图表和正文中，这些变体标记为不含技术 M的 KM 权重和 GAM 权重。由于在原始队列中，匹配因子与癌症风险之间的交互作用在统计学上均不显著，因此我们在计算 GAM 权重时不纳入交互项。

图 6 补充材料中的内容展示了EPIC-ENDO对照组中权重的经验分布和两两相关性。权重的大小差异显著，且相关性较低，唯有两种基于GAM的方法所得到的权重几乎完全相同。正如预期，KM 权重通常小于排除技术因素后得到的对应权重。GAM 权重总体上最大，部分参与者的权重值超过1000。

接下来，我们评估使用不同类型权重进行加权分析所得的结果，重点关注以下三个目标估计量：子宫内膜癌的（无条件）绝对风险、衡量代谢物与子宫内膜癌之间关联的对数风险比（log-HR）（针对相关协变量进行调整），以及暴露因素之间的关联。补充材料中的图7 展示了子宫内膜癌绝对生存概率的估计值。正如第2.2节所述，基于 KM 权重的估计值存在严重偏倚，而使用 GAM 权重得出的估计表现良好。剔除技术匹配因素可改善 KM 权重估计值，但仍存在一定偏倚。相比之下，剔除这些因素几乎不会影响 GAM 权重估计值。

接下来，图 7 比较了条件逻辑回归与加权分析所得的对数风险比估计值。条件逻辑回归模型针对体重指数和所有卡钳匹配因素进行了调整，而加权Cox比例风险模型还纳入了分类匹配因素。针对条件逻辑回归模型仅调整体重指数，或调整体重指数及匹配因素的不同子集所进行的敏感性分析得出了相似结果，因此不再列出。将条件逻辑回归估计值视为金标准时，使用任一版本 GAM 权重的加权分析所得估计值都高度相似，且在均方根误差（RMSE）方面表现最佳；该误差定义为 117 种代谢物的对数风险比估计值在两种方法间差值的平方和。相比之下，使用 KM 权重时，RMSE 增加约 33%；使用不含技术匹配因素的 KM 权重时，RMSE 几乎翻倍。这表明，与绝对风险估计中观察到的情况不同，在 EPIC-ENDO 中，计算 KM 权重时省略技术匹配因素会降低对数风险比估计的准确性。

最后，我们评估暴露因素之间的相关性。首先，我们关注原始队列中可获取的 11 项暴露因素：纳入时年龄、总能量摄入、体重指数、地中海饮食评分、习惯性酒精摄入、序数吸烟评分（同时反映吸烟习惯和吸烟强度），以及四项身体活动评分——职业性身体活动、休闲性身体活动、高强度身体活动和总体身体活动。我们将原始队列中计算出的相应 55 组两两相关性作为参考值，并与通过几种方法从 EPIC-ENDO NCC研究得出的估计值进行比较：对完整 NCC研究进行未加权分析、仅对对照进行未加权分析，以及使用四种权重的加权分析。对于每种方法，表1 列出了偏倚（与参考相关性的平均差值）和残差平方和（RSS）。就偏倚和 RSS 而言，对完整 NCC研究进行未加权分析的表现最佳，其次是对照的未加权分析。这表明选择偏倚的影响很小。在加权方法中，GAM 权重的偏倚略低于 KM 权重，但 RSS 表现较差。值得注意的是，GAM 权重 wo Technical 的 RSS 是完整 NCC 未加权分析的三倍。 M 。

这一发现印证了文献中此前提出的担忧：使用逆概率加权来推断暴露分布 [11]。具体而言，可能会为随访时间较短的对照分配极端权重，从而大幅增加方差。[4] 为解决这一问题，我们还考虑了GAM 权重的阈值化版本，将超过100的权重限制为100。该阈值化版本产生的偏倚与原始 GAM 权重相近，但显著降低了残差平方和，使其降至对照未加权分析所观察到的水平。补充材料中呈现了全面模拟研究的其他结果。这些结果证实，阈值化GAM权重通常能产生更稳定（即变异性更小）的估计值，但可能会增加偏倚。补充材料还包括一项关于EPIC-ENDO 研究中体重指数与117种代谢物之间关联的分析。具体而言，补充材料中的图8 以热图展示了这些相关性，特别突显了以下估计结果之间的高度相似性：

