arXiv:2303.02201v1 [stat.ME] 3 Mar 2023

基于纵向数据的多变量广义线性混合效应模型因果推断

Yizhen Xu、Jisoo Kim、Laura K. Hummers、Ami A. Shah、Scott Zeger

2023年3月7日

摘要

在给定个体特征和纵向历史的条件下，动态预测不同治疗方案下的因果效应，是精准医学中的一项重要问题。在实践中，这颇具挑战，因为观察性研究中的结局和治疗分配机制未知，个体的治疗效果属于反事实，而且选择偏倚往往难以避免。

我们通过调整贝叶斯 g-计算算法（J. Robins，1986年；Zhou、Elliott 和 Little，2019年），将多变量广义线性混合效应模型纳入其中，提出一个贝叶斯框架，用于识别动态治疗方案的亚组反事实获益。在所假设的结局、时变混杂因素和治疗分配的联合分布中，未测量的时间不变因素被表示为个体特异性随机效应。现有方法大多假设不存在未测量混杂，并着重于平衡不同治疗组中观测到的混杂因素分布；而我们的方法允许存在时间不变的未测量混杂。我们基于治疗分配异质性提出一项序贯可忽略性假设，其原理类似于平衡由未测量的时间不变因素（超出可观测因素的范围）导致的对各治疗的潜在倾向。我们通过模拟研究评估所提方法的表现对各种模型假设的敏感性。我们将该方法应用于观察性临床数据，研究接受该药物治疗的不同硬皮病患者亚组持续使用吗替麦考酚酯的疗效。

关键词： 纵向　　因果推断、潜变量建模、混合效应模型、g计算

1 引言

精准医学（Kosorok 和 Laber，2019年；Rosen、Zeger 等，2019年）是一种临床决策过程，它利用患者的病史、当前及既往健康状况，以及来自大规模总体的观测数据，在疾病进展过程中为患者制定个体化治疗方案和护理建议。例如，Wang 等（2022年）预测了个体未来的生物标志物轨迹和重大临床事件，以改善 COVID-19 的护理；Coley 等（2017年）则利用纵向生物标志物测量，改进有关是否切除或放射治疗患者前列腺癌的临床决策。在纵向情境下研究个体治疗效应的异质性，是精准医学关注的诸多问题之一。这涉及将患者当前的信息映射到潜在行动下的生物标志物轨迹，例如娱乐主管

治疗的选择和时机。我们尤其关注利用观察性数据回答因果问题："对于具有 h 天病史的患者，如果实施了特定的动态治疗方案，那么在 τ 天后会发生什么？"，其中动态治疗方案是指可根据观测到的患者病史而改变的处理。随后，我们可以针对具有相似特征或病史的患者亚组，评估不同治疗方案下的平均治疗效应（ATE），从而确定哪种处理方案最适合患者。

我们的示例应用旨在利用约翰斯·霍普金斯硬皮病中心研究登记库中的临床观测数据，研究免疫抑制药物吗替麦考酚酯（MMF）（Omair、Alahmadi 和 Johnson，2015年；Zamora 等，2008年）对系统性硬化症（硬皮病）患者的疗效。硬皮病是一种罕见的多系统自身免疫性疾病，其特征是纤维化过度、血管病变以及免疫系统紊乱。该疾病的纤维化表现包括皮肤增厚，这可能限制关节活动范围并导致残疾；还包括间质性肺病（ILD），后者是发病率和死亡率的主要原因。根据有限的观察性数据和临床试验数据，MMF 是目前针对活动性弥漫型皮肤系统性硬化症和/或 ILD 患者的标准治疗。然而，必须指出，硬皮病具有很强的异质性；尽管许多患者有影像学检查显示 ILD 的证据，但肺病进展的患者比例相对较低。鉴于这种临床差异以及 MMF 的风险（例如感染风险增加和胃肠道不良反应），目前仍不清楚如何在恰当的时间将这种治疗用于恰当的患者。由于缺乏证明疗效显著的研究，且存在相关的不良事件特征，目前尚无针对该疾病所致皮肤增厚的普遍认可的治疗方法。

