

# 基于 g-computation 联合混合效应模型控制未测混杂因素的因果推断方法模拟研究及实例验证\*

天津医科大学流行病与卫生统计学系(300070) 孙博然 芦文丽 陈永杰<sup>△</sup>

**【摘要】目的** 通过模拟实验和实例验证探讨基于 g-computation 的联合混合效应模型(joint mixed-effects model, JMM)控制纵向研究未测混杂因素进行因果推断时的效果及性能特点。**方法** 通过计算机模拟产生包含基线及两次随访时点的纵向数据,模拟条件包括样本含量、有无未测混杂因素及未测混杂效应大小,分别利用基于 g-computation 的 JMM、线性混合效应模型、固定效应模型和纵向目标极大似然估计方法估计因果效应,通过平均绝对偏差(mean absolute deviation, MAD)、标准误、均方根误差(root mean square error, RMSE)、95% 置信区间覆盖率(95% confidence interval coverage, 95% CI coverage)评价比较各方法因果推断的效果。利用绝经期女性队列体检数据,应用四类模型分别估计绝经期女性血清卵泡刺激素(follicle-stimulating hormone, FSH)水平与腰椎骨密度间因果关系,对各模型在真实纵向数据中的因果推断效果进行验证。**结果** JMM 控制未测混杂因素的因果推断准确性最佳,但稳定性略差。当研究中存在较强未测混杂效应时,仅 JMM 可准确估计因果效应,且其在大样本量时估计的精确性和真实性较好。**结论** 基于 g-computation 的 JMM 可有效控制纵向研究中未测混杂因素进行近似无偏因果推断。

**【关键词】** 纵向研究 未测混杂因素 g-computation 联合混合效应模型

**【中图分类号】** R195.1 **【文献标识码】** A **DOI** 10.11783/j.issn.1002-3674.2024.05.012

## Simulation Study and Case Validation on Causal Inference of g-computation-based Joint Mixed-effects Model for Controlling Unmeasured Confounders

Sun Boran, Lu Wenli, Chen Yongjie (Department of Epidemiology and Health Statistics, School of Public Health, Tianjin Medical University, Tianjin 300070)

**【Abstract】Objective** A simulation study was conducted to explore the effect and performance of g-computation-based joint mixed-effects model(JMM) on causal inference for controlling unmeasured confounders in longitudinal studies. **Methods** Longitudinal data including baseline and two follow-up visits were generated by computer simulations. The simulation scenarios included different sample sizes, the presence or absence of unmeasured confounders, and effects of unmeasured confounders. Causal effects were estimated using g-computation-based JMM, linear mixed-effects model, fixed effects model, and longitudinal target maximum likelihood estimation, respectively. Indicators including mean absolute deviation(MAD), standard error, root mean square error(RMSE), and 95% confidence interval coverage(95% CI coverage) were used to evaluate and compare the causal inference performance. Based on the physical examination cohort data of the menopausal women, four models were used to estimate the causal association between serum follicle-stimulating hormone(FSH) levels and lumbar bone density in menopausal women respectively, verifying the causal inference performance of models in the real longitudinal data. **Results** JMM had a better accuracy of causal inference with controlling unmeasured confounders. But its estimation stability was slightly worse. When strong unmeasured confounders existed, only JMM can accurately estimate the causal effect, and its precision and authenticity were better in scenarios with large sample sizes. **Conclusion** JMM can effectively control the unmeasured confounders and perform approximately unbiased causal estimation in longitudinal studies.

**【Key words】** Longitudinal studies; Unmeasured confounders; g-computation; Joint mixed-effects model

在生物医学研究中,因果证据一直是循证医学、药物疗效评价、临床及人群干预决策制定的依据<sup>[1]</sup>。在新时代的医疗健康大数据背景下,观察性研究中的纵向研究逐渐受到学者们的青睐与重视。与随机对照实验不同,纵向研究缺少随机化分组,其因果推断易受到依时暴露因素、依时混杂因素以及未测混杂因素<sup>[2]</sup>的影响。其中,依时暴露与依时混杂因素均随时间变化,其产生的生物滞后效应及过度调整偏倚均可造成有偏的因果估计;未测混杂因素为实际研究中被忽略或未测量的混杂因素,在观察性研究中不可避免,直接对结

局和依时混杂因素产生影响,导致因果推断偏倚的出现。因此,如何在纵向研究中有效控制各类偏倚进而无偏估计因果效应,已成为真实世界研究中亟待解决的因果推断问题。

目前,纵向数据因果推断方法主要集中于半参数模型,包括 g-computation、g-estimation 的结构嵌套模型、边际结构模型的 G 方法、工具变量法以及纵向目标极大似然估计(longitudinal targeted maximum likelihood estimation, LTMLE)等<sup>[3-8]</sup>。G 方法和 LTMLE 可针对性地控制依时混杂因素,但均要求研究中不存在未测混杂因素。工具变量法主要局限于实际研究中较难找到理想的工具变量及易出现无效工具变量。近几年来,共享参数联合模型开始被用于控制纵向研究

