Biometrics , 1–
20 2023年

arXiv:2303.03512v1 [stat.ME] 6 Mar 2023

一种高效的数据整合方案：整合多个次要数据集的信息，以推断主要分析的参数

Chixiang Chen,[1]* Ming Wang,[2] Shuo Chen[1]

[1] 生物统计学与生物信息学部，

马里兰大学医学院，美国马里兰州巴尔的摩

[2] 人口与定量健康科学系，凯斯西储大学医学院，美国俄亥俄州克利夫兰*联系邮箱：chixiang.chen@som.umaryland.edu

Summary: 许多观察性研究和临床试验会收集各种可能与主要终点高度相关的次要结局。这些次要结局通常会在与主数据分析分开的次要分析中分别进行分析。然而，这些次要结局可用于提高主要分析中的估计精度。我们提出一种称为多重信息借用（MinBo）的方法，该方法借用次级数据（包含次要结局和协变量）中的信息，以提高主要分析的效率。所提出的方法对次级数据的模型设定错误具有稳健性。理论研究和案例研究均表明，就效率增益而言，MinBo优于现有方法。我们将MinBo应用于社区动脉粥样硬化风险研究的数据，以评估高血压的风险因素。Key words: 数据整合；经验似然；估计精度；多个次要结局；稳健推断。

本文已提交，供《生物统计学》考虑是否发表

1. 引言

近几十年来，统计学家一直在研究能够充分利用主数据集的（半参数）有效估计量（Shao, 2003；Tsiatis, 2006）。随着当代研究数据日益丰富，通过利用并整合多个数据源，可以进一步提高估计精度（Qin等人，2022）。数据整合的应用已在流行病学、临床试验、遗传学和基因组学等诸多领域得到广泛研究（Lee等人，2020；Li和Song，2020）。本文提出一种新的视角，并构建一个新颖的框架，有效整合来自多个次级数据源的信息，以提高主要分析模型中的参数估计精度。在本文中，如果一个数据集包含与主要终点高度相关的次要结局及一些协变量，我们就将其视为次级数据集。

我们的研究深受 Atherosclerosis Risks in Communities（ARIC）研究的启发（ARIC 研究者，1989；Gonz´alez 等人，2018）。我们的目标是识别高血压发生的风险因素。高血压影响着超过三分之一的美国成年人，也是全球心血管疾病的主要病因之一。原发性高血压定义为收缩压（SBP）$\geqslant 140$ mm Hg 或舒张压（DBP）$\geqslant 90$ mm Hg，或需要服用降压药物。传统上，人们使用以高血压状态为主要终点的逻辑回归来研究吸烟、饮酒和年龄等风险因素（Al-Nozha 等人，2007）。然而，这种成熟的最大似然估计（MLE）方法可能效率不高，因为它在主要分析中未纳入有关收缩压、舒张压或药物使用的信息。由于这些次要结局与主要终点高度相关，因此被认为包含额外信息，有助于改进主模型中的参数估计。其他流行病学研究和临床试验中也常见次要结局。例如，在以中风风险预测为主要关注点的研究中，心肌梗死和外周

血管疾病等其他事件可能是次要结局，并可提供与中风相关的信息（Wilson 等人，2011）。这些实际情境促使我们开发一种新的数据整合工具，以有效综合多个次要数据集中的信息，从而改进参数估计。

文献中，现有的数据整合技术大多旨在纳入外部或独立研究的信息。例如，广义荟萃分析（Kundu等人，2019）和约束最大似然方法（Chatterjee等人，2016；Han和Lawless，2019）利用外部研究的汇总信息。经验似然估计（Qin 和 Lawless，1994；Qin，2000；Qin等人，2015）和校准估计（Lumley等人，2011；Yang和Ding，2019）则考虑不同研究之间某种形式的共享效应或已知效应。贝叶斯整合（Xie等人，2013；Cheng等人，2018）利用外部来源的先验知识。然而，由于主要终点与次要结局高度相关，且二者来自同一项研究，这些方法无法直接应用于我们的数据框架。一种可能适用的方法是通过似然联合建模所有结局的均值及其相关性，从而为所有结局指定联合似然。尽管在概念上可以对多个结局进行联合建模，但由于结局特征和现实世界数据收集方案的复杂性，计算过程较为复杂。此外，若结局之间的任何分布假设不成立，主模型中的参数估计量就会不相合。为避免这些问题，Chen 等人（2021，2022）提出了一种重新加权估计方案，无需联合建模即可高效、稳健地利用次要数据集中的信息。然而，该方法仅适用于只有一个次要数据集的情况。实际中通常有多个次要数据集可用，高效整合这些信息可能会进一步改善主模型中的参数估计。我们在此介绍一种名为**多重信息借**用（MinBo）的高效稳健方案

，该方案利用多种次要结局。通过所提出的方案，可以有效地从多个信息来源借用信息，从而提高主要分析中的估计效率，并具有良好的计算效率。在温和条件下可以保证估计一致性，且对任何次要数据的“工作”模型错设并不敏感。本文其余部分安排如下。第 2 节介绍 MinBo 的一般估计框架。第3节展示详尽的数值评估。第4节说明 MinBo 在 ARIC研究中的应用，旨在识别高血压发生的风险因素。第5节作总结。详细证明、补充数值结果以及进一步讨论见补充材料。

2. 拟议框架：MinBo

2.1 基本设定与现有方法

图1对MinBo工作流程进行了示意性总结。简而言之，MinBo应用经验似然框架计算每位受试者的权重。随后，这些权重被整合到统一得分中，通过对主要估计方程重新加权，用于主要参数估计。

在介绍方法之前，我们先描述基本记号。对于受试者 i 从 $1,\ldots,n$，令 $\boldsymbol{D}_{0i}$ 表示主要研究收集的数据。该数据集包含主要终点 $\boldsymbol{Y}_{0i}$ 和作为协变量的风险因素 $\boldsymbol{X}_{0i}$。除主要终点 Y_{0i} 外，同一受试者还记录了 $K \geqslant 1$ 个次要结局。这些结局与主要终点高度相关。对于 $k = 1,\ldots,K$，令 $\boldsymbol{D}_{ki}$ 表示第 k个次要数据集，其中包含次要结局 $\boldsymbol{Y}_{ki}$ 及各种协变量 $\boldsymbol{X}_{ki}$。次要结局 $\boldsymbol{Y}_{ki}$在分布和数据结构方面可能与主要终点不同。次级数据中的协变量$\boldsymbol{X}_{ki}$ 可以与主数据集中的协变量相同、部分重叠，也可以完全不同

于主数据集中的协变量 $\boldsymbol{X}_{0i}$ 。假设我们主要关注的模型是对主数据集 $\boldsymbol{D}_{0i}$中的 $\boldsymbol{X}_{0i}$ 关于 $\boldsymbol{Y}_{0i}$ 进行回归。通常，回归中关注的参数$\boldsymbol{\beta}$ 可通过以下一般估计方程求解：

$$\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) = \mathbf{0}, \tag{1}$$

其中，函数 $\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta})$ 可以是最小二乘损失的导数、最大似然法或广义估计方程法中的得分函数等。给定某些参数值 $\boldsymbol{\beta}_0$ ，使得 $E\{\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta}_0)\} = \mathbf{0}$, 可基于矩估计法的理论推导出渐近标准误（Newey 和 McFadden，1994年）。为便于说明，我们假设主数据是完全观测的，且主模型被正确设定。主数据集中缺失数据这一更复杂的问题将在第5节和补充材料的第F节中讨论。然而，公式（1）中的估计程序可能并不高效，因为它没有使用任何次级数据。为考虑主要终点与次要结局之间的强关联，Chen 等人（2021年）提出了一种计算高效的估计方法，将一个次要结局的信息纳入主模型。具体而言，增强估计量可通过以下重新加权估计方程获得：

$$\sum_{i=1}^{n} \hat{p}_{ki} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) = \mathbf{0}, \tag{2}$$

其中，非负权重ˆp_{ki} 通过在以下约束条件下，针对 p_{ki}和 $\boldsymbol{\theta}_k$ 最大化 $\prod_{i=1}^{n} p_{ki}$ 来估计：

$$\sum_{i=1}^{n} p_{ki} = 1, \quad \sum_{i=1}^{n} p_{ki} R_{ki} \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k) = \mathbf{0}. \tag{3}$$

上述约束优化问题可通过采用拉格朗日乘子求解（Qin 和 Lawless，1994年；Owen，2001年）。 R_{ki} 是指示变量，用于表示是否观测到受试者 i 的第 k 个次要数据集。估计函数 $\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)$ 由针对数据 $\boldsymbol{D}_{ki}$、参数化为 $\boldsymbol{\theta}_k$ 的工作模型推导得出，例如广义估计方程（Liang 和

Zeger，1986年）。为提高主要参数 $\boldsymbol{\beta}$ 的估计效率，函数 $\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)$ 的维度应大于相应干扰参数向量 $\boldsymbol{\theta}_k$ 的维度，从而形成过度识别函数（Chen 等人，2021年）。补充材料 G 节列出了文献中过度识别函数的具体示例，其中讨论了纵向或横断面次级结局。

请注意，估计方程（2）和（3）与 Chen 等人（2021年）提出的方程略有不同。在补充材料的 D 节和 H 节中，我们证明了：(a) 这两种估计程序是等价的；(b) 在主要函数 $\boldsymbol{g}$ 和过度识别函数 $\boldsymbol{h}$ 已知的情况下，通过经验似然（Qin 和 Lawless，1994年）求解的估计量与由方程（2）求解的估计量渐近等价于半参数有效估计量。此外，上述方法仅适用于存在一个次级数据集的情况。这凸显了当有多个次级数据集可用时，方程（2）中的方法在实际应用中的局限性。为弥补这一不足，我们开发了一种高效的数据整合框架，以提高主模型中的估计效率（图1）。在接下来的章节中，我们将通过三种整合方案阐述该框架，并比较它们的表现。

2.2 *MinBo* 中的平均方案

第一种方案是所谓的平均方案。具体而言，我们首先将方程（2）和（3）应用于每个次级数据集。然后，根据从第 k 个工作模型计算得到的权重 $\hat{}p_{ki}$ ，我们构建一个新的得分函数，并通过求解以下重新加权估计方程来得到平均估计量：

$$\sum_{i=1}^{n} \bar{p}_i \boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}) = \mathbf{0}, \tag{4}$$

其中， $\bar{p}_i$ 由$\hat{}p_{ki}$的凸组合得到，即$\bar{p}_i = \sum_{k=1}^{K} \omega_k \hat{}p_{ki}$ ，其中常数 ω_k 非负，且满足 $\sum_{k=1}^{K} \omega_k = 1$。

为便于说明，我们先从一个简单情形入手：研究中有两个次级数据集（{v1}）

。推广到 $K > 2$ 的情形可见第2.4节。设 $\hat{\boldsymbol{\beta}}$ 为方程（4）的解。注意，由此得到的估计量是相合的，并且对次级数据的任何工作模型错设均具有稳健性。因此，只需证明第2.1节定义的真实参数值 $\boldsymbol{\beta}_0$ 是方程（4）的解，其中 $n \to \infty$。通过计算拉格朗日乘子 $\boldsymbol{\lambda}_k$ 和经验似然估计量 $\boldsymbol{\theta}_k$ （见补充材料的 B 节），可将每个个体的权重重新表示为 $p_{ki} = (1/n)/\{1 + \hat{\boldsymbol{\lambda}}_k^T R_{ki} \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \hat{\boldsymbol{\theta}}_k)\}$，其中 $k = 1, 2$。然后，估计方程 $\sum_{i=1}^n \bar{p}_i \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0) = \mathbf{0}$ 可改写为

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\omega_1}{1+\hat{\boldsymbol{\lambda}}_1^T R_{1i}\boldsymbol{h}_1(\boldsymbol{D}_{1i};\hat{\boldsymbol{\theta}}_1)}+\frac{\omega_2}{1+\hat{\boldsymbol{\lambda}}_2^T R_{2i}\boldsymbol{h}_2(\boldsymbol{D}_{2i};\hat{\boldsymbol{\theta}}_2)}\right\}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0) \\ =&\frac{1}{n}\sum_{i=1}^{n}(\omega_1+\omega_2)\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)+o_p(\mathbf{1}) \xrightarrow{p} E\big\{\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)\big\}=\mathbf{0}. \end{aligned} \tag{5}$$

由于对于所有 k 均有 $\hat{\boldsymbol{\lambda}}_k = O_p(\boldsymbol{n}^{-1/2})$ 和 $\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^* = O_p(\boldsymbol{n}^{-1/2})$，且某个 $\boldsymbol{\theta}_k^*$ 满足 $E\{R_k \boldsymbol{h}_k(\boldsymbol{D}_{ki}$；根据经验似然理论（Qin 和 Lawless，1994年；Owen，2001年；Han，2014），可以看出上述结果成立。根据（5），当 $n \to \infty$时，$\boldsymbol{\beta}_0$ 是（4）的解。这一发现不仅意味着 $\boldsymbol{\beta}$ 的一致性，也表明所提出方案的稳健估计。只要对于某个 $\boldsymbol{\theta}_k^*$ 的值（不一定是真实值），$E\{R_k \boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)\} = \mathbf{0}$成立，对 $\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)$ 的设定对主要参数 $\boldsymbol{\beta}$的估计一致性影响很小。这是一项理想性质，因为在实践中很难为所有次级数据正确设定模型。需要指出的是，模型设定容易受到各种因素的影响，例如均值结构设定错误，或未能考虑次级数据中的信息性缺失。以下定理总结了所得估计量的渐近正态性性质。

THEOREM 1: 在支持性信息 *A*节中的正则条件下，来自平均方案的估计量 $\boldsymbol{\beta}$ 渐近服从正态分布，其方差-协方差矩阵为

方差-协方差矩阵等于

$$Var\left\{n^{\frac{1}{2}}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right\}=\boldsymbol{\Gamma}^{-1}\big\{\boldsymbol{\Sigma}-(2\omega_1-\omega_1^2)\boldsymbol{\Lambda}_1\boldsymbol{S}_1\boldsymbol{\Lambda}_1^T-(2\omega_2-\omega_2^2)\boldsymbol{\Lambda}_2\boldsymbol{S}_2\boldsymbol{\Lambda}_2^T$$
$$+\omega_1\omega_2\boldsymbol{\Lambda}_1\big\{\boldsymbol{S}_1\boldsymbol{S}_{h_1h_2}\boldsymbol{S}_2^T\boldsymbol{\Lambda}_2^T+\omega_2\omega_1\boldsymbol{\Lambda}_2\boldsymbol{S}_2\boldsymbol{S}_{h_2h_1}\boldsymbol{S}_1^T\boldsymbol{\Lambda}_1^T\big\}(\boldsymbol{\Gamma}^T)^{-1},$$

其中 $\boldsymbol{\Gamma}=E\{\partial\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}^T\}$, $\boldsymbol{\Lambda}_k=E\{\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta}_0)\tilde{\boldsymbol{h}}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)^T\}$, $\boldsymbol{\Sigma}=E\{\boldsymbol{g}^{\otimes2}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta}_0)\}$ 和 $\boldsymbol{S}_k=\boldsymbol{S}_{k11}^{-1}(\boldsymbol{\theta}_k^*)-\boldsymbol{S}_{k11}^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{S}_{k12}(\boldsymbol{\theta}_k^*)\boldsymbol{\Omega}_k(\boldsymbol{\theta}_k^*)\boldsymbol{S}_{k21}(\boldsymbol{\theta}_k^*)\boldsymbol{S}_{k11}^{-1}(\boldsymbol{\theta}_k^*)$, 对于 $k=1,\ 2$。这里, $\tilde{\boldsymbol{h}}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)=R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)$, $\boldsymbol{\Omega}_k(\boldsymbol{\theta}_k^*)=\{\boldsymbol{S}_{k21}(\boldsymbol{\theta}_k^*)\boldsymbol{S}_{k11}^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{S}_{k12}(\boldsymbol{\theta}_k^*)\}^{-1}$, $\boldsymbol{S}_{k11}(\boldsymbol{\theta}_k^*)=E\{\tilde{\boldsymbol{h}}_k^{\otimes2}(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)\}$, $\boldsymbol{S}_{k12}(\boldsymbol{\theta}_k^*)=E\{\partial\tilde{\boldsymbol{h}}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)/\partial\boldsymbol{\theta}_k^T\}$, $\boldsymbol{S}_{k21}(\boldsymbol{\theta}_k^*)=\boldsymbol{S}_{k12}^T(\boldsymbol{\theta}_k^*)$, $\boldsymbol{S}_{h_kh_{k'}}=E(\tilde{\boldsymbol{h}}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)\tilde{\boldsymbol{h}}_{k'}^T(\boldsymbol{D}_{k'i};\ \boldsymbol{\theta}_{k'}^*))$, 对于 $k\neq k'$。记号 $\boldsymbol{f}^{\otimes2}$对任意函数向量 $\boldsymbol{f}$均为 $\boldsymbol{f}\boldsymbol{f}^T$。

Remark 1: 容易证明, 所得渐近方差不大于 $\tilde{\boldsymbol{V}}=\boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Gamma}^T)^{-1}$通过无权估计方程（1）求解得到的估计量的方差。

在定理 1 的协方差表达式的括号中, 第一项 $\boldsymbol{\Sigma}$来自主数据集 $\boldsymbol{D}_{0i}$。第二项和第三项, 即 $\boldsymbol{\Lambda}_1\boldsymbol{S}_1\boldsymbol{\Lambda}_1^T$和 $\boldsymbol{\Lambda}_2\boldsymbol{S}_2\boldsymbol{\Lambda}_2^T$, 表示整合次要数据集 $\boldsymbol{D}_{1i}$ 和 $\boldsymbol{D}_{2i}$所带来的方差缩减。剩余两项源于两个次要结局之间的关联。因此, 次要结局之间的关联强度会影响信息借用。接下来, 我们考虑两种极端情形。