[4]此类权重对对数风险比或绝对风险估计的影响较小，正是因为相应个体的随访时间较短。

准则	未加权	未加权	对照	KM-	KM- 不含技术变量 M	GAM-	GAM- 不含技术变量 M	GAM- 阈值。
Bias	0.009	0.015	0.024	0.018	0.016	0.015	0.015	
RSSE	0.107	0.130	0.236	0.164	0.288	0.318	0.139	

表1： EPIC-ENDO 中 11 种暴露因素两两相关性的估计偏倚和残差平方和（RSSE）。

对照的未加权分析结果与采用阈值化GAM权重进行加权分析的结果。

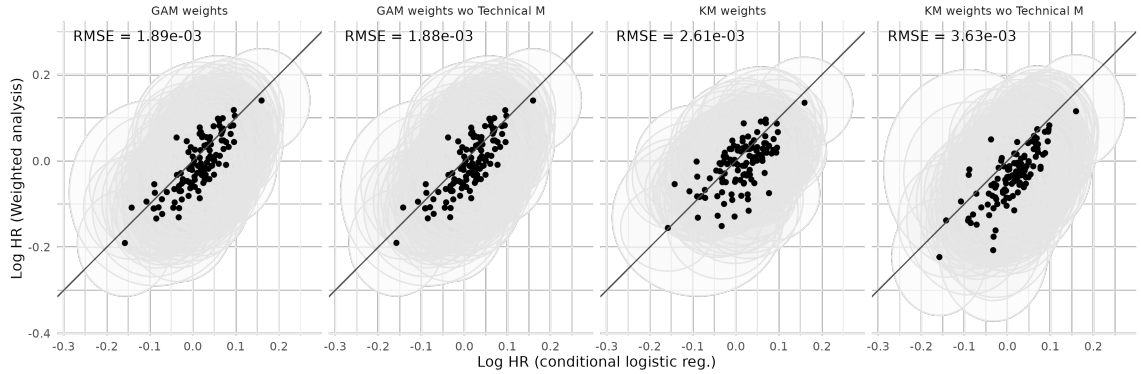

图7： 对 EPIC ENDO 中测量的 117 种代谢物进行加权分析所得的对数风险比估计值；并与条件逻辑回归模型的结果进行比较。对于每种代谢物，周围的椭圆反映了估计的精度。具体而言，椭圆半轴的长度对应于 1.96 个标准差，其中“sd”分别表示条件逻辑回归（y 轴）和加权分析（x 轴）中估计值的标准差。

5 结论

本文展示了巢式病例对照研究加权分析的表现，以及其如何受权重类型（KM 权重或 GAM 权重）和计算权重时所考虑变量的影响。通过关注三个目标估计量，我们表明，在某些情况下可能会产生一定偏倚。

首先，我们的结果表明，当原始队列中的受试者实际上不可能被选入嵌套病例对照研究时，需要谨慎处理。在这类情形下（例如某些特定匹配情形，或匹配因素较多而病例数较少时），KM 权重可能导致有偏估计，尤其是（条件）绝对风险估计。GAM 权重基于更平滑的纳入概率估计，在这类情形下可能得出更准确的结果。

其次，GAM 权重在特定情形下也存在局限，尤其是当匹配因素之间和/或匹配因素与其他暴露因素之间可能对疾病风险产生交互作用时，除非在计算纳入概率时恰当地考虑了这些交互作用。更一般地说，我们表明，在计算权重时，应考虑哪些变量以及如何对其建模（尤其是是否纳入交互作用），不仅取决于这些变量如何影响NCC研究中的选择，还取决于它们与所关注的暴露因素和疾病风险之间的关系。

在缺乏选择最佳权重类型的既定标准时，我们建议比较使用不同加权方案进行加权分析的结果（例如 KM 权重、GAM 权重、包含交互作用的 GAM 权重）。在可能的情况下，还可以将简单分析的结果与在完整队列中开展的相应分析结果进行比较。

最后，我们的结果证实了先前对嵌套病例对照研究加权分析局限性的担忧，尤其是在关注暴露分布的泛函（如暴露之间的相关性）时。我们发现，方差可能很大，但某些权重变体，特别是阈值化版本以及忽略变量 T 和 D的 GAM 权重，可能会带来更好的表现。仍需进一步研究，才能全面评估不同加权方案在这一情境下的表现。