\* 基金项目:国家自然科学基金青年项目(81903416)

<sup>△</sup>通信作者:陈永杰,E-mail: chen Yongjie@tmu.edu.cn

中的未测混杂因素<sup>[9]</sup>,但其不能对因果效应进行有效解释,故无法做出真正的因果推断。Shardell<sup>[10]</sup>等人在共享参数联合模型基础上提出基于 g-computation 的联合混合效应模型 (joint mixed-effects model, JMM),该模型既可有效控制未测混杂偏倚,也使因果估计具有实际意义。鉴于该模型尚未在国内医学研究中涉及,本研究旨在介绍此模型的构建过程及原理,并通过模拟实验和实例验证将 JMM 与基于 g-computation 的线性混合效应模型 (linear mixed-effects model, LMM)、固定效应模型 (fixed effects model, FEM) 及 LTMLE 三种方法进行比较,评价其在控制纵向研究未测混杂因素进行因果推断时的效果,为建立相关因果推断体系提供参考。

### 对象与方法

#### 1. 基于 g-computation 的 JMM 基本原理

基于 g-computation 的 JMM 将未测混杂因素作为随机效应进行控制,对纵向数据中的结局变量与暴露因素分别建立混合效应模型,通过共享随机效应将结局变量模型和暴露因素模型进行联合,利用自适应高斯求积法进行参数估计,最后通过 g-computation 对联合混合效应模型的参数构建因果关系函数,进而得到近似无偏的因果效应估计值<sup>[10]</sup>。

#### 2. 模型构建及参数估计

假定纵向研究中共纳入  $n$  名参与者,对每名参与者  $i$  分别进行  $J_i$  次随访; $Y_{ij}$  表示参与者  $i$  在第  $j$  次随访时的结局变量, $A_{i,j-1} = \{A_{i0}, A_{i1}, \dots, A_{i,j-1}\}$  表示参与者  $i$  自基线起至第  $j-1$  次随访间的暴露情况, $X_i$  为基线协变量。

将随机效应向量  $b_i$  引入结局变量模型,代表研究中的未测混杂因素,结局变量模型构建见式(1)。

$$Y_{ij} = \beta_0 + X_i' \beta_1 + \beta_2 Y_{i,j-1} + \bar{A}_{i,j-1}' \beta_3 + (X_i \bar{A}_{i,j-1})' \beta_4 + \beta_5 j + b_i + \varepsilon_{ij} \quad (1)$$

其中, $j=1, \dots, J_i, \beta_2$  为  $Y_{ij}$  与  $Y_{i,j-1}$  间的序列相关系数,模型通过引入序列相关系数解决依时暴露因素的生物滞后效应并控制依时混杂因素产生的交叉影响, $\varepsilon_{ij}$  为误差项。

将随机效应向量  $b_i$  引入暴露因素模型,暴露因素模型构建见式(2)。

$$\text{logit}(p_{i,j-1}^A) = \alpha_0 + X_i' \alpha_1 + \alpha_2 Y_{i,j-1} + \bar{A}_{i,j-2}' \alpha_3 + (X_i Y_{i,j-1})' \alpha_4 + \alpha_5 j + \alpha_6 b_i \quad (2)$$

其中, $j=1, \dots, J_i, p_{i,j-1}^A$  为参与者  $i$  在第  $j-1$  时刻的暴露情况。

通过共享同一随机效应  $b_i$ ,将上述结局变量模型与暴露因素模型进行联合,构建联合混合效应模型。利用自适应高斯求积法进行联合极大似然估计,通过

双拟 Newton 算法对参数估计进行优化,最终得到参数  $\beta$  的估计值。

由于依时混杂因素的存在,此时参数  $\beta$  并无真正的因果意义,除非研究中不存在未测混杂因素。因此对于含有未测混杂因素的纵向研究,需利用 g-computation 算法估计因果效应。g-computation 算法的基本原理是将暴露因素作为固定效应,在考虑到所有随访时点协变量和随机效应  $b_i$  的情况下对结局变量的条件期望值构建因果关系函数进行积分运算,不同暴露水平间的差异即为因果效应,见式(3)。

$$g_j(\bar{a}_{j-1} | X_i, b_i) = \int E[Y_{ij} | X_i, \bar{Y}_{i,j-1}, \bar{A}_{i,j-1} = \bar{a}_{j-1}, b_i] \prod_{k=0}^{j-1} f(Y_{i,k} | X_i, \bar{Y}_{i,k-1}, \bar{A}_{i,k-1} = \bar{a}_{k-1}, b_i) dY_{i,k} \quad (3)$$

通过计算  $g_j(\bar{a}_{j-1} | X_i, b_i) - g_j(\bar{a}'_{j-1} | X_i, b_i)$  即可得到因果效应。其中,  $\bar{a}_{j-1}$  为参与者的实际暴露情况,  $\bar{a}'_{j-1}$  为参与者的反事实暴露情况。