弥漫性皮肤型系统性硬化症（dcSSc）是硬皮病的一种亚型，其特征是皮肤增厚范围更广，且改良Rodnan皮肤评分更高。改良Rodnan皮肤评分是在17个身体部位评估皮肤增厚程度的连续指标。在硬皮病肺部研究II（Tashkin 等，2016）中，患者被随机分配接受吗替麦考酚酯（MMF）或环磷酰胺治疗硬皮病相关ILD；至24个月时，MMF使用力肺活量（FVC，即肺功能指标）和改良Rodnan皮肤评分（针对弥漫性皮肤型系统性硬化症患者）均获得了相当程度的改善。本研究比较了含MMF与不含MMF治疗方案对患者皮肤和肺部指标的疗效，这些患者均已证明能够耐受MMF，无论其患有弥漫型还是局限型／无皮肤硬化型硬皮病。在这项观察性研究中，存在多项实际挑战：治疗分配并非基于已测量因素进行随机化；生物标志物的测量时间不规则；缺失模式可能反映生物标志物的取值；受试者之间还存在观测数据无法解释的自然异质性。为应对这些问题，我们在潜在结果框架下采用贝叶斯方法（Rubin，1974）。该框架将因果效应定义为同一组受试者在不同治疗方案下潜在结果的比较。这种方法能够处理结构性缺失，运用具有灵活性的贝叶斯模型应对复杂数据，并自然地量化不确定性；这些优点对于精准医学中的决策都至关重要。

在评估治疗效果时，无论是在本文还是在精准医学中比较治疗方案的许多其他情景下，主要因素都是个体异质性，或治疗分配和生物标志物动态中的未测量因素。个体治疗决策很容易受到可能混杂疾病进展的未测量变量影响。通常，决定是否以及何时为患者治疗的临床医生，可以获取关于患者潜在结局的私人信号，例如虚弱程度、接受治疗的意愿以及潜在的不良反应风险等。我们并非总能收集到一组观测变量，作为所有可用信息的代理，来

自所有信号中获取的信息。未测量变量不仅会影响随时间变化的决策，也会影响生物标志物进展。Heckman 和 Willis（1977年）认为，当存在未观测到的永久性成分时，观测特征相似的受试者可能具有不同的反应分布，即个体的连续反应与群体的平均行为存在系统性差异。

现有用于比较时变治疗的因果推断方法大多假设无混杂性，也称为无未测量混杂因素假设或序贯可交换性，即在给定某些观测变量的条件下，治疗分配与潜在结局相互独立。可能存在会混杂治疗分配与生物标志物动态的未测量因素，这违反了这一基本假设，因而削弱了这些方法的有效性，包括 g估计（J. Robins，1986年；Zhou 等人，2019年）、结构嵌套模型（He、Stephens-Shields 和 Joffe，2015年）、历史限制边际结构模型（Neugebauer、van der Laan、Joffe 和 Tager，2007年），以及纵向定向最大似然估计（Van der Laan、Rose 等人，2011年）。另一方面，计量经济学文献采用未观测效应模型（UEM）或个体固定效应模型（Gunasekara、Richardson、Carter 和 Blakely，2014年；Kaufman，2008年），通过纳入个体特异性截距并让每位受试者作为自身对照，消除时间不变的未测量混杂。Imai 和 Kim（2019年）将 UEM 用于匹配，以估计同期治疗效应，即在较短时间内比较治疗状态发生变化前一刻和变化后立即测得的结局。使用 UEM 的主要缺点是，由于它假设严格外生性，因此难以同时处理反向因果和时变混杂所造成的偏倚（Allison、Williams 和 Moral-Benito，2017年）；而这些问题在动态治疗方案的因果比较中很常见。

从建模角度看，我们通过多变量广义线性混合效应模型（MGLMM）（Achana 等人，2021年；Zeger 和 Karim，1991年），对治疗分配和生物标志物动态中的未测量患者异质性进行建模。该模型利用较大总体中每个个体的重复测量数据，对未观测到的永久性成分进行部分识别。行为与社会科学研究者长期以来一直在涉及纵向数据的研究中使用混合效应模型（Agresti、Booth*、Hobert* 和 Caffo*，2000年；Berger 和 Tan，2004年；Laird 和 Ware，1982年；Luger、Suls 和 Vander Weg，2014年；Raudenbush 和 Bryk，2002年）。混合效应模型能够估计个体特异性随机效应，从而对由未观测因素导致的个体间异质性进行定量描述（Bolger 和 Laurenceau，2013年；Schwartz 和 Stone，2007年）。此外，这些模型还能描述时变结果中的个体内依赖性，从而提高参数估计效率。然而，由于多变量广义线性混合效应模型中的链接函数是非线性的，广义模型中的估计参数通常只有在给定随机效应的条件下才具有因果解释；也就是说，即使所有协变量均为外生变量，固定效应系数也无法基于潜在结局得出边际因果效应（Greenland、Pearl 和 Robins，1999年）（Heagerty，1999年；Zeger、Liang 和 Albert，1988年）。