参考文献

[1] Thomas DC. 对 F. D. K. Liddell、J. C. McDonald 和 D. C. Thomas 所著《队列分析方法：通过石棉开采案例评估》的补遗。《英国皇家统计学会杂志 *A*辑》1977; **140**(4): 483–485.[2] Breslow NE. 流行病学统计学：病例对照研究。美国统计学会会刊1996; **91**(433): 14–28.[3] Borgan Ø, Breslow NE, Chatterjee N 等。《病例对照研究统计方法手册》。佛罗里达州博卡拉顿：CRC出版社，2018.[4] Riboli E, Hunt KJ, Slimani N 等。欧洲癌症与营养前瞻性调查（EPIC）：研究人群与数据收集。《公共卫生营养》2002; **5**(6B): 1113–1124.[5] Breeur M, Ferrari P, Dossus L 等。欧洲癌症与营养前瞻性调查（EPIC）中诊断前血液代谢物浓度的泛癌分析。《*BMC*医学》2022;**20**(1): 351.[6] Langholz B 和 Clayton D。巢式病例对照研究中的抽样策略。《环境健康展望》1994;**102**(增刊 8): 47–51.[7] Borgan Ø, Goldstein L 和 Langholz B。Cox 比例风险模型中抽样队列数据的分析方法。《统计年刊》1995;**23**(5): 1749–1778.[8] Samuelsen SO. 巢式病例对照研究分析的伪似然方法。《生物统计学》1997;**84**(2): 379–394.[9] Chen K. 一般化病例队列抽样。《英国皇家统计学会杂志 *B*辑：统计方法论》2001; **63**(4): 791–809.[10] Kim RS. 嵌套病例对照研究设计分析：重新审视逆概率加权法。《统计学应用与方法通讯》2013; **20**(6): 455–466.[11] Støer N 和 Samuelsen S. multipleNCC：嵌套病例对照数据的逆概率加权。《*R*期刊》2016;**8**(2): 5.

[12] Mark SD 和 Katki HA. 在缺失病例数据的两阶段（嵌套）队列研究中指定并实施非参数和半参数生存估计量。美国统计学会会刊2006; **101**(474): 460–471.[13] Kim RS 和 Kaplan RC. 嵌套病例对照研究设计中次要结局的分析。医学统计学2014;**33**(24): 4215–4226.[14] Zhou QM, Wang X, Zheng Y 等。当巢式病例对照研究中的病例仅为事件子集时的新加权方法。生物计量学杂志2022;**64**(7): 1240–1259.[15] Edelmann D, Ohneberg K, Becker N 等。致编辑的信，回应 Qian M. Zhou、Xuan Wang、Yingye Zheng 和 Tianxi Cai 的论文“当巢式病例对照研究中的病例仅为事件子集时的新加权方法”。《生物计量学杂志》2023; 65(4): 2200360.[16] Samuelsen SO, ˚Anestad H 和 Skrondal A. 一般队列抽样设计的分层病例队列分析。斯堪的纳维亚统计学杂志2007; **34**(1): 103–119.[17] Saarela O, Kulathinal S, Arjas E 等。用于多个结局的巢式病例对照数据：似然方法及其替代方案。医学统计学2008;**27**(28): 5991–6008.[18] Støer N 和 Samuelsen S. 具有额外匹配的巢式病例对照研究中的逆概率加权——一项模拟研究。医学统计学2013;**32**(30): 5328–5339.[19] Dossus L, Kouloura E, Biessy C 等。循环代谢物与子宫内膜癌风险的前瞻性分析。妇科肿瘤学2021; **162**(2): 475–481.[20] Rabe-Hesketh S 和 Skrondal A. 忽略不可忽略的缺失。心理计量学2023; 88(1): 31–50.[21] van der Vaart AW 和 Wellner JA. 弱收敛与经验过程。第1版。施普林格统计学丛书，纽约：施普林格，1996.[22] Andersen PK, Borgan Ø, Gill RD 等。《基于计数过程的统计模型》。施普林格统计学丛书，纽约：施普林格美国，1993.