Remark 2: 当两个次要数据集 $\boldsymbol{D}_{1i}$ 和 $\boldsymbol{D}_{2i}$ 相同时, 即 $\boldsymbol{D}_{1i}=\boldsymbol{D}_{2i}=\boldsymbol{D}_i$, 定理 1 得出的方差矩阵会简化为仅纳入这两个次要数据集中的任意一个时所得的矩阵。因此, 信息得以完全保留。直观地说, 重构得分函数能够恢复高度关联的次要结局所共有的潜在且占主导的量, 从而最大限度地保留来自这些高度关联次要结局的信息。

Remark 3: 假设次级数据中的两个次要结局相互独立。此时

方差矩阵简化为 $\boldsymbol{\Gamma}^{-1}\{\boldsymbol{\Sigma}-(2\omega_1-\omega_1^2)\boldsymbol{\Lambda}_1\boldsymbol{S}_1\boldsymbol{\Lambda}_1^T-(2\omega_2-\omega_2^2)\boldsymbol{\Lambda}_2\boldsymbol{S}_2\boldsymbol{\Lambda}_2^T\}(\boldsymbol{\Gamma}^T)^{-1}$。如果两个常数 ω_1 和 ω_2 均为正，则该平均方案无法汇总两个次要数据集中的全部信息，即 $2\omega_1-\omega_1^2=2\omega_2-\omega_2^2=1$。

最后，公式（4）中的平均方案取决于先验权重 ω_k 的选择。所需权重应能够平衡各次要结局的贡献。对主要参数估计贡献更大的次要数据集应赋予更高的权重。因此，我们采用所谓的信息借用指数（IIB）来确定权重：

$$\omega_k=\frac{\mathrm{IIB}_k}{\sum_{k'=1}^{K}\mathrm{IIB}_{k'}},\quad with\,\mathrm{IIB}_k=\mathrm{trace}\{\mathrm{Diag}(\hat{\tilde{\boldsymbol{V}}})^{-\frac{1}{2}}\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\Lambda}}_k\hat{\boldsymbol{S}}_k\hat{\boldsymbol{\Lambda}}_k^T(\hat{\boldsymbol{\Gamma}}^T)^{-1}\mathrm{Diag}(\hat{\tilde{\boldsymbol{V}}})^{-\frac{1}{2}}\},\tag{6}$$

其中，项 $\boldsymbol{V}$、$\boldsymbol{\Gamma}$、$\boldsymbol{\Lambda}_k$ 和 $\boldsymbol{S}_k$ 分别是 $\boldsymbol{V}$、$\boldsymbol{\Gamma}$、$\boldsymbol{\Lambda}_k$ 和 $\boldsymbol{S}_k$ 的相合估计量，且 $\mathrm{Diag}(\boldsymbol{V})$ 表示一个对角元素为 $\boldsymbol{V}$ 、其他元素均为零的矩阵。注意，项 $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}_k\boldsymbol{S}_k\boldsymbol{\Lambda}_k^T(\boldsymbol{\Gamma}^T)^{-1}$ 是由于纳入第 k 个次要数据集而降低的估计变异性。项 $\mathrm{Diag}(\tilde{\boldsymbol{V}})^{-1/2}$ 使效率增益能够与公式（1）中的未加权估计量相比较。由于 $\boldsymbol{S}_k$ 是半正定的，因此定义的权重始终非负。数值越大，表示从第 k 个次要数据集向主要分析传递信息的效果越好。因此，式（6）中的权重将给予对主要分析贡献更大的次要数据集更高的权重。

2.3 *MinBo* 中的聚合方案

MinBo 中的第二种方案称为聚合方案。假设第 k 个数据集的个体权重 $\hat{p}_{ki}$ 已按第2.1节所述计算，其中 $i=1,\ldots,n$。接下来，我们考虑以下重新加权方案：

$$\sum_{i=1}^{n}\tilde{p}_i\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta})=\mathbf{0},\tag{7}$$

其中，聚合得分 $\tilde{p}_i$ 定义为 $\tilde{p}_i=\prod_{k=1}^{K}\hat{p}_{ki}$。同样，为便于说明，我们重点讨论

˜有两个次要数据集 ($K = 2$) 的情形。我们用 $\boldsymbol{\beta}$ 表示式 (7) 的解。与第2.2节一样，我们可以证明，真实的 $\boldsymbol{\beta}_0$ 也是式 (7) 的解，其中 $n \to \infty$ 为

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{1+\hat{\boldsymbol{\lambda}}_1^T R_{1i}\boldsymbol{h}_1(\boldsymbol{D}_{1i};\hat{\boldsymbol{\theta}}_1)}\right\}\cdot\left\{\frac{1}{1+\hat{\boldsymbol{\lambda}}_2^T R_{2i}\boldsymbol{h}_2(\boldsymbol{D}_{2i};\hat{\boldsymbol{\theta}}_2)}\right\}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)\\ =&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)+o_p(\mathbf{1})\xrightarrow{p}E\big\{\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)\big\}=\mathbf{0}.\end{aligned}\tag{8}$$

˜因此，推导表明，如果 $E\big\{R_k\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)\big\}=\mathbf{0}$，估计量 $\boldsymbol{\beta}$ 是相合估计量，且对次级数据工作模型的错设具有稳健性。以下定理总结了通过聚合得到的估计量的渐近性质。

THEOREM 2: 在补充信息 $\boldsymbol{A}$ 节中的正则条件下，通过聚合方案得到的估计量 $\boldsymbol{\beta}$ ˜渐近服从正态分布，其方差-协方差矩阵为

$$\begin{aligned}Var\big\{n^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\big\}=\boldsymbol{\Gamma}^{-1}\big\{&\boldsymbol{\Sigma}-\boldsymbol{\Lambda}_1\boldsymbol{S}_1\boldsymbol{\Lambda}_1^T-\boldsymbol{\Lambda}_2\boldsymbol{S}_2\boldsymbol{\Lambda}_2^T\\&+\boldsymbol{\Lambda}_1\boldsymbol{S}_1\boldsymbol{S}_{h_1h_2}\boldsymbol{S}_2^T\boldsymbol{\Lambda}_2^T+\boldsymbol{\Lambda}_2\boldsymbol{S}_2\boldsymbol{S}_{h_2h_1}\boldsymbol{S}_1^T\boldsymbol{\Lambda}_1^T\big\}(\boldsymbol{\Gamma}^T)^{-1}.\end{aligned}$$

所有符号均在定理 *1* 中定义。

REMARK 4: 让我们重新考察注释3中的情形。假设来自次要数据集 $\boldsymbol{D}_{1i}$ 和 $\boldsymbol{D}_{2i}$ 的两个次要结局相互独立。由此得到的渐近方差简化为 $\boldsymbol{\Gamma}^{-1}\big\{\boldsymbol{\Sigma}-\boldsymbol{\Lambda}_1\boldsymbol{S}_1\boldsymbol{\Lambda}_1^T-\boldsymbol{\Lambda}_2\boldsymbol{S}_2\boldsymbol{\Lambda}_2^T\big\}(\boldsymbol{\Gamma}^T)^{-1}$。这意味着聚合方案能够恢复来自两个独立次要数据集的全部信息。因此，若次要数据集/结局之间的关联较弱，则聚合方案优于平均方案。

REMARK 5: 假设两个次要数据集完全相同，即 $\boldsymbol{D}_{1i}=\boldsymbol{D}_{2i}=\boldsymbol{D}_i$。由此˜得到的渐近方差简化为 $\boldsymbol{V}$，即公式(1)中无权重估计程序所得的方差。因此，聚合两个高度相关的次要数据集/结局不会带来效率增益。

2.4 *MinBo* 的推广：综合方法

我们介绍了两种从多个次要数据集中整合信息的竞争性方案。总之，当次要结局高度相关时，平均方案更为可取（注释2）；而当次要结局的相关性较弱时，聚合方法更为可取（注释4）。现在，我们通过引入第三种方案——综合方案——对 MinBo 进行推广。该方案适用范围更广，并兼具前两种方案的理想特性。我们首先介绍一般设定，然后在本节末尾说明三个实际案例。

对于 $k = 1, \ldots, K$ 和 $k' = 1, \ldots, K'$，给定某个整数 $K' \leqslant K$，我们定义一个用户指定的数组，其元素均为非负值 $\omega_{k'k}$，且对每个 k'，都有 $\sum_{k=1}^{K} \omega_{k'k} = 1$。假设针对每个次要数据集，已根据公式（3）中的流程得到个体权重 $\hat{p}_{ki}$。那么，综合估计量可通过求解以下重新加权估计方程得到：

$$\sum_{i=1}^{n} \check{p}_i \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) = \mathbf{0}, \tag{9}$$

其中，综合得分函数的计算方式为

$$\check{p}_i = \prod_{k'=1}^{K'} \left(\sum_{k=1}^{K} \omega_{k'k} \hat{p}_{ki} \right). \tag{10}$$

Remark 6: 综合估计方案结合了平均和聚合方案。因此，根据第2.2节和第2.3节中的论述，我们可以指出，在满足温和的一阶矩条件时，对于任何次级数据，所得的综合估计量 $\boldsymbol{\beta}$ 也具有相合性，并且对工作模型的错设具有稳健性。

Theorem 3: 在补充材料 *A* 节中的正则条件下，$\check{}$ 所得估计量 $\boldsymbol{\beta}$ 渐近服从方差-协方差

矩阵等于

$$Var\big(n^{\frac{1}{2}}(\check{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\big)=\boldsymbol{\Gamma}^{-1}\Big\{\boldsymbol{\Sigma}-\sum_{m=1}^{K}(2\tilde{\omega}_m-\tilde{\omega}_m^2)\boldsymbol{\Lambda}_m\boldsymbol{S}_m\boldsymbol{\Lambda}_m^T + \sum_{m\neq m'}^{K}\tilde{\omega}_m\tilde{\omega}_{m'}\boldsymbol{\Lambda}_m\boldsymbol{S}_m\boldsymbol{S}_{h_mh_{m'}}\boldsymbol{S}_{m'}^T\boldsymbol{\Lambda}_{m'}^T\Big\}(\boldsymbol{\Gamma}^T)^{-1},$$

其中 $\tilde{\omega}_m=\sum_{k'=1}^{K'}\omega_{k'm}$。所有符号，尤其是 $\boldsymbol{\Lambda}_m$、 $\boldsymbol{S}_m$ 和 $\boldsymbol{S}_{h_mh_{m'}}$，均与定理 ***1*** 相同，但现在索引 m 和 m' 可以大于二。

当 $K'=1$ 时，方程（9）简化为公式（4）中的平均方案。当 $\omega_{k'k}=1$ 对于 $k'=k$，而在其他情况下 $\omega_{k'k}=0$ 时，方程（9）就变为公式（7）中的聚合方案。此外，综合方案更加灵活，因为它能够处理实践中更复杂的情况。

我们来考虑三种常见情形。假设除主要研究数据集 $\boldsymbol{D}_{0i}$外，还有三个次要数据集 $\boldsymbol{D}_{1i}$、$\boldsymbol{D}_{2i}$和 $\boldsymbol{D}_{3i}$。如果次要结局之间的相关性较弱，我们将 $\omega_{11}=\omega_{22}=\omega_{33}=1$ 和$\omega_{k'k}=0$ 设为 $k'\neq k$，采用聚合方案。相反，如果所有次要结局之间的相关性都很强，则将 $K'=1$ 设为相应值，采用平均方案。如果前两个数据集中的次要结局彼此高度相关，但与第三个次要数据集中的结局仅有较弱相关性，则可以采用一种组合方案，使得$\check{}p_i=(\omega_{11}\hat{}p_{1i}+\omega_{12}\hat{}p_{2i})\hat{}p_{3i}$，其中我们将 $K'=2$、$\omega_{23}=1$和 $\omega_{13}=\omega_{21}=\omega_{22}=0$设为相应值。给定这样的重构得分函数后，可以通过对$\hat{}p_{1i}$和$\hat{}p_{2i}$进行凸组合，缓解 $\boldsymbol{D}_{1i}$ 与 $\boldsymbol{D}_{2i}$ 之间高度关联所带来的问题。因此，可以将 $\boldsymbol{D}_{3i}$ 中的信息大量整合到主要分析中。这三种实际情形下的权重可通过第2.2节中的公式（6）确定。在实际应用中，可先根据次要结局的临床特征及其关联情况，确定综合方法的形式；随后，再将其与一两种备选形式进行比较，并根据估计的效率增益加以调整。

3. 模拟

本节中，我们评估了所提出的估计量在 MinBo 中的数值表现。为模拟 ARIC研究中的情形，同时使设定保持一般性，我们考虑了以下情景（补充材料中的图 S1）。一个主数据集 $\boldsymbol{D}_{0i}$ 包含一个以二元尺度表示的主要终点。此外，还有三个来自同一研究的次要数据集，分别为 $\boldsymbol{D}_{1i}$、 $\boldsymbol{D}_{2i}$和 $\boldsymbol{D}_{3i}$。前两个次要数据集 $\boldsymbol{D}_{1i}$ 和 $\boldsymbol{D}_{2i}$ 包含连续尺度的重复测量。第三个次要数据集 $\boldsymbol{D}_{3i}$ 包含一个以横断面形式收集的二元变量。下文中，我们生成了三种次要结局之间不同的关联模式，以评估所提出的三种数据整合方案，并检验当三个次要数据集的所有工作模型均存在模型错设时 MinBo 的稳健性。

3. 数据生成

首先，我们生成了三个次要数据集。第一个和第二个次要数据集中的结局是纵向测量，其模型为 $\boldsymbol{Y}_{ki} = \boldsymbol{X}_{ki}\boldsymbol{\theta}_k + \boldsymbol{\epsilon}_{ki}$ for k =1、2，其中 $\boldsymbol{\theta}_1 = (-1、2、1、1)^T$、 $\boldsymbol{\theta}_2 = (1、-2、-1、-1)^T$，以及 $\boldsymbol{X}_{ki} = (1、\tilde{\boldsymbol{X}}_{ki1}、\tilde{\boldsymbol{X}}_{ki2}、\tilde{\boldsymbol{X}}_{ki3})$，其中 $\tilde{\boldsymbol{X}}_{kij} = (\tilde{X}_{kij1}、.....、\tilde{X}_{kij4})^T$。$\boldsymbol{X}_{ki}$ 和 $\boldsymbol{\epsilon}_{ki}$ 的详细设定见补充材料的 E 节。我们使用 ρ 来量化 $\boldsymbol{\epsilon}_{1i}$ 与 $\boldsymbol{\epsilon}_{2i}$ 之间的相关性，取三个不同的值 $\rho = 0$、0.4或0.8，分别代表无相关性、弱相关性和强相关性。此外，在第三个数据集中 $\boldsymbol{D}_{3i}$，结局 Y_{3i} 服从伯努利分布，成功概率为 $p_i^* = \{1 + \exp(-\boldsymbol{X}_{3i}^T\boldsymbol{\theta}_3)\}^{-1}$，其中 $\boldsymbol{\theta}_3 = (-1、1)^T$ 和 $\boldsymbol{X}_{3i} = (1、\tilde{X}_{1i11})^T$，且 X_{1i11} 定义见第一个数据集 $\boldsymbol{D}_{1i}$。

为构建结果 Y_{3i} 与结果 $\boldsymbol{Y}_{1i}$ 和 $\boldsymbol{Y}_{2i}$ 之间可能存在的关联，我们考虑了以下生成机制。对于每个 i，若 $\epsilon_{1i1} \geqslant x_i^*$，二元结果 Y_{3i} 等于 1，否则等于 0；其中 ϵ_{1i1} 是残差向量 $\boldsymbol{\epsilon}_{1i}$ 的第一个元素，而 x_i^* 是标准正态分布的第 $(1 - p_i^*)$ 个百分位数。因此，

生成的结果 Y_{3i} 始终与重复测量 $\boldsymbol{Y}_{1i}$ 相关；而当 $\rho = 0$ 时，它与结果 $\boldsymbol{Y}_{2i}$ 独立，但当 $\rho = 0.8$ 时，它与 $\boldsymbol{Y}_{2i}$ 高度相关（补充材料中的图 S1）。

接下来，我们生成主数据集 $\boldsymbol{D}_{0i}$。主数据集中 $\boldsymbol{D}_{0i}$ 最受关注的结局 Y_{0i} 被视为二元变量，其成功概率为 $p_{0i} = \{1 + \exp(-\boldsymbol{X}_{0i}^T \boldsymbol{\beta}_0)\}^{-1}$，其中 $\boldsymbol{\beta}_0 = (1, \ -0.5, \ -1, \ 0.5)^T$ 和 $\boldsymbol{X}_{0i} = (1, \ \tilde{X}_{1i11}, \ \tilde{X}_{1i21}, \ \tilde{X}_{1i31})^T$。为确保它与三个次要结局相关，我们采用了以下生成过程。对于每个 i，我们构造一个变量 $\bar{x}_i = (\epsilon_{1i4} + \epsilon_{2i4})/\alpha$，其标量 α 经过指定，使 $\bar{x}_i$ 服从标准正态分布。然后，当 $\bar{x}_i \geqslant x_{0i}$ 时，主要终点 Y_{0i} 等于 1；否则等于 0。值 x_{0i} 是标准正态分布的第 $(1-p_{0i})$ 个百分位数。因此，主要终点 Y_{0i} 与所有三个次要结局均相关。我们还考虑并评估了次要结局与主要终点仅略微相关的情况。（详情见补充材料 E 节。）