为了解决这一问题，并能够在总体和亚组层面使用多元广义线性混合模型估计比较不同治疗方案的边际因果效应，我们采用了 g 计算算法，这是大多数贝叶斯因果推断方法的基础。该方法基于时间变化的混杂因素与结局在给定患者病史条件下的联合分布，直接模拟治疗路径下的潜在结局；如果所有条件分布均设定正确，就能一致地估计潜在结局，进而估计因果效应。当存在未测量混杂因素时，标准的 g 估计方法会产生有偏的效应估计值，因为未观测到的潜在结局并非随机缺失。从敏感性分析的角度来看，Yang 和 Lok（2018）假设存在一个不可识别的偏倚函数，用以量化未测量混杂对结构嵌套均值模型下平均潜在结果的影响

。Sitlani、Heagerty、Blood 和 Tosteson（2012年）以及 Qian、Klasnja 和 Murphy（2020年）比较了仅在某一时间点存在差异的治疗路径，并讨论了似然分解；这支持了对线性混合模型所估计固定效应系数的因果解释，即将其视为结构嵌套模型的“瞬时效应”。Shardell 和 Ferrucci（2018年）将联合混合效应模型纳入 g 计算算法，以估计治疗方案随时间变化的总体平均效应。

本文放宽了无混杂假设，并利用多元广义线性混合模型（MGLMM）提出了一个用于因果比较治疗路径的框架。该框架将未测量的时不变因素作为治疗分配、纵向结局和时变混杂因素中的潜在个体异质性加以处理。我们的目标是综合与个体临床决策相关的总体证据，并解释个体轨迹的动态进展，同时处理选择偏倚中未观测的永久性因素，并遵循因果推断的一般理念，即仅利用过去的信息推断当前状态。现有使用混合效应模型开展纵向数据因果推断的研究，通常会对潜在成分进行边际化，并将因果估量确定为治疗路径、协变量和固定效应系数的函数。尽管影响疾病进展的未观测稳定特质因素随时间保持不变，我们的方案仍基于受试者随时间累积的观测或反事实历史，通过在纵向结局模型和时变混杂因素模型中依次估计个体特异性潜变量，动态更新与这些因素相关的信息。此外，我们注意到，当治疗路径为二元且单调时，治疗分配异质性的分布并不能被完全识别，因为它不是一个重复发生的过程。我们的讨论聚焦于二元单调治疗过程，并将治疗分配异质性总体分布中的方差引入为治疗方案比较的内置敏感性参数。

我们的方案将治疗分配模型纳入更大的整体框架，旨在弥合选择偏倚中的混杂性与患者疾病进展动态异质性之间的鸿沟。这项工作具有若干优势。首先，现有将倾向评分（PS）纳入贝叶斯因果推断的方法（D. Li、Iddi、Thompson、Donohue 和 Initiative，2019年）包括：指定以倾向评分为条件的结局分布（Zhou 等人，2019年）、在倾向评分模型与结局模型之间采用共享先验分布，或使用逆概率加权或双重稳健估计量（Schnitzer、Sango、Ferreira Guerra 和 Van der Laan，2020年）；我们的方法通过模型各组成部分中受试者特异的未观测异质性的依赖结构，建立倾向评分与结局及随时间变化的混杂因素之间的关联，提供了一种新的方法。其次，我们的方法通过多元广义线性混合模型（MGLMM）中的随机效应，自然地纳入未测量的时间不变因素；其估计协方差可为未测量混杂因素可能存在提供部分信息。第三，我们提出一种新的视角来考察潜在的时间不变未测量混杂的影响：将治疗分配异质性的分布作为因果估计中涉及的敏感性参数，而不是在事后敏感性分析中量化未测量混杂因素（J. M. Robins、Rotnitzky 和 Scharfstein，2000年；Yang 和 Lok，2018年）。结局和混杂因素模型组成部分中的随机效应反映了疾病进展中生理因素等未观测的稳定特征，而治疗分配异质性通常取决于具体情境，并可能依据对数据收集和常规实践的了解加以处理。因此，敏感性参数可以根据临床医生的需求进行调整，作为可控组成部分来检验因果估计值的敏感性。最后，在某些条件下，例如假设不存在治疗分配异质性、因而也不存在未测量混杂因素时，我们的方法无需对敏感性参数作出额外假设，即可识别边际亚组治疗效应。