#### 3. 模拟实验

##### (1) 模拟条件

本研究主要考虑样本含量、有无未测混杂因素及未测混杂效应大小三个模拟条件,具体设置如下:样本含量:考虑到实际纵向研究的样本含量,共设置 100、200、500、1000 四种情况;有无未测混杂因素:设置有、无未测混杂因素两种情况, $\alpha_6 = 1$  代表研究中有未测混杂因素, $\alpha_6 = 0$  代表研究中无未测混杂因素;未测混杂效应大小:通过调整  $b_i$  的方差  $\text{Var}(b_i)$  来控制未测混杂效应的大小,分别设定  $\text{Var}(b_i) = 0.5$  和  $\text{Var}(b_i) = 1.0$  两种情况。各模拟条件组合均重复模拟 1000 次。

##### (2) 模拟数据

本模拟实验共设定三个时间点,分别代表基线及两次随访,即  $j=0, 1, 2$ ,并基于如下假定生成模拟数据:①各随访时点结局变量  $Y_{ij}$  为连续型变量,服从正态分布  $N(\mu, \sigma^2), \sigma^2 = 1$ ;②各时点暴露因素  $A_{ij}$  为二分类变量,服从二项分布  $B(n, p)$ ;③基线协变量  $X_i$  服从标准正态分布  $N(0, 1)$ ;④随机效应项  $b_i$  服从正态分布;⑤误差项  $\varepsilon_{ij}$  服从标准正态分布  $N(0, 1)$ 。最后通过对模型系数进行初始赋值产生模拟数据,其中  $\alpha_0 = 0, \alpha_1 = 0.5, \alpha_2 = 0.75, \alpha_3 = 1.0, \alpha_4 = 0.25, \alpha_5 = 0; \beta_0 = 1, \beta_1 = -1, \beta_2 = 0.5, \beta_3 = 2, \beta_4 = 3, \beta_5 = 0, \beta_6 = 0, \beta_7 = -1$ 。

##### (3) 因果效应真实值

当  $j=1$  时,为基线暴露水平与第一次随访结局间的因果效应,即  $a_0$  对  $y_1$  的因果作用,此时因果效应真实值为  $g_1(a_0 = 1) - g_1(a_0 = 0) = \beta_3 + \beta_4 E[X_i] = 2$ 。当  $j=2$  时,包含两类情况:①第一次随访暴露水平相同时,为基线暴露水平与第二次随访结局间的因果效应,即  $a_0$  对  $y_2$  的因果作用,此时因果效应真实值为  $g_2(a_0 = 1, a_1) - g_2(a_0 = 0, a_1) = \beta_3 + \beta_2 \beta_3 = 3$ ;②基线暴露水

平相同时,为第一次随访时暴露水平与第二次随访结局间的因果效应,即  $a_1$  对  $y_2$  的因果作用,此时因果效应真实值为  $g_2(a_0, a_1 = 1) - g_1(a_0, a_1 = 0) = \beta_{31} = 3$ 。

(4) 评价指标

模拟实验评价指标包括平均绝对偏差(mean absolute deviation, MAD)、均方根误差(root mean square error, RMSE)、标准误(s.e)和 95%置信区间(confidence interval, CI)覆盖率。MAD 是因果效应估计值与真实值间差值绝对值的平均数,用于反映模型因果效应估计的准确性,该值越接近 0 表示估计的准确性越好;RMSE 综合反映估计的方差与偏倚,该值越接近于 0,表示估计的精确性越好;s.e 用于反映估计的稳定性,该值越小表示估计越稳定;95% CI 覆盖率为多次重复实验中,因果效应估计值的置信区间包含因果效应真实值的比例,越接近 95% 表示因果效应估计的真实性越好。

4. 实例验证

既往研究已表明血清卵泡刺激素(follicle-stimulating hormone, FSH)水平是导致绝经期女性骨量流失的原因之一,FSH 可综合调控骨量代谢,高水平 FSH 会诱导绝经期妇女骨质疏松症的发生<sup>[11-12]</sup>。本研究应用一项绝经期女性纵向体检人群队列数据,调整包

括年龄、BMI、雌激素、生育子女数、既往史、家族史、体力活动、饮酒、喝茶、喝咖啡等协变量后,通过分析绝经期女性 FSH 水平与骨密度间的因果关系对基于 g-computation 的 JMM、LMM、FEM 和 LTMLE 四种方法的因果推断效果进行实例验证。

5. 统计学软件

本研究应用 R 4.1.1 软件产生不同条件下的模拟数据。通过 SAS 9.4 软件的 PROC NL MIXED 过程构建 JMM、LMM 及 FEM 并进行因果效应估计,通过 R 4.1.1 中 ltmle 软件包进行 LTMLE 的因果效应估计。