3.2 方法评估

基于主数据集 $\boldsymbol{D}_{0i}$，无权估计方程（1）是逻辑回归的得分函数。前两个次要数据集 $\boldsymbol{D}_{1i}$ 和 $\boldsymbol{D}_{2i}$ 的权重 $\hat{}p_{1i}$ 和 $\hat{}p_{2i}$ 通过（3）计算，其中工作函数 $\boldsymbol{h}_1(\boldsymbol{D}_{1i}; \ \boldsymbol{\theta}_1)$ 和 $\boldsymbol{h}_2(\boldsymbol{D}_{2i}; \ \boldsymbol{\theta}_2)$ 根据补充材料 G 节中的设定确定。权重 $\hat{}p_{3i}$ 根据工作函数 $\boldsymbol{h}_3(\boldsymbol{D}_{3i}; \ \boldsymbol{\theta}_3)$ 计算，其中 $\tilde{Z}_{3i} = (\tilde{X}_{1i21}, \ \tilde{X}_{1i31})^T$ 在补充材料 G 节中被视为冗余变量。我们考虑了样本量组合 $n = 300$ 或 600，以及 $\rho = 0$、0。4 或 0。8。研究设置中同时考虑了完全观测和部分观测的次级数据。三个次要数据集的观测比例分别设为 $\eta_1 = 0.6$、$\eta_2 = 0.7$ 和 $\eta_3 = 0.5$。

随后，我们在不同设置下比较了总共八种估计量。具体而言，

估计量 single100、single010 和 single001 分别纳入了第一个、第二个或第三个次要数据集。估计量 ave110 和 agg110 分别通过平均方案或聚合方案纳入了 $\boldsymbol{D}_{1i}$ 和 $\boldsymbol{D}_{2i}$。估计量 ave111 和 agg111 分别通过平均方案或聚合方案纳入了三个次要数据集。最后，估计量 omn111 通过综合方案纳入了三个次要数据集。估计量 omn111 使用了定义为 $\check{p}_i = (\omega_{11}\hat{p}_{1i} + \omega_{13}\hat{p}_{3i})\hat{p}_{2i}$ 的组合得分。平均方案和综合方案中的所有权重均根据 (6) 指定。

表1汇总了在给定 $\rho = 0.8$时，基于1000次蒙特卡洛模拟得到的八种估计量的估计偏差、蒙特卡洛标准差、渐近标准误和95%覆盖概率；此时三个次要结局之间具有较强的相关性。对于小样本量和大样本量，我们观察到所有估计量的偏差都很小，覆盖概率也令人满意，接近95%。样本量从300增加到600时，偏差趋近于零。我们还观察到，完全观测的次级数据对应的标准误小于部分观测的次级数据对应的标准误。 $\rho = 0.4$ 和 $\rho = 0$ 的结果分别列于补充材料中的表S1和表S2。这些表格也呈现出相同的模式。

我们使用经验相对效率（ERE）评估八种竞争估计量相较于最大似然估计法的效率增益。ERE是通过逻辑回归得到的最大似然估计量的经验方差（不考虑任何次级数据）与所提出估计量的经验方差之比。数值越大（大于1）越好。表2汇总了完全观测的次级数据对应的结果。

首先，我们考察了基于现有方法、仅纳入一个次要数据集的估计量（single100、single010 和 single001）。纳入 $\boldsymbol{D}_{1i}$ 或 $\boldsymbol{D}_{3i}$时， β_3 的经验相对效率仅略高于 1。整合 $\boldsymbol{D}_{2i}$时， β_1 的经验相对效率较低，这意味着

仅纳入一个次要数据集对主要估计的贡献有限。接下来，我们通过比较 ave110 和 agg110，评估次要结局之间的关联如何影响数据整合。当 $\rho = 0$时，即 $\boldsymbol{Y}_{1i}$ 与 $\boldsymbol{Y}_{2i}$之间不存在关联，agg110 的经验相对效率略优于 ave110。然而，随着 ρ 越来越接近 1，agg110 的表现变差；而相关系数越大，ave110 的表现则越好。这些结果与理论性质（备注2、3、4和5）相符。请注意，对于 ρ的所有取值，ave110 的经验相对效率也优于 single100 和 single010。这些发现支持第2.2节和第2.3节中的结论：次要结局之间相关性较弱时，聚合方案拟合效果更好；次要结局之间相关性较强时，则宜采用平均方案。最后，我们通过整合三个次要数据集评估了 MinBo。在这种情况下，当 $\rho = 0$ 或 0。4 时，带有 $\check{}p_i = (\omega_{11}\hat{}p_{1i} + \omega_{31}\hat{}p_{3i})\hat{}p_{2i}$ 的 omn111 在经验相对效率方面表现最佳。当$\rho = 0.8$时，ave111 估计量表现最佳。

值得注意的是，在所有设置中，ave111、agg111 和 omn111 中表现最佳的估计量，其 ERE 也高于仅纳入一个或两个次要数据集的所有估计量。因此，正确整合多个次要数据集有助于提高主模型的估计效率。这些发现同样适用于可获得部分观测次级数据的情况（表S3）。

3.3 错误设定的工作模型

如备注6所述，在 MinBo 中，任何次级数据的工作模型错设都可能对主模型的估计一致性影响甚微。实际情况下，此类模型错设可能包括未能正确设定均值结构，或未能妥善处理由设计问题、参与者拒绝、患者退出等导致的复杂缺失。

为检验 MinBo 的稳健性，我们开展了一项案例研究，在拟合三个次要数据集时，所有工作模型中的均值结构均设定错误。数据的生成方法与第3.1节相同。具体而言，对于次要˜ ˜ ˜结局 $\boldsymbol{Y}_{1i}$和 $\boldsymbol{Y}_{2i}$，我们没有使用 $\boldsymbol{X}_{ki} = (\mathbf{1},\ \boldsymbol{X}_{ki1},\ \boldsymbol{X}_{ki2},\ \boldsymbol{X}_{ki3})$ 拟合数据，而是对 $k = 1,\ 2$ 使用了设定错误的协变量向量 $\boldsymbol{X}_{ki}^{\text{mis}} = (\mathbf{1},\ \tilde{\boldsymbol{X}}_{ki1},\ \tilde{\boldsymbol{X}}_{ki2})$。拟合次要结局 Y_{3i}的数据时，我们将真实值 $\boldsymbol{X}_{3i} = (1,\ \tilde{X}_{1i11})^T$ 替换为协变量向量 $\boldsymbol{X}_{3i}^{\text{mis}} = (1,\ \tilde{X}_{1i12})^T$。因此，冗余协变量 $\tilde{Z}_{3i}$ 仅包含 $\tilde{X}_{1i31}$。

完全观测的次要数据结果汇总于表3。均值结构设定错误时，主模型中提出的估计量仍表现出较小的估计偏差，覆盖概率令人满意，接近95%名义水平。在经验相对效率（ERE）方面，MinBo仍能提高估计效率。部分观测的次级数据也观察到类似模式（表S4）。

上述结果假设次级数据完整，或因数据完全随机缺失而仅部分观测。我们还考察了次级数据中的信息性缺失（Enders，2010）对主要参数估计的影响。数据生成方案和结果汇总于补充材料的表S5。此外，我们评估了MinBo在各种设定下的表现，包括次要结局与主要结局轻度相关的情形（表S6），以及主要结局与次要结局之间相关性较高的设定（表S7）。我们还将MinBo与可指定多个结局联合似然时的联合似然估计进行了比较（表S8）。所有结果均表明MinBo稳健且表现良好，偏差较小，估计效率有所提高。

4. 真实数据应用

本节中，我们将MinBo应用于ARIC研究B中心的白人男性人群，以识别原发性高血压发生的基线风险因素。原发性高血压定义为在2周内至少发生以下三种情况之一：SBP $\geqslant 140$ mm Hg、DBP $\geqslant 90$ mm Hg，或服用抗高血压药物。为聚焦于新发高血压病例，我们排除了基线时已患高血压的受试者。主要终点定义为随访期间发生高血压（二元且横断面）。我们考察了以下潜在基线风险因素：体质指数（kg/m^2）、当前饮酒状态（1=是， 0=否）、当前吸烟状态（1=是， 0=否）、年龄（岁）和血红蛋白（g/dL）。主数据集 $\boldsymbol{D}_{0i}$纳入了1138名受试者。风险因素效应的常规估计值通过逻辑回归（使用最大似然估计）得出。为获得未加权估计值，所用得分函数为公式（1）中的得分函数。

我们考虑了以下三个具有不同次要结局的次要数据集：SBP（$\boldsymbol{D}_{1i}$）、DBP（$\boldsymbol{D}_{2i}$）以及在随访期间是否曾服用过降压药物（$\boldsymbol{D}_{3i}$）。需要注意的是，前两个次要结局是纵向测量值，记录于四次医生访视中。我们采用了补充材料 G 节中的估计函数，并将每次访视时的体质指数、当前饮酒状态、当前吸烟状态、年龄和血红蛋白水平作为协变量。正如第3节所述，我们使用四个基矩阵 $\boldsymbol{V}_1$、 $\boldsymbol{V}_2$、 $\boldsymbol{V}_3$ 和 $\boldsymbol{V}_4$ 来构建方程。对于主要分析中的所有受试者，每次访视均测量了 SBP 和 DBP。相比之下，对于第三个工作模型，我们采用了补充材料 G 节中的估计函数，并将体质指数、年龄和血红蛋白作为 $\boldsymbol{X}_{3i}$ 中的主要协变量。当前饮酒状态和当前吸烟状态在 $\boldsymbol{Z}_{3i}$ 中是冗余˜协变量。 $\boldsymbol{D}_{3i}$ 中的所有协变量均在基线时测量。需要注意的是，主要

分析中以及补充材料 G 节中的冗余变量是使用赤池信息准则选择的。主要分析中的所有受试者均被纳入$\boldsymbol{D}_{3i}$。

我们计算并比较了七种估计量。其中一种基于最大似然估计（MLE）方法，且未纳入任何次级数据。三种估计量基于整合单个次要数据集中的信息（Chen等，2021）。这三种估计量分别为 single100、single010 和 single001，分别纳入第一个、第二个或第三个次要数据集。其余三种估计量来自 MinBo。估计量 agg111 和 ave111 分别通过聚合方案或平均方案整合$\boldsymbol{D}_{1i}$、 $\boldsymbol{D}_{2i}$和 $\boldsymbol{D}_{3i}$ 。估计量 omn111 通过综合方法整合 $\boldsymbol{D}_{1i}$、 $\boldsymbol{D}_{2i}$和 $\boldsymbol{D}_{3i}$ ，其统一得分计算为 $\check{p}_i = (\omega_{11} \hat{p}_{1i} + \omega_{12} \hat{p}_{2i}) \hat{p}_{3i}$。该综合估计量利用了收缩压（SBP）和舒张压（DBP）之间因临床特征而存在的强关联（条件相关性$= 0.66$）。由于药物使用与另外两个协变量的相关性较弱（条件相关性 $= 0.2$），该估计量对药物使用的信息进行了聚合。所有权重均依据第2.2节中的（6）指定。

表4比较了六种估计量（不包括最大似然估计）的结果。估计相对效率定义为不使用次级数据时最大似然估计的估计方差与基于数据整合的估计量的估计方差之比。我们发现，所有整合次级数据的估计量都优于仅基于主数据的最大似然估计量（即相对效率大于1）。此外，整合多个次要数据集的估计量（ave111、agg111和omn111）优于仅使用单个次要数据集的现有方法所得到的估计量（single100、single010和single001）。在使用多个次要数据集的估计量中，MinBo的agg111和omn111表现最佳，因为它们在所有参数上的效率增益最高。与最大似然估计法相比，agg111和omn111在饮酒方面的效率增益接近三倍，

在吸烟方面接近三倍，在血红蛋白方面约为两倍。这些结果凸显了MinBo在整合多个次级数据集方面的实用性。我们在表5中展示了agg111和omn111估计量的详细结果。风险因素体质指数、饮酒和年龄具有统计学显著性。这些发现与既往文献报道相符（Shihab等人，2012）。值得注意的是，在仅基于主数据的最大似然估计方法中，只有体质指数和年龄显示出显著性证据，这进一步证明了MinBo的优越性。

5. 结论

将次级数据中的信息整合到主数据分析中，是流行病学和临床试验领域一个及时且重要的研究课题，因为随着技术进步，数据收集已变得容易得多。MinBo 是一种高效地将多个次要数据集纳入主要分析的初步尝试。即使任何次级数据所采用的工作模型存在设定错误，该方法仍能提高主要数据模型中估计的效率以及估计量的一致性。此外，与联合似然估计（如适用）相比，MinBo 可以针对每个工作模型分别生成估计值，从而降低计算负担。

在实践中，整合多个次要数据集颇具挑战，且取决于多种因素，例如次要结局之间的关联。MinBo 中的综合整合方案能够灵活应对这类情况。综合方法的具体形式可根据所关注次要结局的临床特征确定，也可通过比较考虑和不考虑次要结局时估计值的估计方差所计算出的相对效率来验证。在本研究中，我们假设所有主数据点均已观测到，或为完全随机缺失；但在某些应用中，情况可能并非如此。不过，我们的方法可以轻松嵌入成熟的方案中，例如逆概率加权技术

（Enders，2010年；Chen 等人，2021年）。估计程序的更多细节见《补充材料》F 节。此外，将我们的数据整合方案扩展到其他统计领域，例如生存分析、因果推断和高维估计，将十分有意义，也值得未来进一步研究。

数据可用性声明

支持本文研究结果的数据可向通讯作者提出合理请求获取。

参考文献

Al-Nozha, M. M., Abdullah, M., Arafah, M. R., Khalil, M. Z., Khan, N. B., Al-Mazrou, Y. Y., Al-Maatouq, M. A., Al-Marzouki, K., Al-Khadra, A., Nouh, M. S., 等（2007）。沙特阿拉伯的高血压。《沙特医学杂志》**28,** 77。

ARIC 研究者（1989）。社区动脉粥样硬化风险（ARIC）研究：设计与目标。《美国流行病学杂志》**129,** 687–702。

Chatterjee, N., Chen, Y.-H., Maas, P. 和 Carroll, R. J.（2016）。利用外部大数据来源的汇总级信息进行模型校准的约束最大似然估计。《美国统计学会杂志》**111,** 107–117。

Chen, C., Han, P. 和 He, F.（2021）。通过借用辅助数据中的信息改进主要分析。《医学统计学》。

Chen, C., Shen, B., Liu, A., Wu, R. 和 Wang, M.（2021）。一种用于间歇性缺失数据纵向分析的多重稳健倾向评分方法。《生物统计学》**77,** 519–532。

Chen, C., Yu, T., Shen, B. 和 Wang, M.（2022）。整合二手数据至生存分析以提高估计效率。《生物统计学杂志》。

Cheng, W., Taylor, J. M., Vokonas, P. S., Park, S. K., and Mukherjee, B.（2018）。改进

利用来自既有简化模型的外部信息，改进线性回归中的估计和预测。《医学统计学》**37**, 1515–1530。Enders, C. K.（2010）。应用缺失数据分析。Guilford press。

Gonz´alez, H. M., Tarraf, W., Harrison, K., Windham, B. G., Tingle, J., Alonso, A., Griswold, M., Heiss, G., Knopman, D., and Mosley, T. H. (2018). 中年期心血管健康与20年认知衰退：社区动脉粥样硬化风险研究结果。《阿尔茨海默病与痴呆》**14**,579–589.

Han, P. (2014). 缺失数据回归分析中的多重稳健估计。《美国统计学会杂志》**109**,1159–1173.

Han, P. and Lawless, J. F. (2019). 使用辅助汇总信息进行经验似然估计，适用于不同协变量分布。*Statist. Sinica* **29**, 1321–1342.

Kundu, P., Tang, R., and Chatterjee, N. (2019). 针对协变量信息各异的研究，对多元回归模型进行广义荟萃分析。《生物统计学》**106**,567–585.

Lee, S. Y., Lei, B., and Mallick, B.（2020）。整合全球数据并借鉴信息来估计COVID-19传播曲线。*PLOS ONE* **15**, e0236860。

Li, X. and Song, Y.（2020）。通过多来源数据整合进行目标人群统计推断——缓解罕见病临床试验信息短缺的一种方法。《生物制药研究统计学》**12**, 322–333。

Liang, K.-Y. and Zeger, S. L.（1986）。使用广义线性模型进行纵向数据分析。《生物统计学》**73**,13–22。

Lumley, T., Shaw, P. A., and Dai, J. Y.（2011年）。不完整数据的调查校准估计量与半参数模型之间的联系。《国际统计评论》**79**,200–220。

Newey, W. K. 和 McFadden, D.（1994年）。大样本估计与假设检验。

计量经济学手册 **4,** 2111–2245。Owen, A. B.（2001年）。经验似然。CRC出版社。Qin, J.（2000年）。杂论：结合参数似然与经验似然。《生物统计学》**87,** 484–490。Qin, J. 和 Lawless, J.（1994年）。经验似然与一般估计方程。《统计年鉴》第300–325页。Qin, J.、Liu, Y. 和 Li, P.（2022年）。利用相似研究中的校准信息的统计方法综述。《统计理论及相关领域》第1–16页。Qin, J.、Zhang, H.、Li, P.、Albanes, D. 和 Yu, K.（2015年）。利用协变量特异性疾病患病率信息提高病例对照研究的效能。《生物统计学》**102,** 169–180。Shao, J.（2003年）。数理统计。Shihab, H. M.、Meoni, L. A.、Chu, A. Y.、Wang, N.-Y.、Ford, D. E.、Liang, K.-Y.、Gallo, J. J. 和 Klag, M. J.（2012年）。体质指数与整个生命历程中新发高血压风险：约翰·霍普金斯前驱研究。《循环》**126,** 2983–2989。Tsiatis, A. A.（2006年）。半参数理论与缺失数据。Wilson, K. M.、Kasperzyk, J. L.、Rider, J. R.、Kenfield, S.、van Dam, R. M.、Stampfer, M. J.、Giovannucci, E. 和 Mucci, L. A.（2011年）。卫生专业人员随访研究中的咖啡消费与前列腺癌风险及进展。《美国国家癌症研究所杂志》**103,** 876–884。Xie, M.、Liu, R. Y.、Damaraju, C. 和 Olson, W. H.（2013年）。在具有二元结局的临床试验分析中纳入外部信息。《应用统计年鉴》**7,** 342–368。Yang, S. 和 Ding, P.（2019年）。结合多个观察性数据来源估计因果效应。《美国统计学会杂志》第1–46页。