确保嵌套病例对照研究加权分析一致性的权重计算中需考虑的变量

在本附录中，我们首先介绍一个有向无环图，用以描述嵌套病例对照研究中涉及的主要变量之间的因果关系（图8）。随后，我们证明方程（1）中陈述的一般结果。最后，我们给出一些简单的结果，说明根据目标估计量的不同，在特定情形下可以忽略某些匹配因素。

A.1 表示底层数据生成机制的有向无环图

根据定义，NCC研究中的选择取决于所有匹配因素 M。然而，匹配因素影响结局 Y 及其成因的方式可能各不相同。令 X 表示 p维向量，其中$p \geq 1$，由在 NCC研究中测量的、感兴趣的候选风险因素 (*i*)，以及 Y 的其他所有成因 (*ii*) 组成；这些成因不属于匹配因素 M。一般而言，M 中的匹配因素可分为三个子组，如图8所示：(*i*) W 表示可能同时导致 X 和 Y的因素；(*ii*) V 表示可能导致 Y、但不会导致 X 的因素；(*iii*) Z 表示既不会导致 X、也不会导致 Y的因素。尽管有人可能认为，匹配因素应从同时导致 X 和 Y的变量中选取，但我们在第 4 节给出的例子说明，某些“技术性”因素（例如采血时的空腹状态）可能被认为既不会导致 X，也不会导致Y。若计划在 NCC研究分析中使用条件逻辑回归，纳入此类匹配因素可能很有意义，因为它们有助于确保病例及其对照具有可比性，例如在空腹状态方面。不过，如下文所述，在计算加权分析所用的权重时，通常无需考虑这些因素。

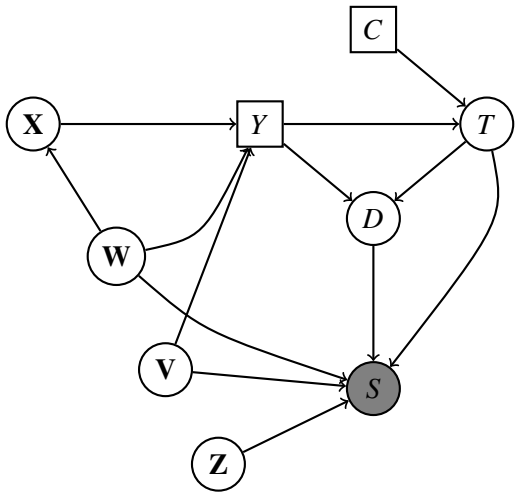

图8：嵌套病例对照研究中的典型有向无环图，其中匹配因素可分为三组：可能同时影响**W**和**X** 及 Y 的因素、仅可能影响 Y 的因素 **V**，以及既不影响 **X** 也不影响 Y的因素 **Z**。右删失变量用方框表示，非删失变量用圆圈表示。变量 S 表示 NCC 研究中的选择，以阴影标出。

根据因果有向无环图的标准规则，我们有

$$\begin{aligned}(Y, C, \mathbf{X}) &\perp\!\!\!\perp S \mid (D, T, \mathbf{W}, \mathbf{V}) \\ \mathbf{X} &\perp\!\!\!\perp S \mid (D, T, \mathbf{W}).\end{aligned} \tag{7}$$

这些条件独立性对应于选择的条件可忽略性的陈述 [20],这对于确定在计算权重时必须考虑哪些匹配因素、哪些可以忽略非常有价值，具体取决于目标估计量。以下段落将详细介绍我们工作中考虑的估计目标。我们首先详细证明公式（1），这对后续段落至关重要。

A.2 证明 (1)

我们证明公式（1）中陈述的结果，将 M 替换为 $\mathrm{M}_0 = (\mathrm{W}, \mathrm{V})$。注意到，在图8所示的有向无环图下，我们还有 $(Y, C, \mathrm{X}) \perp\!\!\!\perp S \mid (D, T, \mathrm{M})$，其中 $\mathrm{M} = (\mathrm{W}, \mathrm{V}, \mathrm{Z})$，因此，公式（1）中陈述的结果可通过类似论证得到。通过连续应用全期望定律和公式（7）中的第一个条件独立性陈述，并假设 $0 < \mathbb{P}(S = 1 | D, T, \mathrm{M}_0) < 1$几乎处处成立，可知，对于任意可积函数 ϕ，我们有

$$\begin{aligned}&\mathbb{E}\left[\frac{\phi(\mathbf{X}, Y)\mathbb{I}(S=1)}{\mathbb{P}(S=1|D,T,\mathbf{M}_0)}\right] \\ &\quad= \mathbb{E}\left\{\mathbb{E}\left[\frac{\phi(\mathbf{X}, Y)\mathbb{I}(S=1)}{\mathbb{P}(S=1|D,T,\mathbf{M}_0)}\middle| D, T, \mathbf{M}_0\right]\right\} \\ &\quad= \mathbb{E}\left\{\frac{\mathbb{E}[\phi(\mathbf{X}, Y)|D,T,\mathbf{M}_0]\mathbb{E}[\mathbb{I}(S=1)|D,T,\mathbf{M}_0]}{\mathbb{P}(S=1|D,T,\mathbf{M}_0)}\right\} \\ &\quad= \mathbb{E}[\phi(\mathbf{X}, Y)],\end{aligned}$$