2 符号与模型

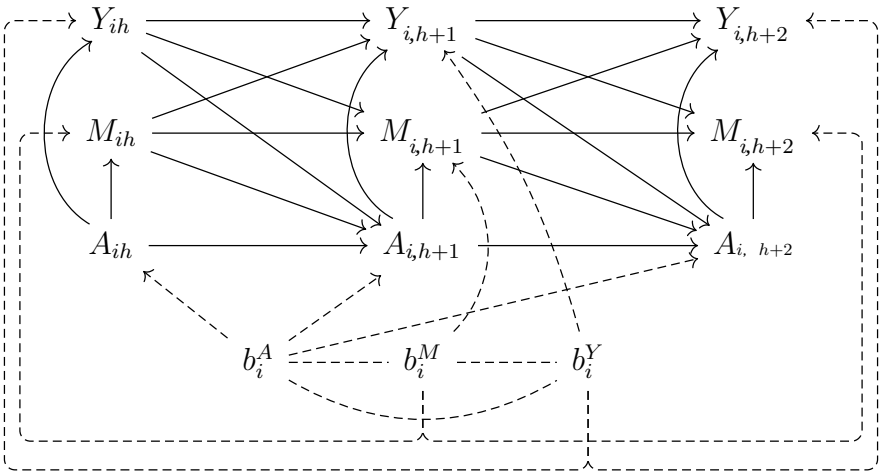

图1：用于广义线性混合模型的有向无环图（DAG），展示观测变量和处理分配及生物标志物动态中时间不变未测量异质性的时间顺序。为简洁起见，基线特征 V_i 未在图中显示。

我们考虑一项涉及序贯分配治疗路径的纵向研究，并假设生物标志物动态和治疗分配中均存在时间不变的未观测异质性。本文按照图1所述的变量时间关系展示该方法：箭头表示可能存在因果关系（单箭头）或协方差（无箭头），缺少箭头则表示影响或协方差为零。存在多元随机过程，$\{(Y_t, M_t, A_t): \ t \geq 0\}$，其中 Y_t、M_t和 A_t 分别表示结局过程、时变混杂因素和序贯治疗，其中 $A_t \in \{0, 1\}$。混杂因素受既往暴露影响，并影响未来结局和治疗分配。令$Y_{i、t_1:t_2}$、$M_{i、t_1:t_2}$以及$A_{i、t_1:t_2}$分别表示受试者 i在时间 $t = t_1$、.....、t_2 观察到的生物标志物、混杂因素和干预纵向路径。任意时间 t，临床医生根据截至时间 t记录的临床病史决定$A_{i、(t+1)}$，即既往治疗路径$A_{i、0:t}$ 以及测量历史$\mathcal{H}_{i、t+1} = (V_i、Y_{i、0:t}、M_{i、0:t})$，其中 V_i为基线信息向量。更新后的临床病史（$\mathcal{H}_{it}、A_{i、0:(t+1)}$）包含最新的治疗决策，随后作为可观测信息，用于解释（$Y_{i、t+1}、M_{i、t+1}$）的动态变化。

本文将讨论范围限定于研究治疗启动，假设治疗在某一时点启动，且受试者在启动后持续接受治疗。为不失一般性，我们假设结局为连续型，时变混杂因素为受试者的访视模式。根据纵向结局的缺失结构，我们将混杂因素建模为二元变量。我们提出采用下文所述的纵向多元广义线性混合模型（MGLMM），以刻画生物标志物和治疗分配在个体层面的时间特异性进展。对于 $t = 1, \ldots, T$，连续结局的线性混合效应模型设定为

$$
\begin{aligned}
Y_{it} &= f_A(\mathcal{H}_{it}, \overline{A}_{i,0:t}, b_i^Y; \theta^Y, \psi_{it}^Y), \quad \mathbb{E}(Y_{it}|\mathcal{H}_{it}, \overline{A}_{i,0:t}, b_i^Y; \theta^Y) = \lambda_Y^{-1}(\eta_{it}^Y) \\
\eta_{it}^Y &= \phi_1^Y(\mathcal{H}_{it})\beta_1^Y + \phi_2^Y(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T\beta_2^Y + \phi_3^Y(\mathcal{H}_{it})b_{i0}^Y + \phi_4^Y(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T b_{i1}^Y, \qquad (1)
\end{aligned}
$$