补充材料

第3.2节和第3.3节中引用的网络附录和表格，以及实现第3节和第4节的R代码，可从Wiley在线图书馆的《生物统计学》网站随本文获取。

[图1见此处。][表1见此处。][表2见此处。]
[表3见此处。][表4见此处。][表5见此处。]

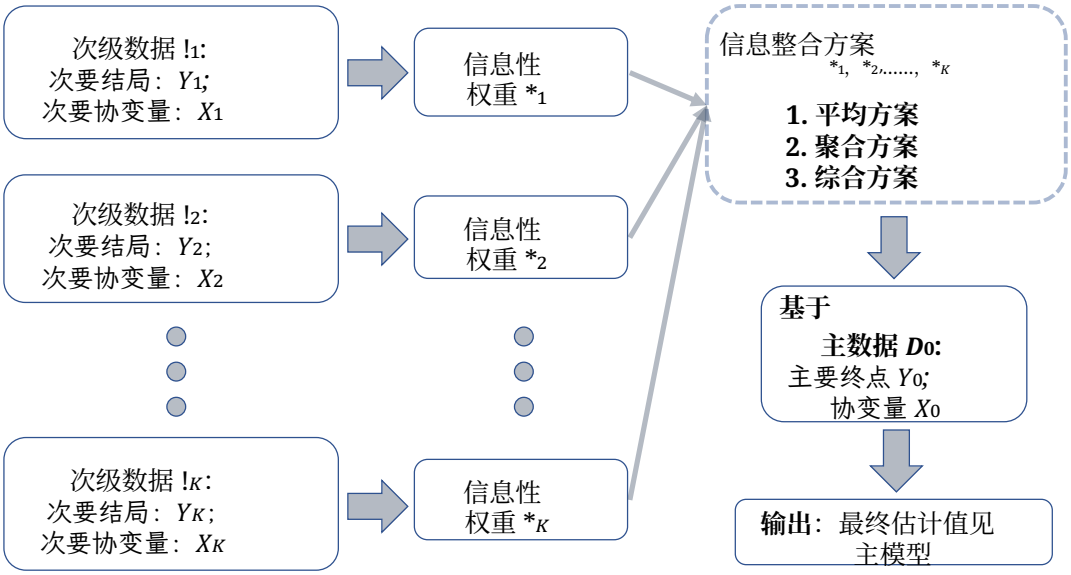

图1. MinBo 的示意工作流程。MinBo 首先根据每个次级数据计算信息性权重，然后根据若干整合方案合并这些权重，生成整合/统一得分。最后，将新得分作为主要估计方程中的新权重，以获得主模型中的参数估计量。

表1

Overall evaluations for eight estimators (defined in Section 3) under the combination of sample sizes ($n = 300, 600$) and proportions for observing three secondary datasets (CASE 1: full observation $\eta_1 = \eta_2 = \eta_3 = 1$; CASE 2: partial observation $\eta_1 = 0.6$ $\eta_2 = 0.7$ $\eta_3 = 0.5$) given the correlation (ρ) between two longitudinal secondary outcomes equal ,,, to 0.8. *MCSD: Monte Carlo standard deviation; ASE: asymptotic standard error; CP: coverage probability. All values have been multiplied by* 100.

			n=300				n=600			
			Bias	MCSD	ASE	95%覆盖概率	Bias	MCSD	ASE	95%覆盖概率
情形 1	single100	β_0	2.5	25	23	94	0.6	17	16	95
		β_1	-0.5	25	24	93	0.2	18	17	94
		β_2	-2.5	15	15	95	-1.0	11	10	94
		β_3	-0.7	29	28	95	0.2	20	20	95
	single010	β_0	2.4	24	24	96	0.9	17	17	95
		β_1	-0.1	28	28	96	-0.2	20	20	95
		β_2	-2.3	15	15	95	-1.0	11	10	94
		β_3	-1.0	25	24	95	0.2	17	17	95
	single001	β_0	2.5	25	23	95	0.7	17	16	95
		β_1	-0.7	26	25	94	-0.1	18	17	94
		β_2	-2.4	15	15	94	-0.9	11	10	94
		β_3	-0.4	29	28	94	0.2	20	19	95
	ave110	β_0	2.5	24	23	94	0.7	17	16	95
		β_1	-0.6	25	24	94	0.1	18	17	94
		β_2	-2.4	15	14	95	-1.0	11	10	95
		β_3	-0.6	29	27	95	0.2	20	19	94
	agg110	β_0	2.2	26	24	93	0.2	17	17	94
		β_1	-0.7	29	27	93	0.4	20	19	93
		β_2	-1.8	17	16	92	-0.7	12	11	94
		β_3	-0.8	30	27	94	0.0	20	19	94
	ave111	β_0	2.4	24	23	95	0.7	16	16	95
		β_1	-0.4	25	24	94	0.0	18	17	94
		β_2	-2.3	15	14	95	-0.9	11	10	94
		β_3	-0.7	27	26	95	0.2	18	18	94
	agg111	β_0	2.4	25	23	93	0.2	17	16	94
		β_1	-0.7	29	27	93	0.4	20	19	93
		β_2	-1.5	19	18	94	-0.6	13	12	95
		β_3	-1.3	26	24	93	0.0	18	17	94
	omn111	β_0	2.3	24	22	93	0.4	16	16	95
		β_1	-0.7	26	25	93	0.2	19	17	94
		β_2	-1.8	16	15	93	-0.7	11	10	94
		β_3	-0.9	26	25	94	0.1	18	18	94

			n=300				n=600			
			Bias	MCSD	ASE	95%覆盖概率	Bias	MCSD	ASE	95%覆盖概率
情况2	single100	β_0	2.6	25	24	95	0.9	17	17	95
		β_1	-0.3	27	26	94	-0.1	19	18	94
		β_2	-2.8	16	15	94	-1.1	11	11	94
		β_3	-0.7	29	28	95	0.3	20	20	95
	single010	β_0	2.6	25	24	95	0.8	17	17	95
		β_1	-0.1	28	28	96	-0.2	20	20	95
		β_2	-2.5	15	16	95	-1.0	11	11	94
		β_3	-1.1	27	26	95	0.4	19	18	95
	single001	β_0	2.4	25	24	95	0.8	17	17	95
		β_1	-0.4	27	26	95	-0.2	19	18	95
		β_2	-3.0	16	15	95	-0.9	11	11	94
		β_3	-0.2	29	28	95	0.3	20	19	95
	ave110	β_0	2.5	25	23	95	0.8	17	17	95
		β_1	-0.4	26	25	94	-0.1	19	18	94
		β_2	-2.9	16	15	95	-1.0	11	11	95
		β_3	-0.5	29	28	95	0.3	20	19	95
	agg110	β_0	2.3	26	24	93	0.5	17	17	95
		β_1	-0.4	29	27	93	0.0	20	19	94
		β_2	-2.7	17	16	93	-0.7	12	11	94
		β_3	-0.7	30	28	94	0.2	20	19	95
	ave111	β_0	2.5	25	23	95	0.8	17	16	96
		β_1	-0.3	27	25	95	-0.2	19	18	94
		β_2	-2.8	15	15	95	-1.0	11	11	95
		β_3	-0.6	28	27	95	0.3	19	19	95
	agg111	β_0	2.6	26	24	93	0.5	17	17	95
		β_1	-0.3	29	27	92	0.0	20	19	94
		β_2	-2.6	18	17	94	-0.7	13	12	93
		β_3	-1.3	28	26	93	0.3	19	18	95
	omn111	β_0	2.4	25	23	94	0.6	17	16	95
		β_1	-0.4	27	26	93	-0.1	19	18	94
		β_2	-2.7	16	15	93	-0.8	12	11	94
		β_3	-0.7	28	26	94	0.3	19	19	95

表2

Empirical relative efficiency for eight estimators (defined in Section 3) under the combination of different sample sizes (n) and correlation values between two longitudinal secondary outcomes (ρ). All secondary data are fully observed.

			经验相对效率							
			single100	single010	single001	ave110	agg110	ave111	agg111	omn111
n=300	ρ=0	β_0	1.01	1.05	1.03	1.06	1.04	1.08	1.08	1.10
		β_1	1.10	1.01	1.08	1.18	1.21	1.15	1.21	1.21
		β_2	1.05	1.10	1.03	1.10	1.07	1.14	1.05	1.15
		β_3	0.98	1.20	1.01	1.02	0.98	1.10	1.15	1.12
	ρ=0.4	β_0	1.03	1.07	1.03	1.07	1.01	1.11	1.08	1.11
		β_1	1.19	1.01	1.16	1.27	1.18	1.23	1.18	1.26
		β_2	1.10	1.18	1.07	1.15	1.03	1.22	0.94	1.15
		β_3	0.98	1.27	1.00	1.01	0.97	1.14	1.20	1.16
	ρ=0.8	β_0	1.04	1.11	1.04	1.06	0.92	1.12	1.02	1.07
		β_1	1.26	1.01	1.23	1.29	0.98	1.26	0.97	1.17
		β_2	1.13	1.20	1.11	1.15	0.87	1.22	0.71	1.01
		β_3	0.98	1.36	0.99	1.00	0.94	1.16	1.22	1.18
n=600	ρ=0	β_0	1.04	1.04	1.06	1.08	1.10	1.09	1.13	1.14
		β_1	1.13	1.00	1.08	1.19	1.24	1.14	1.24	1.21
		β_2	1.05	1.09	1.07	1.12	1.13	1.14	1.12	1.20
		β_3	0.99	1.19	1.03	1.03	1.02	1.12	1.18	1.16
	ρ=0.4	β_0	1.06	1.05	1.07	1.09	1.07	1.12	1.12	1.15
		β_1	1.21	1.00	1.16	1.27	1.19	1.21	1.19	1.26
		β_2	1.12	1.14	1.12	1.18	1.11	1.21	1.02	1.21
		β_3	0.98	1.27	1.02	1.02	1.00	1.15	1.21	1.19
	ρ=0.8	β_0	1.06	1.09	1.08	1.09	0.99	1.14	1.08	1.13
		β_1	1.26	1.00	1.22	1.28	0.99	1.25	0.98	1.17
		β_2	1.17	1.19	1.18	1.20	0.98	1.25	0.81	1.12
		β_3	0.99	1.36	1.02	1.01	0.99	1.17	1.29	1.23

表3

Overall evaluations of MinBo for parameter estimation in the main model when working models are misspecified.

We consider the combination of different sample sizes (n) and correlation values (ρ) given three secondary datasets , are fully observed. MCSD: Monte Carlo standard deviation; ASE: asymptotic standard error; ERE: empirical relative efficiency; CP: coverage probability. All values (except ERE) are multiplied by 100.

			n=300					n=600				
			Bias	MCSD	ASE	ERE	95%覆盖概率	Bias	MCSD	ASE	ERE	95%覆盖概率
ρ=0	ave111	β_0	2.5	24	24	1.06	94	1.2	17	17	1.07	95
		β_1	-0.8	27	26	1.13	94	-0.6	19	18	1.12	95
		β_2	-2.7	16	15	1.08	94	-1.1	11	11	1.08	94
		β_3	-0.4	28	27	1.06	95	0.1	19	19	1.07	95
	平均101	β_0	2.6	25	24	1.05	94	1.2	17	17	1.06	95
		β_1	-0.9	27	26	1.15	94	-0.6	18	18	1.15	95
		β_2	-2.7	16	15	1.09	94	-1.1	11	11	1.10	94
		β_3	-0.4	29	28	0.98	95	0.1	20	20	1.00	95
	agg111	β_0	2.4	24	23	1.08	94	1.1	16	16	1.12	95
		β_1	-0.8	26	25	1.22	94	-0.5	18	17	1.23	95
		β_2	-2.7	16	15	1.08	93	-1.1	11	10	1.14	94
		β_3	-0.3	27	26	1.11	94	0.0	19	18	1.14	94
	聚合101	β_0	2.6	25	23	1.03	93	1.2	17	16	1.08	96
		β_1	-0.9	26	25	1.21	94	-0.5	18	17	1.23	95
		β_2	-2.8	16	15	1.08	93	-1.2	11	10	1.14	94
		β_3	-0.3	29	28	0.94	94	0.1	20	20	0.98	95
	omn111	β_0	2.5	24	23	1.08	94	1.2	16	16	1.11	96
		β_1	-0.9	26	25	1.21	94	-0.6	18	18	1.20	95
		β_2	-2.7	16	15	1.09	94	-1.2	11	11	1.13	94
		β_3	-0.3	28	26	1.06	95	0.1	19	19	1.09	95
ρ=0.4	ave111	β_0	2.3	24	23	1.09	95	0.9	17	16	1.09	95
		β_1	-0.4	26	25	1.19	94	-0.1	18	18	1.18	95
		β_2	-2.6	15	15	1.12	95	-1.1	11	11	1.13	94
		β_3	-0.6	27	27	1.09	95	0.2	19	19	1.10	96
	平均101	β_0	2.5	24	23	1.06	95	0.9	17	16	1.08	95
		β_1	-0.5	26	25	1.24	94	-0.1	18	17	1.23	94
		β_2	-2.6	15	15	1.14	95	-1.1	11	11	1.16	94
		β_3	-0.7	29	28	0.99	95	0.1	20	20	1.00	95
	agg111	β_0	2.2	24	22	1.10	95	0.6	16	16	1.14	94
		β_1	-0.4	25	24	1.26	93	0.1	18	17	1.27	95
		β_2	-2.5	16	15	1.09	94	-1.1	11	10	1.16	93
		β_3	-0.5	26	25	1.18	95	0.1	18	18	1.21	95
	聚合101	β_0	2.6	24	23	1.02	94	0.8	17	16	1.08	95
		β_1	-0.7	25	24	1.26	93	0.0	18	17	1.27	95
		β_2	-2.6	16	15	1.09	94	-1.1	11	10	1.17	93
		β_3	-0.7	29	28	0.95	95	0.1	20	20	0.98	95
	omn111	β_0	2.4	24	23	1.10	94	0.8	16	16	1.14	95
		β_1	-0.6	25	24	1.29	94	-0.1	18	17	1.28	94
		β_2	-2.6	15	15	1.12	94	-1.2	11	10	1.17	94
		β_3	-0.6	27	26	1.10	95	0.1	19	18	1.13	95
ρ=0.8	ave111	β_0	2.3	24	23	1.10	95	0.7	17	16	1.11	95
		β_1	-0.3	26	25	1.23	95	0.0	18	17	1.22	94
		β_2	-2.5	15	15	1.14	95	-1.0	11	10	1.16	95
		β_3	-0.6	27	26	1.11	95	0.3	19	19	1.13	95
	平均101	β_0	2.4	24	23	1.05	94	0.7	17	16	1.08	95
		β_1	-0.4	25	24	1.27	94	0.0	18	17	1.26	94
		β_2	-2.4	15	15	1.15	95	-1.0	11	10	1.19	94
		β_3	-0.5	29	28	0.99	95	0.2	20	20	1.00	95
	agg111	β_0	2.3	24	22	1.08	93	0.4	16	16	1.13	94
		β_1	-0.2	26	25	1.17	93	0.4	19	17	1.16	93
		β_2	-2.2	16	15	1.01	93	-0.8	11	10	1.11	94
		β_3	-1.1	26	24	1.24	94	0.1	18	17	1.29	94
	聚合101	β_0	2.4	25	23	0.98	93	0.5	17	16	1.04	95
		β_1	-0.5	26	25	1.17	93	0.2	19	17	1.17	94
		β_2	-2.2	16	15	1.02	93	-0.9	11	10	1.12	94
		β_3	-0.7	30	28	0.94	94	0.1	20	20	0.98	95
	omn111	β_0	2.3	24	22	1.09	94	0.6	16	16	1.14	95
		β_1	-0.4	25	24	1.27	93	0.1	18	17	1.25	94
		β_2	-2.4	15	15	1.09	93	-0.9	11	10	1.17	94
		β_3	-0.8	27	26	1.13	94	0.1	18	18	1.17	94

表4*Estimated relative efficiency between MLE without considering secondary data and estimators based on different data integration schemes*

	single100	single010	single001	ave111	agg111	omn111
截距	1.15	1.06	1.68	1.39	1.81	1.69
BMI	1.06	1.05	1.07	1.09	1.14	1.12
饮酒	1.14	2.62	1.00	1.96	2.81	2.87
吸烟	1.10	2.87	1.01	2.05	2.88	3.03
血红蛋白	1.16	1.08	2.03	1.53	2.15	2.00
Age	1.13	1.02	1.04	1.08	1.18	1.12

表5

Summary for estimators agg111 and omn111. ASE: asymptotic standard errors; ERE: estimated relative efficiency compared to MLE; OR: odds ratio; LL: lower limit; UL: upper limit; P-value A: P-values for agg111; P-value O: P-values for omn111; P-value MLE: P-values for MLE.

	agg111							
	估计值	ASE	ERE	OR	LL	UL	agg111 的 P 值	P值 MLE
截距	-5.165	0.880	1.805	0.006	0.001	0.032	0.000	0.000
BMI	0.046	0.015	1.142	1.047	1.016	1.078	0.003	0.011
饮酒	0.214	0.077	2.808	1.239	1.066	1.441	0.005	0.368
吸烟	-0.126	0.086	2.880	0.882	0.745	1.044	0.143	0.668
血红蛋白	0.027	0.041	2.146	1.027	0.948	1.113	0.510	0.488
Age	0.055	0.010	1.181	1.056	1.035	1.078	0.000	0.000
	omn111							
	估计值	ASE	ERE	OR	LL	UL	P值 O	P值 MLE
截距	-4.831	0.910	1.687	0.008	0.001	0.047	0.000	0.000
BMI	0.043	0.015	1.118	1.043	1.012	1.075	0.006	0.011
饮酒	0.174	0.076	2.869	1.190	1.025	1.381	0.022	0.368
吸烟	-0.126	0.084	3.030	0.882	0.748	1.039	0.133	0.668
血红蛋白	0.033	0.042	2.002	1.034	0.951	1.124	0.432	0.488
Age	0.048	0.011	1.121	1.049	1.028	1.071	0.000	0.000

arXiv:2303.03512v1 [stat.ME] 6 Mar 2023

《基于 Chixiang Chen、Ming Wang 和 Shuo Chen 的参数推断主要分析：整合多个次级数据集信息的高效数据整合方案》的补充材料

在本补充材料中，我们将介绍正则条件（A节）、论文正文中定理和性质的证明（B、C 和 D 节）、额外的模拟结果（E 节）、对缺失数据问题的扩展（F 节）、过度识别估计函数的示例（G 节），以及关于估计效率的进一步讨论（H 节）。这些证明基于经典的矩条件，这些条件广泛应用于矩估计法（Newey 和 McFadden） [1994] 以及经验似然（Qin 和 Lawless） [1994]。记号 $\|\cdot\|$ 表示 $L2$范数。