这就是方程（1）中给出的结果，其中使用 $\mathbf{M}_0 = (\mathbf{W}, \mathbf{V})$ 代替 $\mathbf{M}$。

A.3 协变量之间关联的估计

为简单起见，我们考虑对其中一个协变量的线性回归进行估计，即 X_k，其中$k \in \{1, \ldots, p\}$，将其对其他协变量 $\mathbf{X}_\ell$进行回归，其中 $l \subseteq \{1, \ldots, p\} \setminus \{k\}$。我们假设 $\mathbb{E}[||\mathbf{X}_\ell||^2] < \infty$和$\mathbb{E}[X_k^2] < \infty$，且矩阵 $\mathbb{E}[\mathbf{X}_\ell^\top \mathbf{X}_\ell]$为正定矩阵。我们的目标是估计总体线性回归系数

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\, \mathbb{E}[(X_k - \mathbf{X}_\ell^\top \theta)^2].$$

我们定义逆概率加权估计量

$$\hat{\theta} = \left(\sum_{i=1}^{n} \tilde{w}_i \mathbf{X}_{\ell,i} \mathbf{X}_{\ell,i}^\top\right)^{-1} \left(\sum_{i=1}^{n} \tilde{w}_i \mathbf{X}_{\ell,i} X_{k,i}\right),$$

其中我们使用符号 $\tilde{w}_i = S_i/\mathbb{P}(S=1|D=D_i,\ T=T_i,\ \mathbf{W}=\mathbf{W}_i)$。我们回顾一下，$n$ 表示原始队列的大小，而 $S_i=1$ 表示被选入嵌套病例对照研究。

假设 $0<\mathbb{P}(S=1|D,\ T,\ \mathbf{W})<1$ 几乎处处成立，则通过与上述类似的论证可得

$$\mathbb{E}\left[\frac{\psi(\mathbf{X})\mathbb{I}(S=1)}{\mathbb{P}(S=1|D,T,\mathbf{W})}\right]=\mathbb{E}[\psi(\mathbf{X})],$$

对于任意可积函数 ψ。这里，我们利用了公式（7）中的第二个条件独立性表述。特别地，

$$\mathbb{E}\left[\frac{S}{\mathbb{P}(S=1|D,T,\mathbf{W})}\mathbf{X}_\ell X_k\right]=\mathbb{E}[\mathbf{X}_\ell X_k],\quad \mathbb{E}\left[\frac{S}{\mathbb{P}(S=1|D,T,\mathbf{W})}\mathbf{X}_\ell^\top\mathbf{X}_\ell\right]=\mathbb{E}[\mathbf{X}_\ell^\top\mathbf{X}_\ell].$$

因此，参数 θ^* 满足：

$$\theta^*=\left(\mathbb{E}[\mathbf{X}_\ell^\top\mathbf{X}_\ell]\right)^{-1}\mathbb{E}[\mathbf{X}_\ell X_k]=\left(\mathbb{E}\left[\frac{S}{\mathbb{P}(S=1|D,T,\mathbf{W})}\mathbf{X}_\ell^\top\mathbf{X}_\ell\right]\right)^{-1}\mathbb{E}\left[\frac{S}{\mathbb{P}(S=1|D,T,\mathbf{W})}\mathbf{X}_\ell X_k\right].$$

根据大数定律以及上述假设，可得

$$\frac{1}{n}\sum_{i=1}^n\tilde{w}_i\mathbf{X}_{\ell,i}X_{k,i}\xrightarrow{\mathbb{P}}\mathbb{E}[\mathbf{X}_\ell X_k],\quad \frac{1}{n}\sum_{i=1}^n\tilde{w}_i\mathbf{X}_{\ell,i}\mathbf{X}_{\ell,i}^\top\xrightarrow{\mathbb{P}}\mathbb{E}[\mathbf{X}_\ell^\top\mathbf{X}_\ell].$$