其中，λ_Y为链接函数，$\phi_A(\overline{A_{i,0:t}})$可以是个体 i 在时间 t 且最大剂量为 K时的剂量信息函数，例如 $(\mathbb{1}\{\sum_{s=1}^{t} A_{is}=1\},\ldots,\mathbb{1}\{\sum_{s=1}^{t} A_{is}=K\})$；$b_i^Y=(b_{i0}^Y,b_{i1}^Y)$为随机效应向量；$\psi_i^Y$为服从零均值分布的随机性，例如 $N(0,1)$；且 $\phi_2^Y(\mathcal{H}_{it})\subseteq\phi_1^Y(\mathcal{H}_{it})$、$\phi_3^Y(\mathcal{H}_{it})\subseteq\phi_1^Y(\mathcal{H}_{it})$，以及 $\phi_4^Y(\mathcal{H}_{it})\subseteq\phi_2^Y(\mathcal{H}_{it})$。结局模型参数可以表示为 $\theta^Y=(\beta_1^Y,\beta_2^Y,\sigma)$，其中 σ 为结局分布的标准差。

治疗启动被建模为

$$
\begin{aligned}
&(A_{it}=1|A_{i,t-1}=0)\sim f_A(\mathcal{H}_{it},b_i^A;\theta^A,\psi_{it}^A),\quad \mathbb{E}(A_{it}|A_{i,t-1}=0,\mathcal{H}_{it},b_i^A;\theta^A)=\lambda_A^{-1}(\eta_{it}^A),\\
&\eta_{it}^A=\phi_1^A(\mathcal{H}_{it})\beta_1^A+\phi_2^A(\mathcal{H}_{it})b_{i0}^A,
\end{aligned}
\tag{2}
$$

其中，λ_A 为logit函数，$\theta^A=(\beta_1^A,\ \beta_2^A)$，$b_i^A=b_{i0}^A$ 为随机效应，且 $\phi_2^A(\mathcal{H}_{it})\subseteq\phi_1^A(\mathcal{H}_{it})$。对于二元因变量，随机性满足 $\psi_{it}^A\sim U(0,\ 1)$，并表示在$A_{i,\ t-1}=0$条件下，通过 $\mathbb{1}\{\psi_{it}^A\leq\lambda_A^{-1}(\eta_{it}^A)\}$ 得到的 A_{it} 的一个实现。我们认识到，当治疗分配不是重复过程，即每位受试者至多发生一次时，治疗分配异质性 b_i^A的分布无法从观测数据中被完全识别。为实现模型估计中的可识别性，在这种情况下有必要对 b_i^A的方差设定取值。

对于时变混杂因素 M_{it}，模型设定与公式（1）类似，

$$
\begin{aligned}
&M_{it}\sim f_M(\mathcal{H}_{it},\overline{A}_{i,0:t},b_i^M;\theta^M,\psi_{it}^M),\quad \mathbb{E}(M_{it}|\mathcal{H}_{it},\overline{A}_{i,0:t},b_i^M;\theta^M)=\lambda_M^{-1}(\eta_{it}^M),\\
&\eta_{it}^M=\phi_1^M(\mathcal{H}_{it})\beta_1^M+\phi_2^M(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T\beta_2^M+\phi_3^M(\mathcal{H}_{it})b_{i0}^M+\phi_4^M(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T b_{i1}^M,
\end{aligned}
\tag{3}
$$

其中，若混杂因素为分类变量，则为 $\theta^M=(\beta_1^M,\ \beta_2^M)$，$b_i^M=(b_{i0}^M,\ b_{i1}^M)$为随机效应向量，且 $\phi_2^M(\mathcal{H}_{it})\subseteq\phi_1^M(\mathcal{H}_{it})$、$\phi_3^M(\mathcal{H}_{it})\subseteq\phi_1^M(\mathcal{H}_{it})$、$\phi_4^M(\mathcal{H}_{it})\subseteq\phi_2^M(\mathcal{H}_{it})$。在示例应用中，$M_{it}$ 表示结局的缺失指示变量，因此我们将 λ_M 设为logit函数。为了使二元混杂因素具有可识别性，η_{it}^M必须采用参数化设定，并且我们假设加性模型。随机性向量（ψ_{it}^Y，ψ_{it}^M，ψ_{it}^A）为独立同分布，且独立于 b_i；（ψ_{it}^Y，ψ_{it}^M，ψ_{it}^A）刻画反事实实现的随机性。为控制随机性，我们在不同治疗方案下使用同一组随机性进行因果估计，以确保投影的潜在结局具有可比性和可重复性。