A 正则条件

条件 1 $E\{\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta})\}=\mathbf{0}$ 当且仅当 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$。

条件 2 对于 $k=1$、.....、K，存在参数 $\boldsymbol{\theta}_k^*$ 的取值，使得 $E\{\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)\}=\mathbf{0}$。

条件 3 函数 $\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta})$ 二阶连续可微，$E\|\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta}_0)\|^2$有限，并且在 $\boldsymbol{\beta}_0$ 的邻域内，$\|\partial^2\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\|$ 可由某个可积函数控制。

条件 4 对于条件 *1* 中定义的 $k=1$、...、K 和 $\boldsymbol{\theta}_k^*$，$E\{\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k^*)\boldsymbol{h}_k^T(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k^*)\}$是正定的，并且 $\{\partial^2\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k)\}/(\partial\boldsymbol{\theta}_k^T\partial\boldsymbol{\theta}_k)$ 在 $\boldsymbol{\theta}_k^*$ 的邻域内连续。此外，在 $\boldsymbol{\theta}_k^*$ 附近，$\|\{\partial\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k)\}/(\partial\boldsymbol{\theta}_k^T)\|$，$\|\{\partial^2\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)\}/(\partial\boldsymbol{\theta}_k^T\partial\boldsymbol{\theta}_k)\|$，和 $\|\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k)\|^3$ 均可由某个可积函数控制。

条件 1 和条件 2 有助于主要关注参数的估计一致性。条件 3 是一项正则性条件，用于保证泰勒展开有效 [Newey 和 McFadden,1994]。条件 4 包含为确保经验似然推断有效而采用的额外矩条件 [Qin 和 Lawless, 1994]。接下来的两节将详细证明主论文中的定理 1 和定理 2。定理 3 的证明可通过结合定理 1 和定理 2 的证明类似地推导得出。因此，我们将省略细节。

B 定理 1 的证明

在此证明中，我们考虑有 K 个次级数据集的一般情形。为了解决主论文式 (3) 中的约束最大化问题，我们采用经验似然理论来辅助证明 [Qin 和 Lawless, 1994]。具体而言，首先将拉格朗日乘子 $\boldsymbol{\lambda}_k$ 引入优化过程。随后，通过对 $\prod_{i=1}^n p_{ki}$ 关于 p_{ki} 和 $\boldsymbol{\theta}_k$进行最大化，可求得估计权重 $\hat{}p_{ki}=$

$(1/n)/\{1+\hat{\boldsymbol{\lambda}}_k^T R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\hat{\boldsymbol{\theta}}_k)\}$，其中，估计的拉格朗日乘子 $\hat{\boldsymbol{\lambda}}_k$ 和经验似然估计量 $\boldsymbol{\theta}_k$ 可通过求解以下估计方程得到

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)}{1+\boldsymbol{\lambda}_k R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)}=\mathbf{0};\\&\frac{1}{n}\sum_{i=1}^{n}\frac{R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)}{1+\boldsymbol{\lambda}_k R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)}\left\{\frac{\partial R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k)}{\partial\boldsymbol{\theta}^T}\right\}^T\boldsymbol{\lambda}_k=\mathbf{0}.\end{aligned}\tag{1}$$

在条件 4 下，将 (1) 中的方程分别在 $\boldsymbol{\lambda}_k=\mathbf{0}$ 和 $\boldsymbol{\theta}_k^*$ 处展开，可推导出 $\boldsymbol{\lambda}_k$ 的渐近表示，即对每个 $k=1$、.....、K，

$$\hat{\boldsymbol{\lambda}}_k=\frac{1}{n}\boldsymbol{S}_k\boldsymbol{Q}_k(\boldsymbol{\theta}_k^*)+o_p(\boldsymbol{n}^{-\frac{1}{2}}),\tag{2}$$

其中 $\boldsymbol{Q}_k(\boldsymbol{\theta}_k^*)=\sum_{i=1}^n R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)$。符号 $\boldsymbol{S}_k$ 在主论文的定理 1 中定义。另一方面，在条件 4 下再次进行泰勒展开，可进一步将每个 k 对应的估计权重表示为 $\hat{p}_{ki}=(1/n)[1-\hat{\boldsymbol{\lambda}}_k^T R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k^*)\{1+o_p(1)\}]$。结合主论文 (6) 中的估计方程，可得

$$\begin{aligned}\mathbf{0}=&\sum_{i=1}^{n}\bar{p}_i\boldsymbol{g}(\boldsymbol{D}_{0i};\bar{\boldsymbol{\beta}})\\=&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_{0i};\bar{\boldsymbol{\beta}})\left[1-\sum_{k=1}^{K}\omega_k\hat{\boldsymbol{\lambda}}_k^T R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k^*)\big\{1+o_p(1)\big\}\right]\\=&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\&-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_k R_{ki}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k^*)^T\hat{\boldsymbol{\lambda}}_k+o_p(\boldsymbol{n}^{-\frac{1}{2}}).\end{aligned}$$

第二个等式是将 $\bar{p}_i$ 替换为 $\sum_{i=1}^K\omega_k\hat{p}_{ki}$，并利用 $\sum_{i=1}^K\omega_k=1$ 得到的。第三个等式基于在条件 3 下关于 $\boldsymbol{\beta}_0$ 的泰勒展开。因此，估计量 $\bar{\boldsymbol{\beta}}$ 的渐近展开可写为

$$\begin{aligned}&n^{\frac{1}{2}}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\=&-\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}\right)^{-1}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)\right.\\&\left.-n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_k\Big\{R_{ki}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)\boldsymbol{h}_k(\boldsymbol{D}_{ki};\boldsymbol{\theta}_k^*)^T\Big\}\hat{\boldsymbol{\lambda}}_k\right]+o_p(\mathbf{1})\\=&-\boldsymbol{\Gamma}^{-1}\left\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_{0i};\boldsymbol{\beta}_0)-n^{-\frac{1}{2}}\sum_{k=1}^{K}\omega_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k\boldsymbol{Q}_k(\boldsymbol{\theta}_k^*)\right\}+o_p(\mathbf{1}).\end{aligned}$$

最后，通过简单的矩阵代数，我们可以得到 $n^{1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$, 的渐近正态性，其中主论文定理 1 中的渐近协方差矩阵可以根据上述推导的影响函数并令 $K=2$ 后轻松计算得到。

C 定理 2 的证明

在此证明中，我们考虑存在 K 个辅助数据集的一般情形。根据主论文中的式 (8)、估计权重 $\hat{}p_{ki}$ 的渐近展开以及 B 节中推导的拉格朗日乘子 $\hat{}\,\boldsymbol{\lambda}_k$ ，我们有

$$
\begin{aligned}
\mathbf{0} =& \sum_{i=1}^{n} \tilde{p}_i \boldsymbol{g}(\boldsymbol{D}_{0i}^{u}; \tilde{\boldsymbol{\beta}}) \\
=& \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_{0i}; \tilde{\boldsymbol{\beta}}) \left[1 - \sum_{k=1}^{K} \hat{\boldsymbol{\lambda}}_k^T R_{ki} \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k^*) \{1 + o_p(1)\} \right] + o_p\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_{0i}; \tilde{\boldsymbol{\beta}}) \right) \\
=& \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0) + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
& - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} R_{ki} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0) \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k^*)^T \hat{\boldsymbol{\lambda}}_k + o_p(\boldsymbol{n}^{-\frac{1}{2}}),
\end{aligned}
$$

第二个等式由将 $\tilde{p}_i$ 替换为 $\prod_{i=1}^{K} \hat{}p_{ki}$ 以及以下事实得到：

$$
\max_{i=1,\ldots,n} \hat{\boldsymbol{\lambda}}_k^T R_{ki} \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k^*) = o_p(1),
$$

这基于条件 4 [下的经验似然理论 Owen, 2001]。第三个等式基于条件 3 下关于 $\boldsymbol{\beta}_0$ 的泰勒展开。因此，估计量 $\boldsymbol{\beta}$ 的˜渐近展开可写为

$$
\begin{aligned}
& n^{\frac{1}{2}} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
=& - \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T} \right)^{-1} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0) \right. \\
& \left. - n^{-\frac{1}{2}} \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ R_{ki} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0) \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k^*)^T \right\} \hat{\boldsymbol{\lambda}}_k \right] + o_p(\mathbf{1}) \\
=& - \boldsymbol{\Gamma}^{-1} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}_0) - n^{-\frac{1}{2}} \sum_{k=1}^{K} \boldsymbol{\Lambda}_k \boldsymbol{S}_k \boldsymbol{Q}_k(\boldsymbol{\theta}_k^*) \right\} + o_p(\mathbf{1}).
\end{aligned}
$$

最后，通过简单的矩阵代数，我们可以得到 $n^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$,的渐近正态性，其中主论文中定理 2 的渐近协方差矩阵可根据上述推导出的影响函数并令 $K = 2$而容易地计- 算得到。

主论文中定理 3 的证明可通过结合定理 1 和定理 2 的证明以类似方式推导得到。此处省略细节。

D 其他估计形式

在主论文第 2 节中，我们提出了重新加权估计式（2）和（3），这与 Chen 等人 [2021] 介绍的重新加权估计相同。简而言之，Chen 等人

[2021] 假设前 m_1 位受试者拥有次级数据，其余受试者没有辅助记录。然后，参数 $\boldsymbol{\beta}$ 可通过以下方式求解

$$\sum_{i=1}^{m_1} m_i \hat{p}_{ki}^{*} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) + \sum_{i=m_1+1}^{n} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) = \mathbf{0}, \tag{3}$$

其中，非负权重 $\hat{p}_{ki}^{*}$ 通过在以下约束条件下最大化 $\prod_{i=1}^{m_1} p_{ki}$ 求得

$$\sum_{i=1}^{m_1} p_{ki}^{*} = 1, \sum_{i=1}^{m_i} p_{ki}^{*} \boldsymbol{h}(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k) = \mathbf{0}. \tag{4}$$

为说明两种估计形式之间的等价性，我们首先注意到，主论文中式 (2) 的估计权重等于 $\hat{p}_{ki} = n^{-1}/\{1 + \hat{\boldsymbol{\lambda}}_k^T R_{ki} \boldsymbol{h}_k(\boldsymbol{D}_{ki}$；应用拉格朗日乘子法 [Qin 和 Lawless, 1994] 对于 $i = 1, \ldots, m_1$可得 $\hat{\boldsymbol{\theta}}_k)\}$。当 $i = m_1 + 1, \ldots, n$时，这些权重将变为 $1/n$。随后，主论文中式 (3) 的约束条件变为 $\sum_{i=1}^{m_1} p_{ki} = m_1/n$ 和 $\sum_{i=1}^{m_1} p_{ki} \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k) = \mathbf{0}$，主论文中式 (2) 的估计方程则变为 $\sum_{i=1}^{m_1} \hat{p}_{ki} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) + \sum_{i=m_1+1}^{n} (1/n) \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) = \mathbf{0}$。因此，注意到 $\hat{p}_{ki} = (m_1/n)\hat{p}_{ki}^{*}$这一事实，主论文中的式 (2) 和 (3) 可简化为上文的式 (3) 和 (4)。

除了主论文中的 (3) 和 (4) 外，所提出的估计程序 (2) 和 (3) 渐近等价于以下经验似然估计，即关于 p_{ki}、$\boldsymbol{\beta}$、$\boldsymbol{\theta}_k$ 最大化 $\prod_{i=1}^{n} p_{ki}$，并满足三个约束条件

$$\sum_{i=1}^{n} p_{ki} = 1, \sum_{i=1}^{n} p_{ki} \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) = \mathbf{0}, \sum_{i=1}^{n} p_{ki} R_{ki} \boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k) = \mathbf{0}. \tag{5}$$

为证明渐近等价性，我们应用 Qin 和 Lawless [1994] 的结果，并引入两个拉格朗日乘子 $\boldsymbol{\lambda}_{k1}$ 和 $\boldsymbol{\lambda}_{k2}$，分别对应上述两个矩约束。为简化记号，在以下推导中省略下标 k。容易验证，$\boldsymbol{\lambda}_1$ 和 $\boldsymbol{\lambda}_2$ 是以下方程的解 [Qin 和 Lawless, 1994]

$$\begin{aligned} \boldsymbol{Q}_{1n} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta})}{1 + \boldsymbol{\lambda}_1^T \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) + \boldsymbol{\lambda}_2^T R_i \boldsymbol{h}(\boldsymbol{D}_i; \boldsymbol{\theta})} = \mathbf{0} \\ \boldsymbol{Q}_{2n} &= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i \boldsymbol{h}(\boldsymbol{D}_i; \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}_1^T \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) + \boldsymbol{\lambda}_2^T R_i \boldsymbol{h}(\boldsymbol{D}_i; \boldsymbol{\theta})} = \mathbf{0}, \end{aligned} \tag{6}$$

$\hat{}\hat{}$和 $\boldsymbol{\beta}$ 和 $\boldsymbol{\theta}$ 是以下方程的解

$$\begin{aligned} \boldsymbol{G}_{1n} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\left\{\partial \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^T\right\}^T \boldsymbol{\lambda}_1}{1 + \boldsymbol{\lambda}_1^T \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) + \boldsymbol{\lambda}_2^T R_i \boldsymbol{h}(\boldsymbol{D}_i; \boldsymbol{\theta})} = \mathbf{0} \\ \boldsymbol{G}_{2n} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\left\{\partial R_i \boldsymbol{h}(\boldsymbol{D}_i; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T\right\}^T \boldsymbol{\lambda}_2}{1 + \boldsymbol{\lambda}_1^T \boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta}) + \boldsymbol{\lambda}_2^T R_i \boldsymbol{h}(\boldsymbol{D}_i; \boldsymbol{\theta})} = \mathbf{0}. \end{aligned} \tag{7}$$

在 $\boldsymbol{\beta}_0$、 $\boldsymbol{\theta}^*$、 $\boldsymbol{\lambda}_1 = \mathbf{0}$ 和 $\boldsymbol{\lambda}_2 = \mathbf{0}$ 处对式 (6) 和 (7) 进行泰勒展开，可得

$$\begin{aligned}
\mathbf{0} =& \boldsymbol{Q}_{1n} + \frac{\partial \boldsymbol{Q}_{1n}}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{\partial \boldsymbol{Q}_{1n}}{\partial \boldsymbol{\lambda}_1^T}\hat{\boldsymbol{\lambda}}_1 + \frac{\partial \boldsymbol{Q}_{1n}}{\partial \boldsymbol{\lambda}_2^T}\hat{\boldsymbol{\lambda}}_2 + o_p(\boldsymbol{n}^{-1/2}) \\
\mathbf{0} =& \boldsymbol{Q}_{2n} + \frac{\partial \boldsymbol{Q}_{2n}}{\partial \boldsymbol{\theta}^T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \frac{\partial \boldsymbol{Q}_{2n}}{\partial \boldsymbol{\lambda}_1^T}\hat{\boldsymbol{\lambda}}_1 + \frac{\partial \boldsymbol{Q}_{2n}}{\partial \boldsymbol{\lambda}_2^T}\hat{\boldsymbol{\lambda}}_2 + o_p(\boldsymbol{n}^{-1/2}) \\
\mathbf{0} =& \boldsymbol{G}_{1n} + \frac{\partial \boldsymbol{G}_{1n}}{\partial \boldsymbol{\lambda}_1^T}\hat{\boldsymbol{\lambda}}_1 + \frac{\partial \boldsymbol{G}_{1n}}{\partial \boldsymbol{\lambda}_2^T}\hat{\boldsymbol{\lambda}}_2 + o_p(\boldsymbol{n}^{-1/2}) \\
\mathbf{0} =& \boldsymbol{G}_{2n} + \frac{\partial \boldsymbol{G}_{2n}}{\partial \boldsymbol{\lambda}_1^T}\hat{\boldsymbol{\lambda}}_1 + \frac{\partial \boldsymbol{G}_{2n}}{\partial \boldsymbol{\lambda}_2^T}\hat{\boldsymbol{\lambda}}_2 + o_p(\boldsymbol{n}^{-1/2}),
\end{aligned}$$

由于在 $\boldsymbol{\beta}_0$、 $\boldsymbol{\theta}^*$、 $\boldsymbol{\lambda}_1 = \mathbf{0}$和 $\boldsymbol{\lambda}_2 = \mathbf{0}$处取值时，
$\partial \boldsymbol{Q}_{1n}/\partial \boldsymbol{\theta}^T$, $\partial \boldsymbol{Q}_{2n}/\partial \boldsymbol{\beta}^T$, $\partial \boldsymbol{G}_{1n}/\partial \boldsymbol{\beta}^T$, $\partial \boldsymbol{G}_{1n}/\partial \boldsymbol{\theta}^T$, $\partial \boldsymbol{G}_{2n}/\partial \boldsymbol{\beta}^T$,和$\partial \boldsymbol{G}_{2n}/\partial \boldsymbol{\theta}^T$ 均为零矩阵，因此上述结论成立。通过求解上述方程组，我们可以得到主要参数估计的渐近展开

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -\frac{\partial \boldsymbol{Q}_{1n}}{\partial \boldsymbol{\beta}^T}(n^{1/2}\boldsymbol{Q}_{1n} + n^{1/2}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{Z}) + o_p(\mathbf{1}), \tag{8}$$