由于根据假设，$\mathbb{E}[\mathbf{X}_\ell^\top\mathbf{X}_\ell]$ 可逆，因此我们也有

$$\left(\frac{1}{n}\sum_{i=1}^n\frac{S_i}{\mathbb{P}(S=1|D_i,T_i,\mathbf{W}_i)}\mathbf{X}_{\ell,i}\mathbf{X}_{\ell,i}^\top\right)^{-1}\xrightarrow{\mathbb{P}}\left(\mathbb{E}[\mathbf{X}_\ell^\top\mathbf{X}_\ell]\right)^{-1}.$$

因此，根据斯卢茨基定理，

$$\hat{\theta}\xrightarrow{\mathbb{P}}\theta^*.$$

由于在计算权重 $1/\mathbb{P}(S=1|D,\ T,\ \mathbf{W})$ 时仍需考虑 D 和 T，随访时间较短的对照通常会获得较大的权重。因此，在嵌套病例对照研究中仍被选中的对照可能会对 θ^* 的估计产生重大影响。值得注意的是，如果 $\mathbf{X}$ 不是 Y 的原因，那么 $\mathbf{X}\perp\!\!\!\perp S\,|\,W$ 在这种情况下，根据 $\mathbb{P}(S=1|\mathbf{W})$ 的估计值推导出的权重，可以有效估计 $\mathbf{X}$ 各组成部分之间的关联；更一般地，对于任意可积函数 ψ，也可以有效估计 $\mathbb{E}[\psi(\mathbf{X})]$，且不会受到基于 T 和 D 的权重所存在局限性的影响。

A.4 条件生存函数的估计

现在，我们关注某些 $t\geq 0, k\in\{1,\ldots,p\}$ 和 $x_0\in\mathbb{R}$ 的条件生存函数 $\mathbb{P}(Y\geq t|X_k=x_0)$ 的估计。我们假设 $Y\perp\!\!\!\perp C|X_k$，并且为简便起见，假设 X_k 是分类变量，且 $p_0=\mathbb{P}(X_k=x_0)>0$。无条件生存函数的证明可通过类似论证得出，故此处省略。

考虑以下加权条件 Kaplan-Meier 估计

$$\hat{S}_n(t) = \prod_{D_i=1, T_i \le t, X_{k,i}=x_0} \Big(1 - \frac{\tilde{w}_i}{\sum_{k:T_k \ge T_i, X_{k,i}=x_0} w_k}\Big),$$

其中，此时，$\tilde{w}_i = S_i/\mathbb{P}(S=1|D_i,\ T_i,\ \mathbf{W}_i,\ \mathbf{V}_i)$。

引入以下两个加权计数过程

$$N_n(t) = \sum_{i=1}^{n} \tilde{w}_i \mathbb{I}(T_i \le t, D_i = 1, X_{k,i} = x_0)$$

$$Y_n(t) = \sum_{i=1}^{n} \tilde{w}_i \mathbb{I}(T_i \ge t, X_{k,i} = x_0),$$

定义在 $[0, \tau(x_0)]$上，其中 $\tau(x_0)$ $=\sup\{t \ge 0\colon\ S_T(t)\ = \mathbb{P}(T \ge t|X_k = x_0)\ > 0\}$。

这两个函数类 $\mathcal{F}_N = \{\mathbb{I}\{T \le t,\ \ D = 1,\ \ X_k = x_0\colon\ \ t \in [0,\tau]\}\}$ 和 $\mathcal{F}_Y = \{\mathbb{I}\{T \ge t,\ \ X_k = x_0\colon$ $t \in [0,\tau]\}\}$均为 VC类；由于权重 $\tilde{w}_i$ 有界，这些函数的加权版本是 Glivenko-Cantelli 函数类 [21]。然后，采用与上述类似的论证，可得

$$\sup_{t \in [0, \tau(x_0)]} |N_n(t) - p_0 \mathbb{P}(T \le t, D = 1|X_k = x_0)| \xrightarrow{\mathbb{P}} 0$$

and

$$\sup_{t \in [0,\tau](x_0)} |Y_n(t) - p_0 S_T(t)| \xrightarrow{\mathbb{P}} 0.$$

根据计数过程理论中的标准论证[22],，我们得到

$$\hat{S}_n(t) = \exp\Big(-\int_0^t \frac{dN_n(u)}{Y_n(u)}\Big) + o_{\mathbb{P}}(1),$$