三个模型组成部分 (1)、(2) 和 (3) 通过随机效应之间的协方差结构相互连接，

$$
b_i=(b_i^Y,b_i^M,b_i^A)^T\sim MVN(0,G_i).
$$

在我们的应用中，我们假设 G_i 在所有受试者间相同，即 $b_i\sim MNV(0,\ G)$。在假定结构下，进一步通过个体层面的参数化，可实现受试者特异的协方差 G_i。这些随机效应可解释为未观测到的、随时间不变的受试者特异异质性；它们代表稳定特征，通过随机截距直接影响临床轨迹和治疗分配过程，并通过随机斜率间接影响各因素的效应。具体而言，b_i^A 是治疗分配中的未测量静态异质性，例如患者的虚弱状态虽在临床中被观察到，却未被记录下来。在不失一般性的前提下，对于本文其余部分，我们假设$\psi_{it}^Y\sim N(0,\ 1)$，对 λ_Y 使用恒等连接函数，并对 λ_A 和 λ_M 使用Logit连接函数。

3 使用多元广义线性混合模型进行贝叶斯g计算

3.1 因果量与目标估计量

到目前为止，我们一直关注如何使用多元广义线性混合模型描述数据生成机制，如图1所示。当多元广义线性混合模型设定正确时，

模型参数的后验预测样本会集中于真实数据分布。由于存在随机效应，多数情况下，多元广义线性混合模型中的模型参数不具有因果解释，但补充材料中介绍了一种例外情况。

治疗方案将患者当前的治疗状态动态定义为观测到的或反事实临床病史的函数 $q(\cdot)$。也就是说，给定截至时间 t 的既往治疗路径和测量历史（$A_{i,0:(t-1)}, \mathcal{H}_{it}$），治疗方案 q 下的治疗序列由 $a_t(q) = q(A_{i,0:(t-1)}, \mathcal{H}_{it})$ 依次确定。对于任意变量 X， $X(q)$ 表示个体在接受治疗方案 q 下的 X 值。我们将 $Y_{i,0:t}(q)$、$M_{i,0:t}(q)$ 和 $a_{0:t}(q)$ 定义为 $=(a_1(q)\ldots, a_t(q))$，即治疗方案 q 下结局、混杂因素和治疗路径的反事实纵向轨迹；并将治疗方案 q 下截至时间 t 之前的反事实测量历史记为 $\mathcal{H}_{it}(q) = \{V_i, Y_{i,t-1}(q)\, M_{i,t-1}(q)\}$。

治疗分配异质性 b_i^A 表示临床医生观察到的私有信号或稳定特质因素，这些因素未被数据捕捉，但与临床医生对患者潜在结局的判断相关。为考虑在临床治疗患者的过程中可能自然出现的时间不变的未测量混杂，我们根据治疗分配异质性对因果估计进行分层。给定一个特定的目标治疗方案 q 以及未观测到的时间不变异质性 b_i^A，我们的目标是识别未来 τ 天反事实轨迹在截至当前时间 h 的观测历史条件下的联合分布，

$$P(\overline{Y}_{(h+1):(h+\tau)}(q), \overline{M}_{(h+1):(h+\tau)}(q) | V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A). \tag{4}$$

基于（4）和 g计算（J. Robins，1986年），我们可以通过对在感兴趣的治疗方案所确定的治疗路径下的观测或反事实历史进行积分，从反事实结局的期望中识别因果效应；这些期望是固定效应参数以及（b_i^Y， b_i^M）随时间变化的估计值的函数。

假设我们关注目标亚组 T 内的条件混合平均治疗效应（F. Li、Ding 和 Mealli，2022年）（CMATE），该亚组的特征为 $\{V_i, A_{i,0:h_i}, Y_{i,0:\ h_i},\ M_{i,\ 0:h_i}\}_{i\in T}$，其中个体 i 为该亚组提供截至时间 h_i 的历史信息。例如，我们的应用研究考察了一个亚组 T，其中包含硬皮病^患者在使用吗替麦考酚酯之前的随访信息，这些患者被观察到接受了吗替麦考酚酯治疗。令 P_T 表示亚组 T 中观测值的经验分布，则目标量 CMATE 定义为

$$\int \mathbb{E}(Y_{h_i+\tau}(q) | V_i, \overline{A}_{0:h_i}, \overline{Y}_{0:h_i}, \overline{M}_{0:h_i}) d\widehat{P}_T = \frac{1}{N_T} \sum_{i\in T} \mathbb{E}(Y_{h_i+\tau}(q) | V_i, \overline{A}_{0:h_i}, \overline{Y}_{0:h_i}, \overline{M}_{0:h_i}), \tag{5}$$