结 果

1. 模拟实验

(1) 因果效应估计的 MAD

无未测混杂因素时,与 JMM 相比,LMM 因果效应估计的 MAD 值较小,因果效应估计准确性较好。随着样本量的增加,JMM 估计的 MAD 下降速度较快。当存在未测混杂因素时,特别是在较大样本量条件下,JMM 因果效应估计的 MAD 值较小,且当未测混杂效应增加时,JMM 仍表现出较好的因果估计准确性,如图 1 所示。

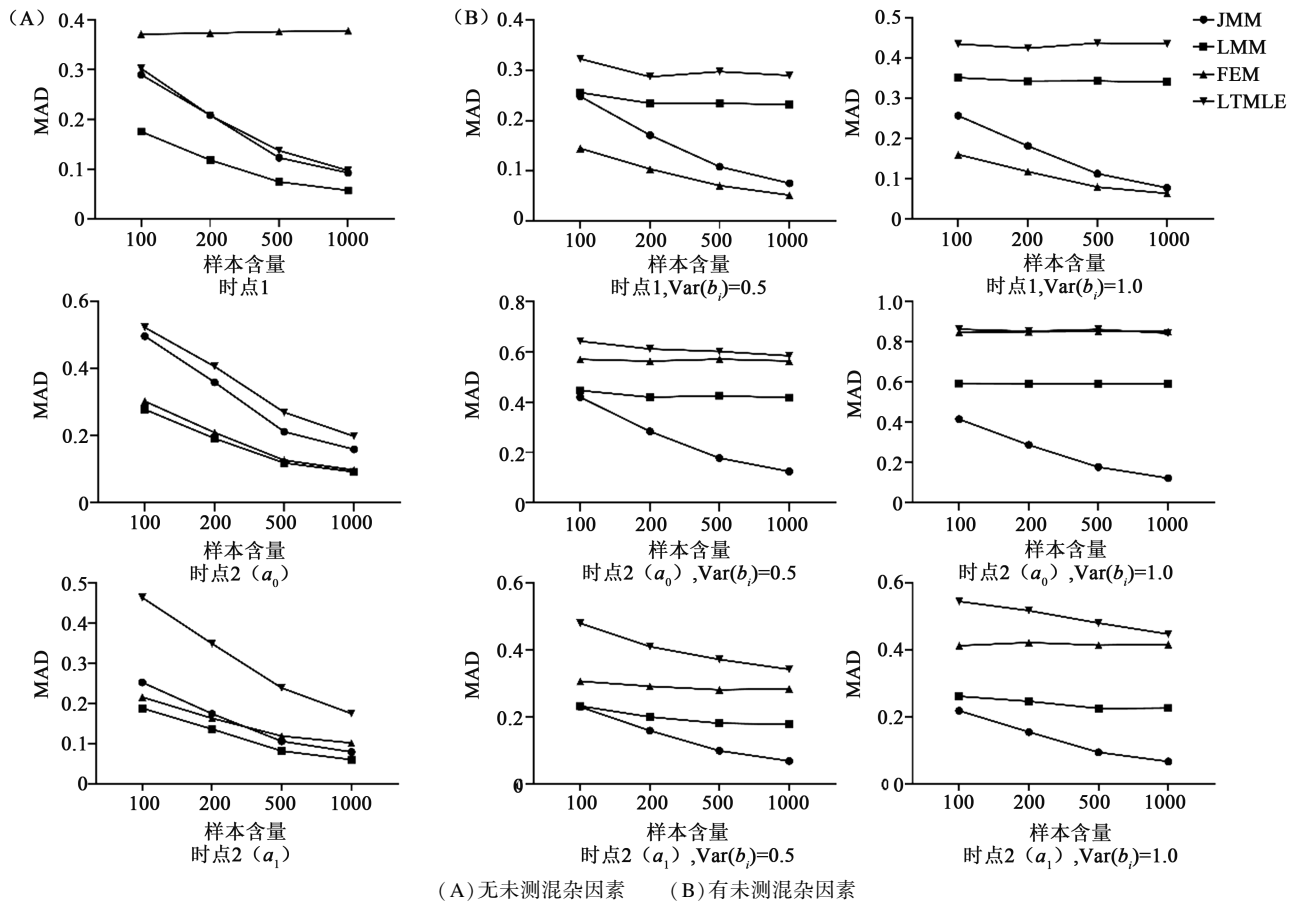


图 1 不同样本含量情形下不同模型因果效应估计的 MAD

(2) 因果效应估计的 RMSE 及 s.e

对于 RMSE, 无未测混杂因素时, LMM 估计的 RMSE 最小, JMM 估计的 RMSE 略高于 LMM 和

FEM。当存在未测混杂因素且样本量较小时, JMM 估计的 RMSE 略高于 LMM 和 FEM, 随着样本量增加, 各模型估计的 RMSE 均呈下降趋势, 其中 JMM 下降