其中 $\boldsymbol{B} = (\partial \boldsymbol{Q}_{1n}/\partial \boldsymbol{\lambda}_2^T, \partial \boldsymbol{Q}_{1n}/\partial \boldsymbol{\lambda}_1^T, \mathbf{0})$, $\boldsymbol{Z} = (-\boldsymbol{Q}_{2n}^T, -\boldsymbol{G}_{1n}^T, -\boldsymbol{G}_{2n}^T)^T$,和

$$\boldsymbol{D} = \begin{bmatrix} \frac{\partial \boldsymbol{Q}_{2n}}{\partial \boldsymbol{\lambda}_2^T} & \frac{\partial \boldsymbol{Q}_{2n}}{\partial \boldsymbol{\lambda}_1^T} & \frac{\partial \boldsymbol{Q}_{2n}}{\partial \boldsymbol{\theta}^T} \\ \mathbf{0} & \frac{\partial \boldsymbol{G}_{1n}}{\partial \boldsymbol{\lambda}_1^T} & \mathbf{0} \\ \frac{\partial \boldsymbol{G}_{2n}}{\partial \boldsymbol{\lambda}_2^T} & \mathbf{0} & \mathbf{0}. \end{bmatrix}$$

经过一些代数运算，项 $\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{Z}$ 最终变为补充材料 [Chen 等人（ 2021], ）中式（2）所定义的$-(1/n)\boldsymbol{\Lambda}\boldsymbol{S}\boldsymbol{Q}_{m_1}(\boldsymbol{\theta}^*)$ ，这是基于研究中前 m_1 名受试者有次级数据，而其余受试者没有次级记录这一假设。因此，式（8）中的表达式也等于本文估计量 $\boldsymbol{\beta}_{EN}$ 的渐近展开。此外，基于 $E\{\boldsymbol{g}(\boldsymbol{D}_{0i}$、 $\boldsymbol{\beta})\} = \mathbf{0}$和 $E\{R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki}$的信息，以上经验似然估计量是半参数有效的； $\boldsymbol{\theta}_k)\} = \mathbf{0}$遵循 Qin [2017]第 6.4 节中的证明策略。因此，主论文中提出的估计方案（2）和（3）也是半参数有效的。与（6）和（7）相比，我们提出的估计方案在经验似然框架中所需的函数维数和调优参数更低，因此数值上更稳定，计算量也更小。此外，对于可获得多个次要结局（$k = 1, \ldots, K$）的情形，可以通过我们的加权估计或经验似然，将过度识别函数$R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k)$ 替换为扩展后的过度识别函数，从而在概念上构造半参数有效估计量

$$\boldsymbol{h}(\boldsymbol{D}_{1i}, \ldots, \boldsymbol{D}_{Ki}, R_{1i} \ldots, R_{Ki}; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) = (R_{1i}\boldsymbol{h}_1^T(\boldsymbol{D}_{1i}; \boldsymbol{\theta}_1), \ldots, R_{Ki}\boldsymbol{h}_K^T(\boldsymbol{D}_{Ki}; \boldsymbol{\theta}_K))^T.$$

然而，当纳入更多次级数据，且这些数据的类型和尺度各不相同时，堆叠多个过度识别函数的维度可能会非常高。由于数据结构多维且复杂，估计算法可能无法收敛（这一点已在我们此前的模拟中发现），从而阻碍该方法的应用。另一方面，我们的 MinBo估计量是一种有效的解决方案，使得在有多个次要结局时，信息整合能够在实践中实现。请注意，如果我们只假设 $\boldsymbol{D}_{0i}$ 具有参数形式，而不对其他所有部分作任何设定，那么半参数理论（如果存在）会变得复杂得多；这值得投入大量精力，可留待未来工作探讨。

E 附加模拟结果

本节给出主论文中更详细的设定和额外的模拟结果。前两个次要模型中的协变量和残差按~ ~如下方式生成：时变协变量向量 $\boldsymbol{X}_{1i1} = \boldsymbol{X}_{2i1}$ 服从多元伯努利分布，成功概率为 0.5，且具有可交换相关矩阵，相关系数为 0.3；另一个时变协变量向量 $\boldsymbol{X}_{1i2} = \boldsymbol{X}_{2i2}$ 服从均值为零、方差为一的多元正态分布，且具有可交换相关结构~ ~，相关系数为 0.3；最后一个协变量向量 $\boldsymbol{X}_{1i3} = \boldsymbol{X}_{2i3}$ 为时间无关变量，其所有分量均相等，并服从成功概率为 0.5 的伯努利分布。两个数据集中的残差向量 $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{1i}^T, \boldsymbol{\epsilon}_{2i}^T)^T$，其中$\boldsymbol{\epsilon}_{ki} = (\epsilon_{ki1}, \ldots, \epsilon_{ki4})^T$，对于 $k = 1, 2$，服从均值为零向量、方差 - 协方差

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}, \tag{9}$$

其中，$\boldsymbol{V}_{11}$ 和 $\boldsymbol{V}_{22}$ 分别是 $\boldsymbol{\epsilon}_{1i}$ 和 $\boldsymbol{\epsilon}_{2i}$的方差-协方差矩阵。$\boldsymbol{V}_{11}$ 的对角线元素均为 1，非对角线元素均为 0。8；$\boldsymbol{V}_{22}$ 的对角线元素均为 1，非对角线元素均为 0。5；矩阵 $\boldsymbol{V}_{12} = \boldsymbol{V}_{21}^T$ 的所有元素均为 ρ，用于控制两个纵向次要结局之间的关联强度。

此外，为了构造过度识别的工作估计函数，我们采用四个基矩阵 $\boldsymbol{V}_1$、$\boldsymbol{V}_2$、$\boldsymbol{V}_3$和 $\boldsymbol{V}_4$，其中 $\boldsymbol{V}_1$ 为单位矩阵，$\boldsymbol{V}_2$ 是对角线元素为 0、非对角线元素为 1 的矩阵，$\boldsymbol{V}_3$ 是两个主非对角线元素为 1、其余元素为 0 的矩阵，$\boldsymbol{V}_4$ 是左上角和右下角元素为 1、其余元素为 0 的矩阵。这些矩阵在以往研究中被广泛使用 [Qu 等人，2000年，Tang 和 Leng，2011年，Chen 等人，2021]。

图 F 展示了本文模拟研究所用数据结构示意图；表 S.1 总结了相关系数为 ρ 的两个纵向辅助变量之间等于 0。4 时八种估计方法的总体评估；表 S.2 总结了 $\rho = 0$ 下的结果；表 S.3 总结了存在部分观测的次级数据时经验相对效率（ERE）的结果；表 S.4 总结了观测到部分次级数据时，均值结构设定错误情况下的总体评估。

E.1 信息性缺失

我们还研究了次级数据中的信息性缺失 [Enders、2010] 对主要参数估计的影响。为了生成具有非随机缺失的部分观测次要数据，我们为 k^{th} 次级数据模拟信息性观测指示变量 R_{ki}。R_{ki}的成功概率建模为 $\{1 + \exp(-\boldsymbol{X}_{0i}^T\boldsymbol{\alpha})\}^{-1}$，其中 $k = 1, 2,3$，且$\boldsymbol{\alpha} = (0.5, 1, 1, 1)^T$和 $\boldsymbol{X}_{0i}$的定义见主论文第3.1节。其他所有设置均与主论文第3.1节相同。为了评估 MinBo 的性能，我们仅使用主论文第 3.2 节所述的错误设定的均值结构来构建三个次要模型，并且仅使用具有 $R_{ki} = 1$ 的次级数据来拟合模型。由于观测指示变量 R_{ki} 取决于潜在结果并导致均值结构设定错误，因此会产生信息性缺失问题。每个次级数据的缺失率

约为 25%，结果汇总于表 S.5。对于存在信息性缺失且均值结构设定错误的次级数据，主模型中提出的估计值仍具有较小的估计偏差，且覆盖概率令人满意，接近 95%名义水平。此外，就经验相对效率而言，在这种情况下，MinBo 仍能提高估计效率。

E.2 主要结局与次要结局之间关联甚微

我们考虑这样一种设定：主数据和三组次级数据的生成方式与主论文第3.1节所示相同，不同之处在于，此时主要结局与三个次要结局之间的相关性非常弱（例如，相关系数约为0.15）。我们还考虑了第一和第二个次要结局之间相关系数 ρ 分别为 0、0.4 和 0.8 的不同情形。样本量为 600 时的结果总结于下方所示的表 S.6。根据表 S.6，我们发现，基于平均方案和综合方案的估计值，其 ERE 值仅略高于 1。请注意，当 ρ 较大时，基于聚合方案的估计值的 ERE 值可能略低于 1。这是预期之中的，因为当次要结局彼此高度关联时，聚合方案可能会降低效率（主论文第2.3节）。此外，请注意，我们提出的方法“MinBo”依赖于主要结局与次要结局彼此高度关联这一假设。这意味着，如果在实践中这些结局之间的关联很弱，确实没有必要将这些次要结局纳入主要分析。此外，即使关联强度非常弱，与不考虑次级数据相比，我们的方法仍可带来相当或略高的效率。

E.3 与联合似然方法进行比较

为将我们的方法与JL方法进行比较，我们基于多元正态分布生成一个横断面主要结局和两个纵向次级结局，其中三个结局的联合似然易于指定。具体而言，我们考虑一个横断面主要结局和两个包含四个时间点的纵向次级结局。三个结局的残差基于均值为零、方差为1的多元正态分布生成。三个结局之间的相关模式如下：每个次级结局内部各时间点之间的相关性为0.8；两个次级结局之间的相关性为0.6；主要结局与次级结局之间的相关性为0.8。主要模型和两个次要模型中所考虑的协变量和效应量与主论文第3.1节中的相同。

我们考虑通过平均和聚合方案从两个次要结局中借用信息，并将其与最大似然估计和 JL估计进行比较。JL方法基于具有非结构化相关模式的多元正态分布，对三个结局进行联合建模。为检验次要模型设定错误时的表现，我们考虑均值结构设定错误的情况：拟合的两个次要结局的均值仅包含两个协变量（真实均值结构中总共有三个协变量）。结果汇总于表 S.8。当次要模型设定正确时，我们观察到，JL估计和 MinBo估计的偏差都很小；就经验相对效率（ERE）而言，JL估计优于MinBo估计量（即最大似然估计与目标估计值的均方误差之比，该比值越大越理想，越大越好）。这是意料之中的，因为根据参数理论，JL估计是效率最高的{v1}Newey 和 McFadden，{v2}。

然而，当次要模型设定错误时，JL估计会产生较大的偏差，且经验相对效率（ERE）极低（接近于零）；而 MinBo估计量的偏差仍然很小，并且在经验相对效率（ERE）方面具有显著的效率增益。这一评估进一步凸显了我们提出的方法的优势，更适合实际应用。

E.4 相关性更高的设置

为评估主要结局与次要结局之间的关联强度对效率增益的影响，我们生成了另一组数据，其设置与主论文中的数据相同，但主要结局与三个次要结局之间的条件相关系数趋于更高（主要结局与两个纵向次要结局之间约为0.7；主要结局与横断面次级结局之间约为0.6。在主论文的设置中，这两个相关系数分别仅约为0.45和0.3）。具体而言，我们首先生成一个服从标准正态分布的随机变量 Z_{0i} ，然后使用该变量生成主要结局 $Y_{0i}=1$ ：当 $Z_0 \geqslant x_{0i}$ 时（其中 x_{0i} 的定义见主论文第3.1节），否则为 $Y_{0i}=0$ 。随后，我们生成两个纵向次要结局，具体为$\boldsymbol{Y}_{ki}=\boldsymbol{X}_{ki}\boldsymbol{\theta}_k+\boldsymbol{Y}_{0i}-\boldsymbol{p}_{0i}+0.4\boldsymbol{\epsilon}_{ki}$，其中 $k=1,2$（ $\boldsymbol{X}_{ki}$、 $\boldsymbol{\theta}_k$、 p_{0i}和 $\boldsymbol{\epsilon}_{ki}$ 的定义见主论文第3.1节。此处， $\boldsymbol{Y}_{0i}-\boldsymbol{p}_{0i}$ 中的值对于受试者 i的所有观测均相同）；并通过以下方式生成二元次要结局 Y_{3i} ：当 $-Z_0 \geqslant x_i^*$时为 $Y_{3i}=1$ ，否则为 $Y_{3i}=0$（其中 x_i^* 的定义见主论文第3.1节）。其他所有设置均与主论文第3.1节中的设置相同。从表S.7可以看出，与主论文（表2）中的结果相比，即使所有次要结局模型均设定错误，所得的MinBo估计量的效率增益仍显著提高（从 $1.2-1.3$ 提高至 $1.5-1.7$）。

F 主模型中缺失数据问题的扩展

在主论文中，我们假设所有主数据均已观测，或完全随机缺失，但在某些应用中，这一假设可能过于理想。然而，只需进行少许修改，我们的方法即可轻松扩展，以处理主模型中的缺失数据问题。为此，我们引入一种称为逆概率权重的技术，这是处理主数据随机缺失问题的常用方法 [Enders, 2010]。具体而言，首先令观测指示变量为 R_{0i}：如果来自 i^{th} 受试者的主数据已观测，则其取值为 1，否则为 0。然后，我们将观测到 i^{th} 受试者数据的概率定义为 $\pi_i(\boldsymbol{\alpha})=E(R_{0i}|\tilde{\boldsymbol{X}}_{0i})$，其中 $\tilde{\boldsymbol{X}}_{0i}$ 和 $\boldsymbol{\alpha}$ 分别为缺失数据模型中涉及的某些协变量和参数。此外，令缺失数据模型中参数 $\boldsymbol{\alpha}$ 的得分函数为 $\sum_{i=1}^{n}\boldsymbol{L}_i(\boldsymbol{\alpha})$。除上述内容外，其他设置与主论文第 2 节相同。然后，我们同时求解

$\boldsymbol{\beta}$ 和 $\boldsymbol{\alpha}$ 所对应的以下修正加权估计方程:

$$\sum_{i=1}^{n} p_i^* \tilde{\boldsymbol{g}}(\boldsymbol{D}_{0i}, \boldsymbol{\alpha}; \boldsymbol{\beta}) = \mathbf{0}, \tag{10}$$

其中，权重 p_i^* 由我们提出的方案（主论文第2.2、2.3和2.4节）重新构造； $\tilde{\boldsymbol{g}}(\boldsymbol{D}_{0i}, \boldsymbol{\alpha}; \boldsymbol{\beta}) = (R_{0i}/\pi_i(\boldsymbol{\alpha})\boldsymbol{g}^T(\boldsymbol{D}_{0i}; \boldsymbol{\beta}), \boldsymbol{L}_i^T(\boldsymbol{\alpha}))^T$，其中 $\boldsymbol{g}(\boldsymbol{D}_{0i}; \boldsymbol{\beta})$ 与主论文中的（1）相同。随后，遵循前两节证明的思路，可以类似地推导出渐近性质。存在高缺失率时，我们可以看到，唯一需要做的修改是为主要分析构造扩展估计函数 $\tilde{\boldsymbol{g}}(\boldsymbol{D}_{0i}, \boldsymbol{\alpha}; \boldsymbol{\beta})$。 $\tilde{\boldsymbol{g}}(\boldsymbol{D}_{0i}, \boldsymbol{\alpha}; \boldsymbol{\beta})$ 中的第一部分是逆概率加权框架下的典型估计函数。第二部分旨在使构造出的估计函数 $\tilde{\boldsymbol{g}}(\boldsymbol{D}_{0i}, \boldsymbol{\alpha}; \boldsymbol{\beta})$ 独立同分布，从而在理论上保证 $\boldsymbol{\beta}$ 的估计具有效率增益。本文并未将对此扩展的评估作为重点，因此将其视为未来工作。

G 过度识别估计函数

本节提供了用于拟合次要数据的过度识别估计函数示例。文献中有两种常用的过度识别估计函数 $\boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k)$。其中一种适用于重复次级结局，即 $\boldsymbol{Y}_{ki}$ 是在随访期间对每位受试者进行的重复测量:

$$\boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k) = \begin{pmatrix} \boldsymbol{Z}_{ki}^T \tilde{\boldsymbol{R}}_{ki}^{-1/2} \boldsymbol{V}_{1_k} \tilde{\boldsymbol{R}}_{ki}^{-1/2} \{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta}_k)\} \\ \cdots \\ \boldsymbol{Z}_{ki}^T \tilde{\boldsymbol{R}}_{ki}^{-1/2} \boldsymbol{V}_{\tau_k} \tilde{\boldsymbol{R}}_{ki}^{-1/2} \{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta}_k)\} \end{pmatrix}, \tag{11}$$

其中， $\boldsymbol{Z}_{ki} = \partial \boldsymbol{\mu}_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k^T$； $\tilde{\boldsymbol{R}}_{ki}$ 是一个对角矩阵，包含 $\boldsymbol{Y}_{ki}$中每个元素的方差； $\boldsymbol{\mu}_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta}_k)$ 是由参数 $\boldsymbol{\theta}_k$索引的 k^{th} 次要结局的条件均值。基于 $\boldsymbol{\theta}_k$、 r_k和 $\tau_k \geqslant 2$的维度， τ_k 矩阵 $\boldsymbol{V}_j$ 得到长度为 $r_k \times \tau_k$的过度识别估计函数。 $\boldsymbol{V}_{1_k}, \ldots, \boldsymbol{V}_{\tau_k}$可以是一系列基矩阵 [（Qu 等人，2000年；Tang 和 Leng，2011年；Chen 等人， 2021]）。另一种形式 $\boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k)$ 适用于横断面次级结局 Y_{ki}，即

$$\boldsymbol{h}_k(\boldsymbol{D}_{ki}; \boldsymbol{\theta}_k) = \boldsymbol{d}(\boldsymbol{X}_{ki}, \tilde{\boldsymbol{Z}}_{ki}; \boldsymbol{\theta}_k)\{Y_{ki} - \mu_k(\boldsymbol{X}_{ki}; \boldsymbol{\theta}_k)\}, \tag{12}$$