其中， N_T 是亚组 T中的个体数。将经验分布替换为相应的总体分布，即可得到总体层面的条件平均处理效应（CMATE）；它可视为条件平均治疗效应的纵向扩展，并可通过联合建模定义亚组 T所涉及的所有变量来实现。

3.2 假设和方法

本节介绍估计条件平均处理效应（CMATE）的假设与步骤。该方法对多元时变成分进行联合建模，并通过对时间不变的未观测特征（b_i^Y， b_i^M）进行随时间演化的更新，考虑个体信息随时间的累积。该方法使我们能够评估纵向因果效应估计对不同未观测处理异质性分布的敏感性，同时允许存在

时间不变的未测量混杂。为简便起见，以下讨论中省略下标 i 。为说明无需指定参数形式即可识别条件反事实联合分布（4），我们作出以下假设:

假设。对于 $t = 0, \ldots\ldots, T$,

1. 一致性：$Y_{0:\overline{t}} = Y_{0:\overline{t}}(q)$ 且当 $\overline{A_{0:t}} = \overline{a_{0:t}}(q)$ 时，$\overline{M}_{0:t} = M_{0:t}(q)$ ；*2.* 正值性：$P(A_{t+1} = a_{t+1}(q)|V, A_{0:t} = a_{0:t}(q), Y_{0:t}, M_{0:t}, b_i^A) > 0$ 对 $t \geq 0$ 以概率 *1* 成立;

3. 给定 b_i^A的序贯可交换性：对于 $\tau > 0$,

$$P(\overline{Y}_{(t+1):(t+\tau)}(q), \overline{M}_{(t+1):(t+\tau)}(q)|V, A_{t+1}, \overline{A}_{0:t} = \overline{a}_{0:t}(q), \overline{Y}_{0:t}, \overline{M}_{0:t}, b_i^A)$$
$$= P(\overline{Y}_{(t+1):(t+\tau)}(q), \overline{M}_{(t+1):(t+\tau)}(q)|V, \overline{A}_{0:t} = \overline{a}_{0:t}(q), \overline{Y}_{0:t}, \overline{M}_{0:t}, b_i^A).$$

一致性假设指出，当观察到的治疗路径遵循所关注的假设治疗方案时，观察到的生物标志物动态与反事实生物标志物动态等价。需要注意的是，这种等价并不意味着取值相同，而是指分布相同。正值性保证在一段时间内不会系统性地排除合理的治疗模式。现有文献通常采用经典序贯可交换性假设（Greenland & Robins，1986年），假定观察到的治疗前历史足以解释当前治疗分配与未来反事实之间的依赖关系。我们将这一假设扩展为以治疗分配中未观察到的时间不变异质性为条件，该异质性由模型 (2) 中的随机效应 b_i^A 量化。在实践中，治疗分配异质性可能源于患者接受治疗的意愿、治疗可能带来的不良反应风险，以及临床医生对治疗的看法。在这些假设成立且模型 (1)、(2) 和 (3) 设定正确的情况下，(4) 可以非参数识别如下（详情见附录 A），

$$\begin{aligned}
&P(\overline{Y}_{(h+1):(h+\tau)}(q), \overline{M}_{(h+1):(h+\tau)}(q)|V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A) \\
&= \prod_{s=h}^{h+\tau-1} \int_{u_s} \int_{v_s} P(Y_{s+1}|V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^Y = u_s) \\
&\qquad P(M_{s+1}|V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^M = v_s) \\
&\qquad P(b_i^Y = u_s, b_i^M = v_s|V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A) du_s dv_s. \qquad (6)
\end{aligned}$$