速度较快。如图 2 所示。对于 s.e,随着样本量的增加,各模型估计的 s.e 均呈下降趋势,无论是否存在未

测混杂因素,JMM 估计的 s.e 均高于 FEM 和 LMM。如图 3 所示。

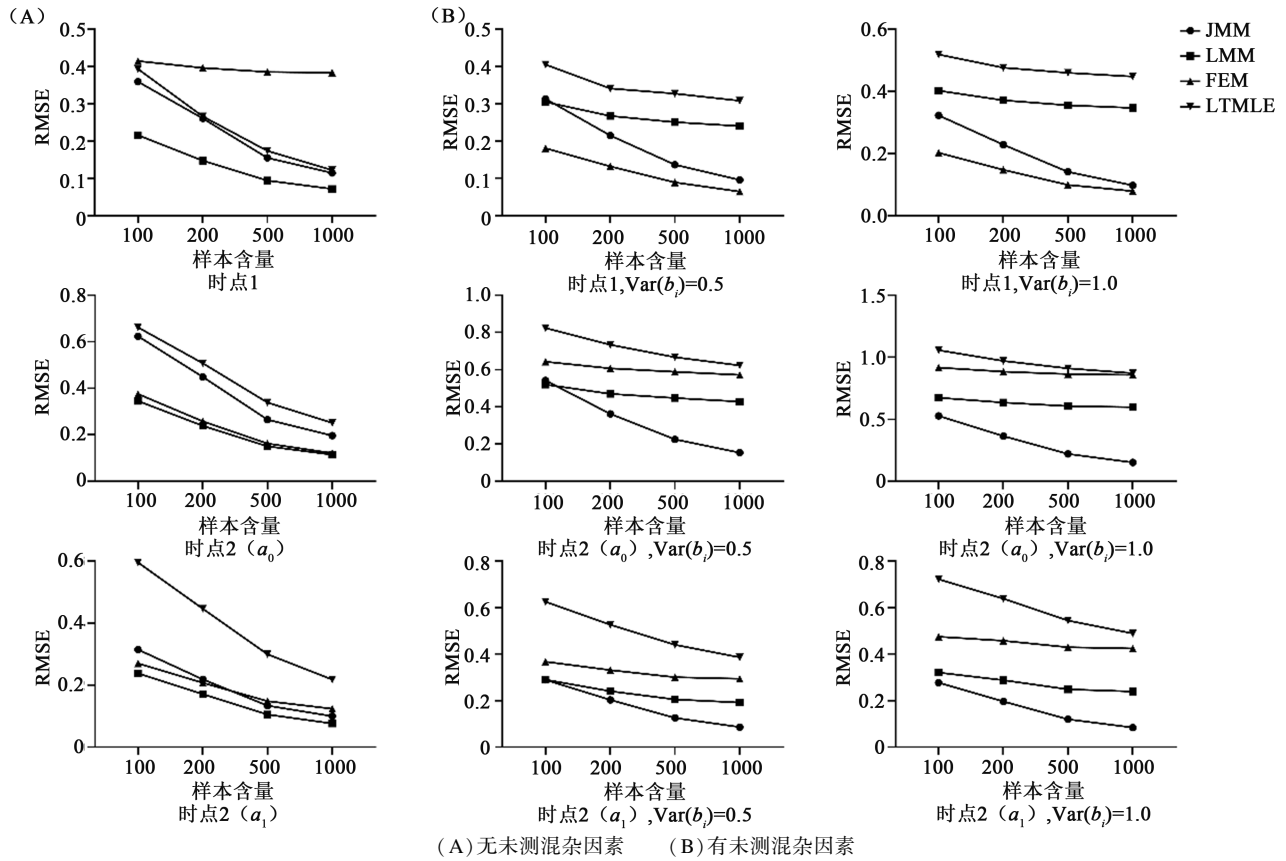


图 2 不同样本含量情形下不同模型因果效应估计的 RMSE

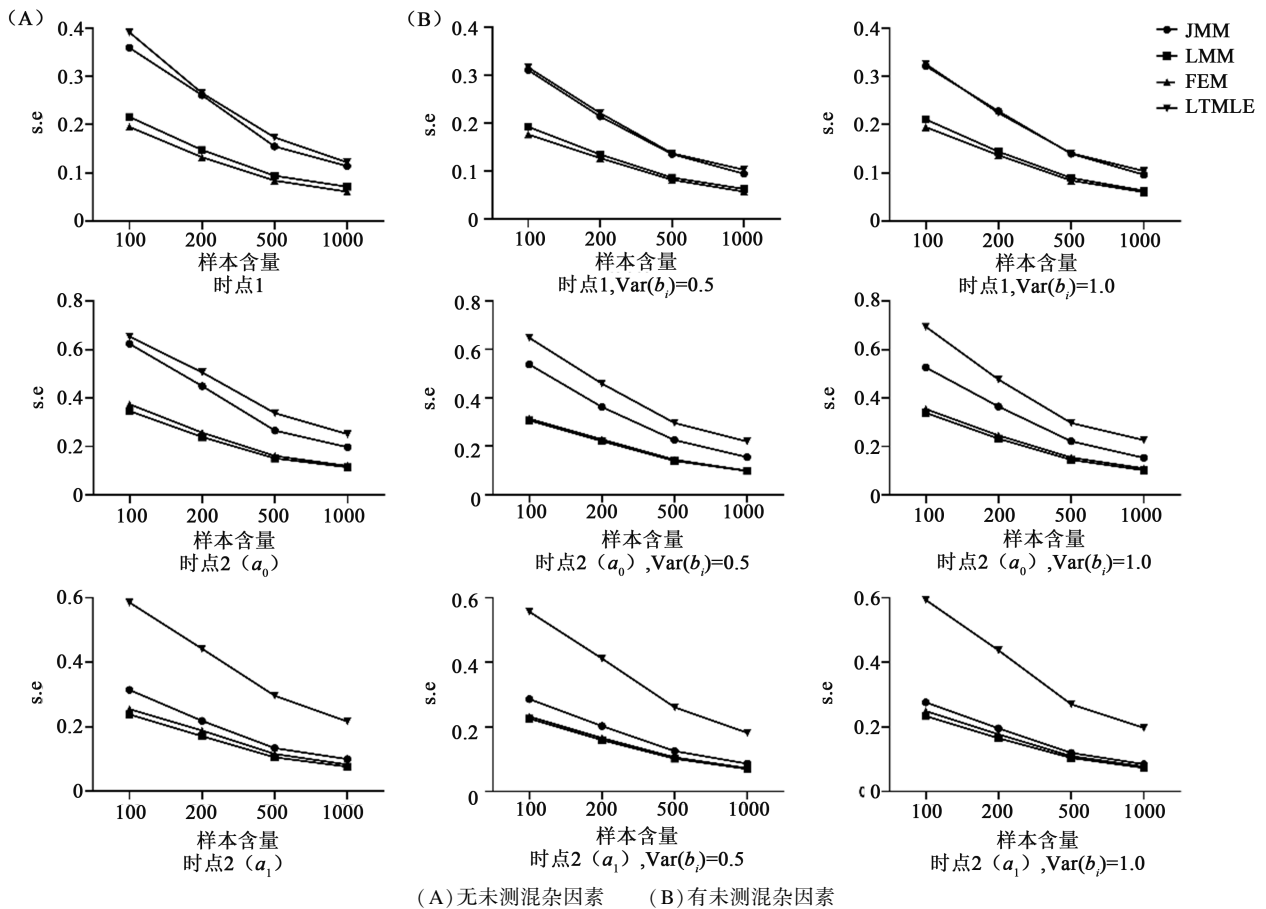


图 3 不同样本含量情形下不同模型因果效应估计的 s.e

(3) 因果效应估计的 95% CI 覆盖率

当未测混杂因素时, FEM 估计的 95% CI 覆盖率波动较大, JMM、LMM 和 LTMLE 估计的 95% CI 覆盖率均保持相对稳定, 与 JMM 和 LTMLE 相比, LMM 估计的 95% CI 覆盖率更接近于 95%, 且波动最小。

当存在未测混杂因素时, JMM 因果效应估计的 95% CI 覆盖率均稳定保持在 95% 附近, 随着样本量的增加, 其估计的 95% CI 覆盖率逐渐接近于 95%, 而其余三种模型估计的 95% CI 覆盖率波动较大, 且随样本量增加呈现下降趋势。如图 4 所示。