˜其中， $\boldsymbol{d}(\boldsymbol{X}_{ki}, \boldsymbol{Z}_{ki}; \boldsymbol{\theta}_k)$ 是一个维度大于$\boldsymbol{\theta}_k$的用户指定的向量函数，通常形式为 $\boldsymbol{d}(\boldsymbol{X}_{ki}, \tilde{\boldsymbol{Z}}_{ki}; \boldsymbol{\theta}_k) = (\boldsymbol{X}_{ki}^T, \tilde{\boldsymbol{Z}}_{ki}^T)^T$。注意，向量 $\tilde{\boldsymbol{Z}}_{ki}$包含满足 $E\{\boldsymbol{h}_k(\boldsymbol{D}_{ki}$的冗余变量； $\boldsymbol{\theta}_k^*)\} = \mathbf{0}$, 对于某些参数值$\boldsymbol{\theta}_k^*$。(11)和(12)中的两个工作函数均可用于连续型和分类结局。

H 关于估计效率的更多讨论

根据D节中的讨论，我们已证明，若具备以下信息，主文中的估计程序(2)和(3)是半参数有效的:

$E\{\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta})\} = \mathbf{0}$ 和 $E\{R_{ki}\boldsymbol{h}_k(\boldsymbol{D}_{ki};\ \boldsymbol{\theta}_k)\} = \mathbf{0}$。本节将进一步讨论该估计量的估计效率。

首先，所提出的估计量仍然是相合的，并且不违反半参数理论。注意，根据 M估计量 [Newey 和 Mc- Fadden，1994]，半参数估计量的估计一致性由 $E[\boldsymbol{W}(\boldsymbol{X}_0)\{Y_0-\mu(\boldsymbol{X}_0$ 保证；$\boldsymbol{\beta})\}] = \mathbf{0}$（$\boldsymbol{W}(\boldsymbol{X}_0)$ 是一个取决于协变量$\boldsymbol{X}_0$的权重函数），而我们所提出的估计量基于一阶矩 $E\{p_*\boldsymbol{g}(\boldsymbol{D}_0;\ \boldsymbol{\beta})\} = \mathbf{0}$，其中$p_*$ 是 $\hat{}p_i$ 的极限值，且 $\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta})$ 是某个有效的估计函数。根据经验似然理论，我们有 $n\hat{}p_i = 1 + o_p(1)$ [Qin 和 Lawless，1994年，Owen，2001],这意味着 $np_* = 1$。因此，就估计一致性所依赖的一阶矩而言，我们所提出的估计量没有引入主要结局与次要结局之间的依赖关系，并且通过实现$E\{np_*\boldsymbol{g}(\boldsymbol{D}_0;\ \boldsymbol{\beta})\} = E\{\boldsymbol{g}(\boldsymbol{D}_0;\ \boldsymbol{\beta})\} = \mathbf{0}$，所提出的估计量是相合的。

第二，所提出的估计量可能比不考虑任何次级数据的半参数估计量更有效率。从理论上讲，我们允许 $\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta})$采用一般形式，该形式可以是$\boldsymbol{\beta}$的任何有效估计函数。在本文中，我们聚焦于横断面研究中的二元结局，因此在数值研究（模拟和真实数据应用）中，通过考虑 logit链接函数，我们采用估计函数 $\boldsymbol{g}(\boldsymbol{D}_{0i};\ \boldsymbol{\beta}) =\boldsymbol{X}_{0i}\{Y_{0i} - \mu(\boldsymbol{X}_{0i};\ \boldsymbol{\beta})\}$ 的形式。请注意，这是逻辑回归中的得分函数，也是具有半参数效率的估计函数（实际上，该函数将得到最大似然估计量，即效率最高的估计量）。在此背景下，通过通过估计方程$\sum_{i=1}^{n}\hat{}p_i\boldsymbol{g}(\boldsymbol{D}_{0i},\ \boldsymbol{\beta}) = \mathbf{0}$考虑来自次要结局的信息，我们在理论和模拟中均证明，所提出的估计量比不进行信息整合的估计量更有效率。这意味着，所提出的估计量可能比在不考虑任何次级数据的情况下通过 $\sum_{i=1}^{n}[\boldsymbol{W}(\boldsymbol{X}_{0i})\{Y_{0i} - \mu(\boldsymbol{X}_{0i};\ \boldsymbol{\beta})\}] = \mathbf{0}$ 求解得到的半参数估计量更有效率，这一点已在横断面二元结局情境下得到证明。

最后，上述讨论可扩展至如下更一般的框架。考虑数据 Y_{0i}、$\boldsymbol{Y}_{si}$ 和 $\boldsymbol{X}_i$，其中 Y_{0i} 为主要结局，$\boldsymbol{Y}_{si}$ 为（多个）次要结局，$\boldsymbol{X}_i$ 为包含主要数据和次级数据中所有协变量的向量。随后，我们定义通过最大化联合似然 $L(Y_{0i},\ \boldsymbol{Y}_{si},\ \boldsymbol{X}_i)$（假设其设定正确）获得的最有效估计量（$E_1$）；通过考虑数据 Y_{0i} 和 $\boldsymbol{X}_i$得到的半参数估计量（E_2）；以及利用带有用于信息借用的权重 $\hat{}p_i$ 的半参数估计函数得到的我们所提出的估计量（E_3）。因此，这三个估计量的估计效率排序为 $E_1 \geqslant E_3 \geqslant E_2$。这意味着，我们所提出的估计量通过有效整合（多个）次要结局的信息，是介于最有效估计量与半参数有效估计量之间的一种折中估计量。请注意，完全似然估计量（E_1）在许多应用中可能无法获得，因为它要求对完整似然进行正确设定，而这在实践中难以实现（表 S.8）。值得注意的是，主要结局和（多个）次要结局可遵循不同的分布（如连续型/计数型/二元），甚至具有不同的数据结构（如横断面、纵向），这给分布设定带来了挑战。我们所提出的估计量为利用次要结局的信息以提高主要参数的效率提供了一种便捷且稳健的方法。

参考文献

Chixiang Chen、Peisong Han 和 FanHe。通过借用辅助数据中的信息改进主要分析。《医学统计学》，2021年。

Craig K Enders。《应用缺失数据分析》。Guilford press，2010年。

Whitney K Newey 和 Daniel McFadden。大样本估计与假设检验。《计量经济学手册》，4：2111–2245，1994年。

Art B Owen。《经验似然》。CRC出版社，2001年。

Jin Qin 和 Jerry Lawless。经验似然与一般估计方程。《统计年鉴》，第 300–325 页，1994年。

Jing Qin。 偏倚抽样、过度识别参数问题及其他，第 5 卷。施普林格，2017。

Annie Qu、Bruce G Lindsay 和 Bing Li。使用二次推断函数改进广义估计方程。《生物统计学》，87(4):823–836，2000年。

Cheng Yong Tang 和 Chenlei Leng。纵向数据分析中的经验似然与分位数回归。《生物统计学》，98(4):1001–1006，2011年。

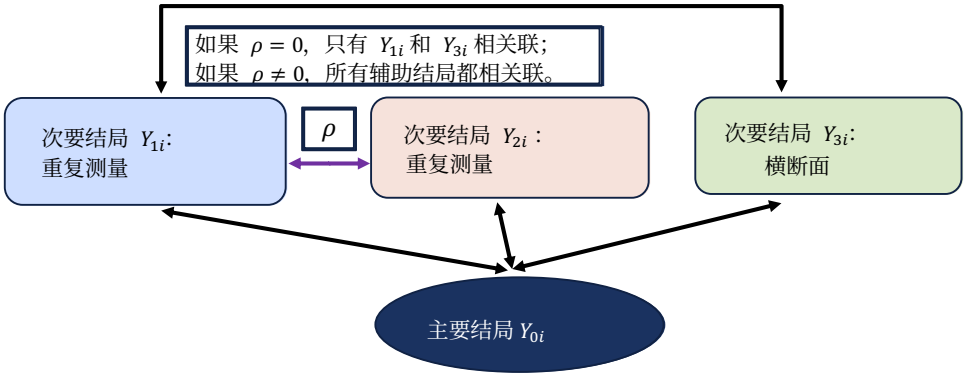

图 S.1：正文模拟研究所用数据结构示意图。第一个和第三个次要结局始终相关，而前两个次要结局之间的相关性由 ρ 控制。所有次要结局均与主要结局相关。

表 S.1：在两项纵向次要结局之间相关性为 0.4 时，针对不同样本量 (n) 和观察到三项次级数据的比例（情形 1：　$\eta_1 = \eta_2 = \eta_3 = 1$；情况2：　$\eta_1 = 0.6$,η_2 =0.7,$\eta_3 = 0.5$）组合，对八种估计方法（定义见主论文的模拟部分）的总体评估。MCSE：蒙特卡洛标准误；ASE：渐近标准误；CP：覆盖概率。所有数值均已乘以 100。

			n=300				n=600			
			Bias	MCSE	ASE	95%覆盖概率	Bias	MCSE	ASE	95%覆盖概率
情形 1	single100	β_0	2.6	24	24	95	0.8	17	17	94
		β_1	-0.5	26	25	93	0.1	18	18	94
		β_2	-2.6	15	15	95	-1.2	11	11	94
		β_3	-0.8	29	28	95	0.1	20	20	95
	single010	β_0	2.2	24	24	95	1.0	17	17	95
		β_1	-0.3	28	28	95	-0.3	20	20	95
		β_2	-2.4	15	15	96	-0.9	11	11	95
		β_3	-0.6	25	25	96	0.2	18	18	95
	single001	β_0	2.7	24	24	95	0.9	17	17	94
		β_1	-0.9	26	26	94	-0.3	19	18	94
		β_2	-2.6	16	15	95	-1.1	11	11	94
		β_3	-0.7	28	28	95	0.1	20	19	96
	平均101	β_0	2.6	24	23	95	0.8	17	16	95
		β_1	-0.7	25	24	94	-0.1	18	17	94
		β_2	-2.5	15	15	95	-1.1	11	10	94
		β_3	-0.7	28	27	95	0.1	20	19	95
	聚合101	β_0	2.6	25	23	94	0.6	17	16	94
		β_1	-0.9	26	25	93	0.2	18	17	94
		β_2	-2.3	16	15	93	-1.0	11	11	94
		β_3	-0.8	29	27	94	0.0	20	19	95
	ave111	β_0	2.4	24	23	95	0.9	17	16	95
		β_1	-0.5	26	25	94	-0.1	18	18	94
		β_2	-2.4	15	14	95	-1.0	11	10	94
		β_3	-0.7	27	26	95	0.1	19	18	95
	agg111	β_0	2.4	24	22	94	0.5	17	16	95
		β_1	-0.8	26	25	93	0.2	18	17	94
		β_2	-2.0	17	16	94	-0.8	12	11	94
		β_3	-0.7	26	25	94	0.1	18	17	94
	omn111	β_0	2.4	24	22	95	0.7	16	16	95
		β_1	-0.8	25	24	93	0.0	18	17	94
		β_2	-2.2	15	14	94	-0.9	11	10	93
		β_3	-0.6	26	25	95	0.1	18	18	95

			n=300				n=600			
			Bias	MCSE	ASE	95%覆盖概率	Bias	MCSE	ASE	95%覆盖概率
情况2	single100	β_0	2.7	25	24	95	1.0	17	17	95
		β_1	-0.4	27	26	95	-0.2	19	18	94
		β_2	-2.9	16	16	95	-1.3	12	11	94
		β_3	-0.8	29	28	95	0.2	20	20	94
	single010	β_0	2.6	24	24	95	0.9	17	17	95
		β_1	-0.3	28	28	95	-0.3	20	20	95
		β_2	-2.6	16	16	95	-1.2	12	11	94
		β_3	-1.2	27	27	96	0.3	19	19	95
	single001	β_0	2.6	25	24	95	1.0	17	17	95
		β_1	-0.7	27	26	94	-0.4	19	19	94
		β_2	-3.0	16	16	94	-1.1	11	11	94
		β_3	-0.5	29	28	95	0.2	20	19	96
	平均101	β_0	2.6	24	24	95	1.0	17	17	96
		β_1	-0.6	27	26	95	-0.3	19	18	94
		β_2	-2.9	16	15	95	-1.2	11	11	94
		β_3	-0.7	28	27	95	0.2	20	19	96
	聚合101	β_0	2.6	25	24	94	0.8	17	17	95
		β_1	-0.7	27	26	94	-0.3	19	18	94
		β_2	-3.0	16	15	93	-1.1	12	11	93
		β_3	-0.6	29	27	94	0.2	20	19	95
	ave111	β_0	2.6	24	23	95	1.0	17	17	95
		β_1	-0.5	27	26	95	-0.3	19	18	95
		β_2	-2.8	15	15	95	-1.2	11	11	94
		β_3	-0.7	28	27	95	0.2	19	19	96
	agg111	β_0	2.8	25	23	94	0.7	17	16	95
		β_1	-0.5	27	26	93	-0.3	19	18	94
		β_2	-2.8	17	16	94	-1.1	12	11	93
		β_3	-1.1	28	26	94	0.3	19	18	94
	omn111	β_0	2.6	25	23	94	0.8	17	16	95
		β_1	-0.6	27	25	94	-0.4	19	18	94
		β_2	-2.9	16	15	94	-1.1	11	11	93
		β_3	-0.7	28	26	94	0.3	19	19	95

表 S.2：在不同样本量（n）和观测三个次级数据的比例组合下（情形 1：$\eta_1 = \eta_2 = \eta_3 = 1$；情况2：$\eta_1 = 0.6$,$\eta_2$ =0。7,$\eta_3 = 0.5$），且两个纵向次级结局之间的相关性（ρ）等于 0 时，八种估计方法（定义见主论文的模拟部分）的总体评估。MCSE：蒙特卡洛标准误；ASE：渐近标准误；CP：覆盖概率。所有数值均已乘以 100。

			n=300				n=600			
			Bias	MCSE	ASE	95%覆盖概率	Bias	MCSE	ASE	95%覆盖概率
情形 1	single100	β_0	2.7	25	24	95	1.2	17	17	95
		β_1	-0.7	27	26	94	-0.4	19	18	95
		β_2	-2.8	16	15	94	-1.2	11	11	94
		β_3	-0.6	29	28	95	0.1	20	20	95
	single010	β_0	2.6	25	24	95	1.4	17	17	96
		β_1	-0.8	28	28	95	-0.7	20	20	95
		β_2	-2.7	15	16	95	-1.1	11	11	95
		β_3	-0.6	26	26	96	0.0	18	18	95
	single001	β_0	2.7	25	24	95	1.3	17	17	95
		β_1	-1.1	27	27	95	-0.7	19	19	95
		β_2	-2.6	16	16	94	-1.1	11	11	94
		β_3	-0.5	28	27	95	0.0	20	19	95
	ave101	β_0	2.7	24	23	94	1.2	17	16	95
		β_1	-0.9	26	25	95	-0.6	18	18	95
		β_2	-2.6	15	15	94	-1.1	11	11	94
		β_3	-0.5	28	27	95	0.1	20	19	95
	聚合101	β_0	2.7	25	23	94	1.1	17	16	95
		β_1	-1.0	26	24	94	-0.4	18	17	94
		β_2	-2.6	16	15	93	-1.1	11	10	94
		β_3	-0.5	29	27	94	0.0	20	19	95
	ave111	β_0	2.6	24	23	94	1.2	17	16	95
		β_1	-0.9	27	26	94	-0.6	19	18	95
		β_2	-2.6	15	15	94	-1.1	11	11	94
		β_3	-0.5	27	26	95	0.1	19	19	95
	agg111	β_0	2.7	24	22	95	1.1	16	16	95
		β_1	-1.0	26	24	93	-0.4	18	17	94
		β_2	-2.6	16	15	94	-1.0	11	10	94
		β_3	-0.6	26	25	94	-0.1	18	18	94
	omn111	β_0	2.6	24	23	94	1.2	16	16	95
		β_1	-1.0	26	25	93	-0.6	18	18	95
		β_2	-2.5	15	14	94	-1.0	11	10	94
		β_3	-0.5	27	26	95	0.0	19	18	95

			n=300				n=600			
			Bias	MCSE	ASE	95%覆盖概率	Bias	MCSE	ASE	95%覆盖概率
情况2	single100	β_0	2.9	25	24	94	1.4	17	17	95
		β_1	-0.7	28	27	94	-0.7	19	19	95
		β_2	-2.9	16	16	95	-1.3	12	11	94
		β_3	-0.7	29	28	96	0.2	20	20	95
	single010	β_0	2.8	25	24	95	1.3	17	17	95
		β_1	-0.7	28	28	95	-0.7	20	20	95
		β_2	-2.8	16	16	95	-1.2	11	11	96
		β_3	-1.0	27	27	96	0.3	19	19	95
	single001	β_0	2.8	25	24	95	1.3	17	17	95
		β_1	-1.1	28	27	95	-0.9	19	19	95
		β_2	-3.0	16	16	94	-1.1	11	11	94
		β_3	-0.3	28	28	95	0.2	20	19	95
	平均101	β_0	2.8	25	24	95	1.3	17	17	95
		β_1	-0.9	27	26	95	-0.8	19	18	94
		β_2	-2.9	16	16	94	-1.2	11	11	94
		β_3	-0.5	28	27	95	0.2	20	19	95
	聚合101	β_0	2.9	25	23	94	1.3	17	17	95
		β_1	-1.0	27	26	94	-0.8	19	18	94
		β_2	-3.1	16	15	93	-1.2	11	11	94
		β_3	-0.4	29	27	95	0.2	20	19	95
	ave111	β_0	2.8	25	24	95	1.3	17	17	95
		β_1	-0.9	27	26	95	-0.8	19	19	95
		β_2	-2.9	16	15	95	-1.2	11	11	94
		β_3	-0.6	28	27	96	0.2	19	19	95
	agg111	β_0	3.1	25	23	94	1.2	17	16	95
		β_1	-0.9	27	26	94	-0.8	19	18	94
		β_2	-3.2	17	15	93	-1.3	11	11	95
		β_3	-0.9	28	26	95	0.3	19	19	95
	omn111	β_0	2.9	25	23	94	1.2	17	16	96
		β_1	-1.0	27	26	94	-0.9	19	18	94
		β_2	-3.1	16	15	93	-1.2	11	11	94
		β_3	-0.5	28	27	95	0.3	19	19	95