对于任意 $t < \tau(x_0)$。最后，将给定 $X_k = x_0$ 时，T在时刻 u 的条件风险率记为$\lambda(u|X_k = x_0)$，我们得到

$$\hat{S}_n(t) \xrightarrow{\mathbb{P}} \exp\Big(-\int_0^t \lambda(u|X_k = x_0)\Big) = \mathbb{P}(Y \le t|X_k = x_0),$$

在 $t \in [0, \tau(x_0)[$ 上一致地。

A.5 Cox比例风险模型下对数风险比的估计

在本段中，我们采用以下 Cox比例风险模型

$$\lambda(t|\mathbf{z}) = \lambda_0(t) \exp(\mathbf{z}^T \alpha^*)$$

对于某个变量$\mathbf{z}\ \in \mathbb{R}^d$ 和 $\alpha^* \in \mathbb{R}^d$，对于某个 $d \ge 1$。

将 $\tilde{w}_i = S_i/\mathbb{P}(S=1|D_i,T_i,\mathbf{W}_i,\mathbf{V}_i)$ 按照 A.4, $Y_j(t) = \mathbf{1}_{\{T_j \geq t\}}$，并令 $(\mathbf{Z}_1,\ldots,\mathbf{Z}_n)$ 表示变量 $\mathbf{z}$ 的 n 观测值。引入，对于 $k = 0,1,2,$

$$S_n^{(k)}(t,\alpha) = \sum_{j=1}^{n} \tilde{w}_j Y_j(t) \exp(\mathbf{Z}_j^\top \alpha)\mathbf{Z}_j^{\otimes k},$$

与 $\mathbf{Z}^{\otimes 0} = 1$、$\mathbf{Z}^{\otimes 0} = \mathbf{Z}$，以及 $\mathbf{Z}^{\otimes 2} = \mathbf{Z}\mathbf{Z}^T$，并且

$$\bar{Z}_n(t,\alpha) = \frac{S_n^{(1)}(t,\alpha)}{S_n^{(0)}(t,\alpha)}$$

将加权部分似然得分函数的根定义为 $\hat{\alpha}$,

$$U_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i \int_0^\tau \left[\mathbf{Z}_i - \bar{Z}_n(t,\alpha)\right] dN_i(t)$$

其中 $\tau = \sup\{t \geq 0 : \mathbb{P}(T \geq t) > 0\}$。

现在考虑上文所引入各量的总体版本。对于 $t \in [0,\tau]$、α 属于围绕 α^* 的紧A，且 $k = 0,1,2,$ 定义

$$s^{(k)}(t,\alpha) = \mathbb{E}\left[Y(t)\exp(\alpha^T\mathbf{Z})\mathbf{Z}^{\otimes k}\right],$$

假设对所有 $t \in [0,\tau]$ 和 $\alpha \in A$，$s^{(0}(t,\alpha) > 0$，进一步设 $\bar{z}(t,\alpha) = s^{(1)}(t,\alpha)/s^{(0)}(t,\alpha)$，并引入矩阵

$$v(t,\alpha) = \frac{s^{(2)}(t,\alpha)}{s^{(0)}(t,\alpha)} - \bar{z}(t,\alpha)^{\otimes 2}.$$

最后引入 $U(\alpha) = \mathbb{E}\left[\int_0^\tau (Z - \bar{z}(t,\alpha))dN(t)\right]$。

注意到函数类 $\{Y(t)\exp(\alpha^T\mathbf{Z})\mathbf{Z}^{\otimes k},\ t \in [0,\tau],\ \alpha \in A\}$ 满足 Glivenko-Cantelli 性质，且权重有界，因此通过与上述类似的论证可得

$$\sup_{t\in[0,\tau],\alpha\in A} |S_n^{(k)}(t,\alpha) - s^{(k)}(t,\alpha)| \xrightarrow{\mathbb{P}} 0, \quad k = 0,1,2.$$

因此，

$$\sum_{\alpha\in A} |U_n(\alpha) - U(\alpha)| \xrightarrow{\mathbb{P}} 0,$$

进而，进一步假设 $I(\alpha) = \int_0^\tau s^{(0)}(t,\ \alpha)v(t,\ \alpha)\lambda_0(t)dt$在 α^* 处正定（因此，α^* 是 U的唯一零点），根据 Z 估计量的标准一致性，最终有 $\hat{\alpha} \xrightarrow{\mathbb{P}} \alpha^*$ 。