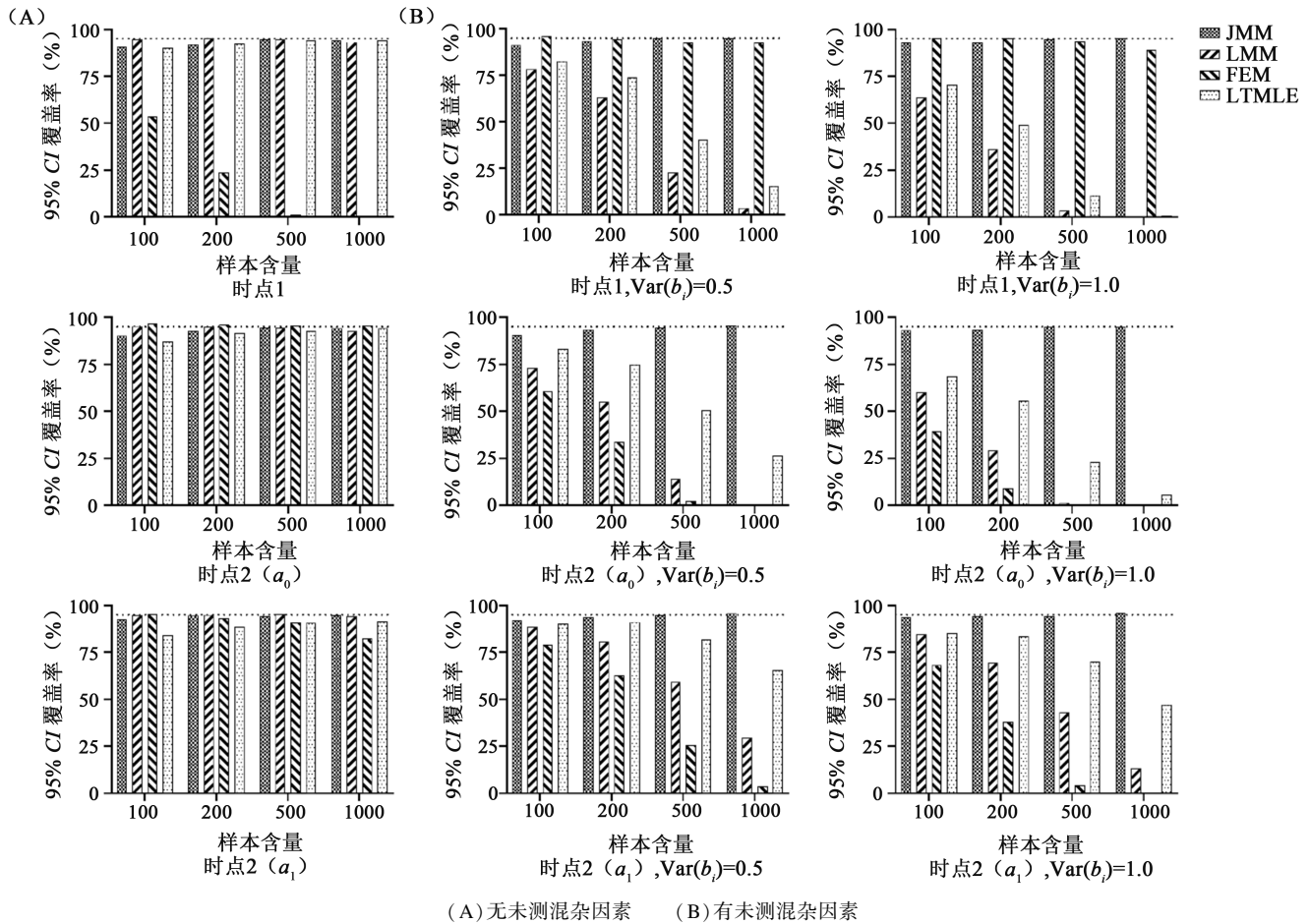


图 4 不同样本含量情形下不同模型因果效应估计的 95% CI 覆盖率

2. 实例验证

已有研究表明 FSH 与骨密度确实存在因果关联, 本实例数据控制协变量后, JMM 和 LMM 分析结果表明血清 FSH 水平与腰椎骨密度间的因果关系具有统计学意义, 较高的血清 FSH 水平可降低腰椎骨密度 [JMM:  $\beta(95\% CI): -0.017(-0.032, -0.002)$ ,  $P = 0.027$ ; LMM:  $\beta(95\% CI): -0.015(-0.027, -0.004)$ ,  $P$

$= 0.011$ ], 与已有研究结果一致; 同时, 鉴于实际研究不可避免存在未测混杂因素, JMM 和 LMM 控制未测混杂因素后结果准确性较高, 而 FEM 和 LTMLE 分析结果未提示血清 FSH 水平与腰椎骨密度间存在因果关系, 与真实情况不符, 准确性较差, 且 JMM 因果效应估计标准误差较 LMM 大, 稳定性略差, 这些结果均与模拟实验一致。详见表 1。

表 1 不同模型估计的血清 FSH 水平与腰椎骨密度间因果关系

模型	估计值	s.e	95% CI	P
JMM	-0.017	0.007	-0.032, -0.002	0.027
LMM	-0.015	0.006	-0.027, -0.004	0.011
FEM	0.037	0.019	0.000, 0.074	0.051
LTMLE	-0.008	0.007	-0.023, 0.007	0.275

\* 依据 FSH 中位数分为低、高水平两组

讨 论

本研究围绕如何有效控制纵向研究未测混杂因素

进而得到近似无偏因果推断这一难题开展模拟研究, 探讨了基于 g-computation 的 JMM 在解决该难题时的效果及性能特点, 对于完善纵向研究因果推断方法体

系具有重要参考意义。

本研究首先通过模拟实验证明了基于 *g-computation* 的 *JMM* 可有效控制纵向研究中的未测混杂因素进行因果推断,且因果效应估计具有较好的准确性,但稳定性略差。相比其他模型,*JMM* 应用混合效应模型中的随机效应项控制未测混杂因素,将其作为共享参数将结局变量模型和暴露因素模型联合,有效控制未测混杂因素的同时也能估计依时暴露因素的生物滞后效应,使得因果效应估计具有较好的准确性<sup>[10]</sup>。精确性和真实性是衡量因果效应估计效果的综合指标。由于 *JMM* 不依赖于研究中无未测混杂因素的假设,当研究中存在较强的未测混杂因素时,*JMM* 因果推断的优势逐渐体现出来,尤其在大样本时,其估计准确性、真实性和精确性均达到最佳。但 *JMM* 为参数模型,对模型参数设定较敏感,当研究中无未测混杂因素时,随机效应分布可能会被错误指定,*JMM* 估计精确性和真实性受参数设定的影响较大。因此,*JMM* 对因果效应的无偏估计依赖于对结局变量模型、暴露因素模型及随机效应的正确指定,当模型参数较多时,易导致估计结果不太稳定。此外,实例验证结果进一步说明在存在未测混杂因素的纵向研究中,*JMM* 具有较好的因果推断准确性及略差的稳定性的特点,对模拟实验结果再次进行了验证。