表 S.3：八种估计量（定义见主论文的模拟部分）在不同样本量（n）及两个纵向次级结局之间的相关性（ρ）组合下的经验相对效率。所有次级数据均为部分观测（η_1 =0.6, $\eta_2 = 0.7$,$\eta_3 = 0.5$）。

			经验相对效率							
			single100	single010	single001	ave110	agg110	ave111	agg111	omn111
n=300	ρ=0	β_0	0.99	1.02	1.00	1.02	0.98	1.03	1.00	1.02
		β_1	1.04	1.00	1.06	1.11	1.11	1.09	1.10	1.12
		β_2	0.99	1.03	0.99	1.03	0.97	1.05	0.95	1.02
		β_3	0.97	1.10	1.01	1.01	0.97	1.05	1.05	1.04
	ρ=0.4	β_0	1.00	1.03	1.00	1.03	0.97	1.05	1.00	1.02
		β_1	1.08	1.00	1.11	1.15	1.09	1.14	1.08	1.14
		β_2	1.04	1.06	1.02	1.07	0.97	1.09	0.91	1.03
		β_3	0.97	1.13	1.00	1.00	0.96	1.05	1.06	1.05
	ρ=0.8	β_0	1.00	1.03	1.00	1.02	0.93	1.05	0.95	1.02
		β_1	1.09	1.00	1.15	1.16	0.98	1.15	0.97	1.14
		β_2	1.03	1.08	1.05	1.07	0.87	1.10	0.77	1.03
		β_3	0.97	1.16	0.99	0.99	0.94	1.05	1.06	1.05
n=600	ρ=0	β_0	1.03	1.02	1.04	1.06	1.06	1.06	1.08	1.08
		β_1	1.07	1.01	1.04	1.10	1.11	1.09	1.12	1.11
		β_2	1.01	1.04	1.05	1.06	1.06	1.07	1.04	1.09
		β_3	0.99	1.10	1.02	1.02	1.01	1.06	1.10	1.08
	ρ=0.4	β_0	1.04	1.02	1.04	1.06	1.04	1.07	1.07	1.08
		β_1	1.11	1.00	1.09	1.15	1.10	1.13	1.10	1.13
		β_2	1.04	1.05	1.08	1.09	1.04	1.10	0.97	1.09
		β_3	0.99	1.13	1.01	1.01	1.00	1.07	1.11	1.09
	ρ=0.8	β_0	1.04	1.05	1.05	1.06	1.00	1.08	1.06	1.08
		β_1	1.12	1.00	1.15	1.17	1.00	1.15	0.99	1.10
		β_2	1.07	1.07	1.11	1.11	0.97	1.12	0.86	1.05
		β_3	0.99	1.16	1.01	1.01	0.99	1.07	1.14	1.10

表 S.4：三个次级数据部分可观测时，在不同样本量（n）和相关性（ρ）组合下，设定错误的工作模型的总体评估（$\eta_1 = 0.6$、$\eta_2 = 0.7$、$\eta_3 = 0.5$）。MCSE：蒙特卡洛标准误；ASE：渐近标准误；ERE：经验相对效率；CP：覆盖概率。除 ERE 外，所有数值均乘以 100。

			n=300					n=600				
			Bias	MCSE	ASE	ERE	95%覆盖概率	Bias	MCSE	ASE	ERE	95%覆盖概率
ρ=0	ave111	β_0	2.7	25	24	1.03	95	1.3	17	17	1.04	95
		β_1	-0.8	28	27	1.06	95	-0.7	19	19	1.06	95
		β_2	-2.9	16	16	1.01	94	-1.2	11	11	1.04	94
		β_3	-0.5	28	27	1.04	95	0.2	19	19	1.06	96
	平均101	β_0	2.7	25	24	1.02	95	1.3	17	17	1.04	96
		β_1	-0.8	28	27	1.07	95	-0.7	19	19	1.08	95
		β_2	-2.9	16	16	1.01	95	-1.2	11	11	1.04	94
		β_3	-0.4	29	28	0.99	95	0.2	20	20	1.00	95
	agg111	β_0	2.9	25	23	1.02	94	1.3	17	16	1.07	95
		β_1	-0.6	28	26	1.07	95	-0.7	19	18	1.10	95
		β_2	-3.2	17	16	0.96	93	-1.2	11	11	1.05	94
		β_3	-0.5	28	26	1.05	94	0.2	19	19	1.12	95
	聚合101	β_0	2.9	25	24	0.99	94	1.3	17	17	1.05	96
		β_1	-0.7	27	26	1.08	95	-0.7	19	18	1.10	94
		β_2	-3.2	17	16	0.96	94	-1.2	11	11	1.05	94
		β_3	-0.2	29	28	0.95	94	0.3	20	20	0.99	95
	omn111	β_0	2.8	25	24	1.02	94	1.3	17	17	1.07	95
		β_1	-0.8	27	26	1.08	95	-0.8	19	19	1.09	95
		β_2	-3.2	16	16	0.98	94	-1.2	11	11	1.05	94
		β_3	-0.3	28	27	1.02	94	0.3	19	19	1.07	96
ρ=0.4	ave111	β_0	2.5	24	24	1.05	95	1.0	17	17	1.06	95
		β_1	-0.3	27	27	1.09	95	-0.3	19	19	1.09	95
		β_2	-2.8	16	16	1.04	95	-1.2	12	11	1.05	94
		β_3	-0.7	28	27	1.06	95	0.2	19	19	1.08	95
	平均101	β_0	2.6	24	24	1.03	95	1.0	17	17	1.04	95
		β_1	-0.3	27	26	1.11	95	-0.3	19	18	1.11	94
		β_2	-2.9	16	16	1.04	95	-1.2	11	11	1.06	94
		β_3	-0.6	29	28	0.99	95	0.2	20	20	1.00	96
	agg111	β_0	2.6	24	23	1.04	95	0.9	17	16	1.07	95
		β_1	-0.1	27	26	1.10	95	-0.2	19	18	1.11	94
		β_2	-3.0	16	15	0.98	94	-1.2	12	11	1.05	93
		β_3	-0.8	27	26	1.08	94	0.2	18	18	1.16	95
	聚合101	β_0	2.6	25	24	1.00	94	0.9	17	17	1.04	95
		β_1	-0.3	27	26	1.10	94	-0.3	19	18	1.11	94
		β_2	-3.1	16	15	0.98	94	-1.2	12	11	1.05	93
		β_3	-0.5	29	28	0.95	95	0.2	20	20	0.98	96
	omn111	β_0	2.6	24	23	1.04	95	0.9	17	16	1.07	96
		β_1	-0.3	27	26	1.12	95	-0.3	19	18	1.12	95
		β_2	-3.1	16	15	1.00	94	-1.2	11	11	1.06	94
		β_3	-0.6	28	26	1.05	95	0.3	19	19	1.10	95
ρ=0.8	ave111	β_0	2.4	24	24	1.06	95	0.8	17	17	1.07	95
		β_1	-0.1	27	26	1.11	95	-0.1	19	18	1.10	94
		β_2	-2.8	16	16	1.05	96	-1.0	11	11	1.07	94
		β_3	-0.6	28	27	1.08	95	0.4	19	19	1.10	95
	平均101	β_0	2.4	25	24	1.02	95	0.9	17	17	1.04	95
		β_1	-0.1	27	26	1.12	95	-0.1	19	18	1.12	94
		β_2	-2.8	16	15	1.05	95	-1.0	11	11	1.08	95
		β_3	-0.5	29	28	0.99	95	0.3	20	20	1.00	95
	agg111	β_0	2.6	25	23	1.04	95	0.6	17	16	1.08	95
		β_1	0.2	27	26	1.08	93	0.0	19	18	1.06	94
		β_2	-2.8	17	15	0.95	93	-0.9	12	11	1.02	94
		β_3	-1.1	27	26	1.12	94	0.3	18	18	1.21	95
	聚合101	β_0	2.5	25	24	0.98	95	0.7	17	17	1.02	96
		β_1	0.0	27	26	1.08	93	-0.1	19	18	1.07	94
		β_2	-2.8	16	15	0.95	93	-0.9	12	11	1.02	93
		β_3	-0.5	30	28	0.94	94	0.3	20	20	0.98	94
	omn111	β_0	2.4	25	23	1.05	95	0.7	17	16	1.08	96
		β_1	0.0	27	26	1.13	94	-0.2	19	18	1.11	94
		β_2	-2.9	16	15	1.00	94	-1.0	12	11	1.05	94
		β_3	-0.6	28	26	1.07	94	0.4	19	18	1.12	95

表 S.5：在三个次级数据因信息性缺失而部分观测的情况下，不同样本量（n）与相关性（ρ）组合下，设定错误的工作模型的总体评估。MCSE：蒙特卡洛标准误；ASE：渐近标准误；ERE：经验相对效率；CP：覆盖概率。除 ERE 外，所有数值均乘以 100。

		n=300					n=600				
		Bias	MCSE	ASE	ERE	95%覆盖概率	Bias	MCSE	ASE	ERE	95%覆盖概率
ave111	β_0	2.1	24.2	23.4	1.08	96	0.4	16.5	16.5	1.10	95
	β_1	-2.5	26.6	25.8	1.16	95	-0.6	18.5	18.2	1.16	95
	β_2	-2.1	16.3	15.3	1.10	94	-0.6	11.1	10.8	1.09	95
	β_3	1.1	27.2	26.6	1.05	95	1.0	19.8	18.8	1.09	95
平均101	β_0	2.3	24.4	23.5	1.06	96	0.6	16.6	16.6	1.08	95
	β_1	-2.3	26.2	25.2	1.20	95	-0.5	18.1	17.8	1.21	95
	β_2	-2.0	16.2	15.1	1.12	94	-0.6	11.0	10.7	1.11	94
	β_3	0.6	28.3	27.6	0.98	95	0.6	20.8	19.5	0.99	95
agg111	β_0	1.0	24.2	22.5	1.09	93	-0.2	16.0	16.0	1.16	95
	β_1	-1.5	26.1	24.6	1.21	95	-0.4	17.8	17.4	1.26	95
	β_2	-2.0	16.2	14.8	1.11	93	-0.7	11.0	10.5	1.10	94
	β_3	2.2	27.2	25.2	1.04	93	1.9	19.0	17.8	1.18	94
聚合101	β_0	2.0	24.7	23.2	1.04	94	0.6	16.5	16.5	1.10	95
	β_1	-1.6	26.0	24.6	1.22	94	-0.4	17.8	17.4	1.26	95
	β_2	-2.1	16.2	14.8	1.11	93	-0.7	11.0	10.5	1.10	93
	β_3	0.4	29.0	27.4	0.93	94	0.6	21.0	19.5	0.97	94
omi111	β_0	2.0	24.2	22.9	1.09	95	0.7	16.2	16.2	1.13	95
	β_1	-2.5	25.9	24.8	1.22	94	-1.1	17.9	17.6	1.24	95
	β_2	-2.7	16.1	14.9	1.11	94	-1.4	11.0	10.6	1.10	94
	β_3	1.3	27.4	26.0	1.03	94	1.3	19.7	18.4	1.10	94

表 S.6：在样本量为 600、次要结局与主要结局的关联极弱（约为 0.15）时，三个估计量的评估；两个纵向次要结局之间的三个相关值为 $(\rho = 0, 0.4, 0.8)$。MCSD：蒙特卡洛标准差；ASE：渐近标准误；ERE：经验相对效率（最大似然估计与目标估计值的均方误差之比）；CP：覆盖概率。除 ERE 外，所有数值均乘以 100。

		参数	Bias	MCSD	渐近标准误	经验相对效率	95%覆盖概率
ave111	$\rho = 0$	β_0	0.28	17	17	1.01	96
		β_1	-0.25	19	19	1.02	95
		β_2	-1.09	11	11	1.03	95
		β_3	0.68	19	19	1.01	95
	$\rho = 0.4$	β_0	0.27	17	17	1	95
		β_1	-0.23	19	19	1.03	95
		β_2	-1.2	12	11	1.04	95
		β_3	0.60	20	19	1.01	95
	$\rho = 0.8$	β_0	0.32	17	17	1	95
		β_1	-0.13	19	19	1.03	95
		β_2	-1.09	12	11	1.04	95
		β_3	0.44	20	19	1.01	95
omn111	$\rho = 0$	β_0	0.21	17	17	1.01	96
		β_1	-0.23	19	19	1.02	95
		β_2	-1.05	11	11	1.06	95
		β_3	0.83	19	19	1.01	95
	$\rho = 0.4$	β_0	0.19	17	17	1	95
		β_1	-0.2	19	19	1.03	95
		β_2	-1.16	11	11	1.06	95
		β_3	0.75	20	19	1.01	95
	$\rho = 0.8$	β_0	0.25	17	17	1	95
		β_1	-0.11	19	19	1.03	95
		β_2	-1.06	11	11	1.07	95
		β_3	0.58	20	19	1.01	95
agg111	$\rho = 0$	β_0	0.16	17	17	0.99	95
		β_1	-0.13	19	19	1.01	95
		β_2	-1.04	11	11	1.04	95
		β_3	0.93	20	19	0.99	94
	$\rho = 0.4$	β_0	0.12	17	17	0.98	95
		β_1	-0.08	20	19	1.01	95
		β_2	-1.14	11	11	1.05	95
		β_3	0.84	20	19	0.99	94
	$\rho = 0.8$	β_0	0.17	17	17	0.96	95
		β_1	0.01	20	19	0.99	94
		β_2	-1.05	12	11	1.02	95
		β_3	0.67	20	19	0.98	94

表 S.7：在次要模型正确设定（正确）和错误设定（错误）的情况下，针对主要终点与次要结局之间相关性较高的情形，MinBo估计量（如 ave111 和 agg111）在两种样本量 $(n = 300, 600)$ 下的评估。MCSD：蒙特卡洛标准差；ASE：渐近标准误；ERE：经验相对效率（MLE 与目标估计值的均方误差之比）；CP：基于 95% 置信区间的覆盖概率。除 ERE 外，所有数值均乘以 100。

		n=300					n=600				
		偏差	MCSD	ASE	ERE	CP	偏差	MCSD	ASE	ERE	CP
ave111 正确	β_0	2.20	22	22	1.21	95	1.13	15	15	1.20	95
	β_1	-1.31	23	22	1.46	94	-0.68	16	16	1.47	94
	β_2	-2.02	14	13	1.52	93	-0.93	10	9	1.59	94
	β_3	1.31	24	24	1.31	95	0.37	17	17	1.32	94
ave111 正确	β_0	1.46	23	21	1.17	94	0.52	15	15	1.19	95
	β_1	-0.84	25	23	1.33	94	-0.10	17	16	1.38	96
	β_2	-0.84	15	15	1.34	93	-0.26	10	10	1.33	95
	β_3	1.25	22	21	1.47	94	0.29	15	15	1.67	94
ave111 错误	β_0	2.52	23	23	1.07	95	1.18	16	16	1.07	95
	β_1	-1.75	26	24	1.23	94	-0.71	17	16	1.36	94
	β_2	-2.64	16	14	1.19	92	-1.16	11	10	1.28	94
	β_3	1.52	27	28	0.99	95	0.44	20	20	1.00	95
ave111 不正确	β_0	2.48	22	21	1.20	93	0.97	15	14	1.29	95
	β_1	-1.86	25	21	1.32	93	-0.46	15	15	1.61	94
	β_2	-1.98	14	13	1.51	92	-0.79	9	9	1.73	94
	β_3	1.58	22	21	1.58	95	0.54	15	15	1.71	94

表 S.8：在次要模型正确设定（正确）和设定错误（错误）的情况下，两种样本量 $(n = 300,\ 600)$ 下三种估计量（ave11、agg11、联合似然（JL））的评估。MCSD：蒙特卡洛标准差；ASE：渐近标准误；ERE：经验相对效率（最大似然估计与目标估计值的均方误差之比）。除 ERE 外，所有数值均已乘以 100。

方法	样本量	参数	正确			错误		
			Bias	MCSD	ERE	Bias	MCSD	ERE
ave11	300	β_0	0.36	8.69	1.41	0.45	9.50	1.18
		β_1	-0.32	8.88	1.94	-0.48	10.5	1.38
		β_2	0.19	4.28	2.06	0.17	4.71	1.71
		β_3	-0.32	8.86	1.62	-0.3	10.58	1.14
	600	β_0	0.16	5.72	1.47	0.22	6.34	1.2
		β_1	-0.09	6.06	1.99	-0.2	7.21	1.4
		β_2	0.08	2.82	2.25	0.08	3.13	1.83
		β_3	-0.35	6.16	1.74	-0.37	7.50	1.17
agg11	300	β_0	0.26	9.94	1.08	0.49	9.30	1.24
		β_1	-0.25	11.29	1.2	-0.75	10.51	1.38
		β_2	0.11	5.19	1.4	0.15	4.85	1.61
		β_3	-0.18	10.94	1.06	-0.11	10.34	1.19
	600	β_0	0.03	6.55	1.12	0.2	6.10	1.29
		β_1	0.10	7.65	1.25	-0.27	7.11	1.44
		β_2	0.01	3.44	1.52	0.01	3.27	1.67
		β_3	-0.26	7.18	1.28	-0.28	7.05	1.33
JL	300	β_0	0.24	6.59	2.46	30.36	7.99	0.11
		β_1	-0.22	4.3	8.24	37.05	9.61	0.01
		β_2	0.04	2.24	7.58	0.01	3.72	2.74
		β_3	-0.10	3.87	8.49	76.22	8.07	0.02
	600	β_0	0.07	4.52	2.36	30.44	5.48	0.05
		β_1	-0.17	3.14	7.37	37.23	6.22	0
		β_2	0.05	1.53	7.65	0.03	2.73	2.4
		β_3	-0.08	2.74	8.8	76.26	5.08	0.01