基于 *g-computation* 的 *JMM* 不依赖研究中无未测混杂因素的假设,也不要求研究中存在非混杂变量,可使用现成的统计软件如 SAS 软件的 PROC NL MIXED 过程步对模型参数进行有效估计,该方法的实用性得到了增强。但该方法也有其局限性。首先,*JMM* 依赖于对各项参数的正确指定,研究证明当错误指定研究中参数分布时,*JMM* 的因果效应估计是有偏的<sup>[10,13]</sup>;其次,与其他基于 *g-computation* 的模型类似,当研究中含有大量的依时混杂因素时,*JMM* 运算量较大,不易收敛,易出现 *g-null* 悖论<sup>[14]</sup>;此外,目前尚无证据表明 *JMM* 可有效控制未测依时混杂因素;最后,*JMM* 只适用于定量结局变量,尚不能处理分类、等级及生存数据结局变量。鉴于 *JMM* 的强参数假设,研究人员应依据自身专业背景充分收集必要的协变量,同时可借助集成机器学习等新兴方法帮助建立正确模型,也可使用依赖于不同假设的方法进行敏感性分析<sup>[15]</sup>。例如,可在一项研究中同时应用 LTMLE 和 *JMM* 方法,利用 LTMLE 对 *JMM* 的函数或变量形式进行适当转换,以满足 *JMM* 的参数假设。

本研究仍存在一些局限性。首先,本研究为简化模型,未纳入随机斜率项,仅通过纳入随机截距项以代

表研究中未测混杂效应的大小;此外,本研究未考虑对缺失数据或删除数据建立相应模型,这可能会使得本研究对于 *JMM* 的因果推断效果评价不够全面。未来应多维度评价各类模型在控制纵向数据中未测混杂因素进行因果推断方面的效果,使得纵向数据因果推断方法的理论体系更加完善。

## 参 考 文 献

- [ 1 ] Rothman KJ, Greenland S. Causation and causal inference in epidemiology[J]. Am J Public Health, 2005, 95(Suppl 1):S144-S150.
- [ 2 ] Vandembroucke JP. When are observational studies as credible as randomised trials[J]. Lancet, 2004, 363(9422):1728-1731.
- [ 3 ] Petersen M, Schwab J, Gruber S, et al. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models[J]. J Causal Inference, 2014, 2(2):147-185.
- [ 4 ] 吴诗蓝,周价,李逊,等. Parametric g-formula 方法在因果分析中的应用[J]. 中华流行病学杂志,2019,40(10):1310-1313.
- [ 5 ] Naimi AI, Richardson DB, Cole SR. Causal inference in occupational epidemiology: accounting for the healthy worker effect by using structural nested models[J]. Am J Epidemiol, 2013, 178(12):1681-1686.
- [ 6 ] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology[J]. Epidemiology, 2000, 11(5):550-560.
- [ 7 ] Sainani KL. Instrumental Variables: Uses and Limitations [J]. PM&R, 2018, 10(3):303-308.
- [ 8 ] Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies[J]. Am J Epidemiol, 2017, 185(1):65-73.
- [ 9 ] Sitlani CM, Heagerty PJ, Blood EA, et al. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance [J]. Stat Med, 2012, 31(16):1738-1760.
- [ 10 ] Shardell M, Ferrucci L. Joint mixed-effects models for causal inference with longitudinal data[J]. Stat Med, 2018, 37(5):829-846.
- [ 11 ] Sun L, Peng Y, Sharrow AC, et al. FSH directly regulates bone mass[J]. Cell, 2006,125(2):247-260.
- [ 12 ] Wang J, Zhang W, Yu C, et al. Follicle-Stimulating Hormone Increases the Risk of Postmenopausal Osteoporosis by Stimulating Osteoclast Differentiation[J]. PLoS One, 2015, 10(8):e0134986.
- [ 13 ] Litière S, Alonso A, Molenberghs G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models[J]. Stat Med, 2008, 27(16):3125-3144.
- [ 14 ] Pantazis N, Touloumi G. Robustness of a parametric model for informatively censored bivariate longitudinal data under misspecification of its distributional assumptions: A simulation study[J]. Stat Med, 2007, 26(30):5473-5485.
- [ 15 ] Gruber S, Logan RW, Jarrín I, et al. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets[J]. Stat Med, 2015, 34(1):106-117.

(责任编辑:邓妍)