在图2中，我们使用单世界干预图（Hern´an & Robins，2010年；Richardson & Robins，2013年）（单世界干预图）来展示导出（6）所需的独立关系，并说明如果将治疗路径设定为治疗方案 q下的路径，将会存在怎样的反事实依赖关系。该图通过拆分图1因果图中的处理节点$A_{i,\ (h+1):(h+\tau)}$，并将指定处理的所有后代替换为其潜在结果而构建，同时用红色标示所有反事实。单世界干预图通过反事实轨迹($Y_{(h+1):(h+\tau)}(q)$ $M_{(h+1):(h+\tau)}(q)$)与$A_{i,\ h+1}$之间的 d-分离，且以($A_{0:h}$, $Y_{0:\ h}$, $M_{0:h}$, b_i^A)为条件，展示了条件序贯可交换性假设。如果不控制 b_i^A，路径$A_{i,\ h+1} \leftarrow b_i^A \leftrightarrow \overline{b_i^Y} \rightarrow \overline{Y}_{i,\ h+1}(q)$和$A_{i,\ h+1} \leftarrow b_i^A \leftrightarrow b_i^M \rightarrow M_{i,\ h+1}(q)$会导致选择偏倚，而按 b_i^A分层则会阻断这些路径。图2矩形框中的变量是在（6）中涉及且与.....相关的量

（b_i^Y， b_i^M）随时间演化的更新。在每个时间点，基于从已观测或反事实生物标志物历史中回流的信息，可以推导出（b_i^Y， b_i^M）的分布，从而对这些受试者特异的未观测永久特征进行序贯更新。假设的处理状态a_{it}（q）， $t \in [h+1, h+\tau]$，不会参与（b_i^Y， b_i^M）的序贯更新，因为除定义所关注的治疗方案外，它不会带来额外信息。

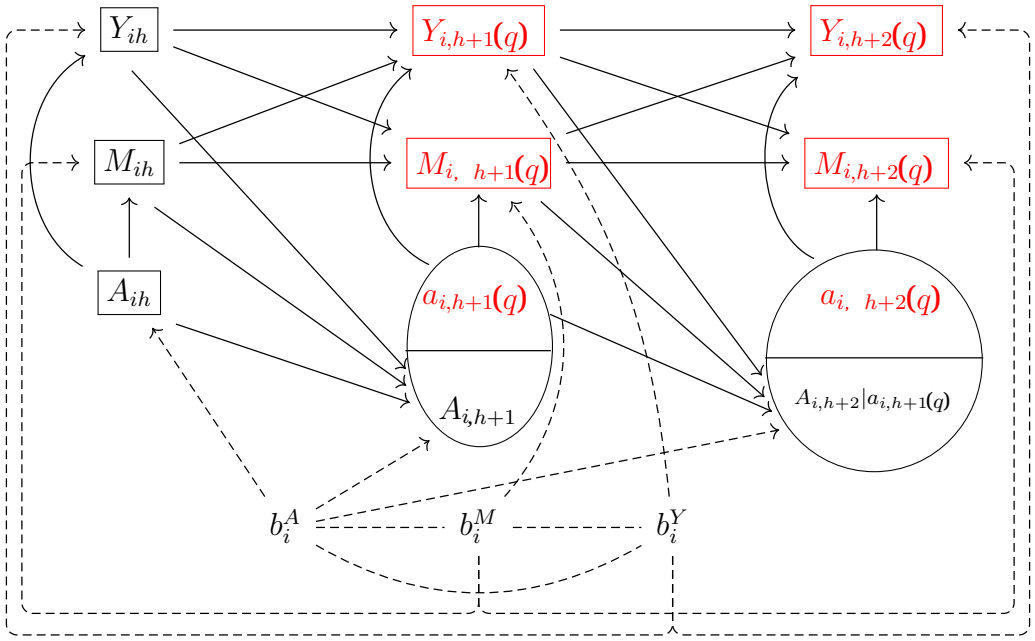

图2：单世界干预图。

在每个时间点 $s \in [h, h+\tau)$，基于式（6）对反事实结局和混杂因素（$Y_{s+1}(q)$ $M_{s+1}(q)$）进行的蒙特卡罗模拟，涉及对 $P(b_i^Y, b_i^M | V, A_{0:h}, Y_{0:s}, M_{0:s}, b_i^A)$ 积分；这是（b_i^Y, b_i^M）的更新条件后验分布。给定的轨迹（$Y_{0:s}, M_{0:s}$）在分布上等价于（$Y_{0:h}, Y_{(h+1):s}(q)$ $M_{0:h}, M_{(h+1):s}(q)$），其中混合了观测变量和反事实变量。请注意，反事实轨迹（$Y_{(h+1):s}(q)$ $M_{(h+1):s}(q)$）具有如下分布

$$\prod_{s=h}^{s-1} P(Y_{s+1}, M_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A).$$

在治疗方案 q 和治疗分配异质性 b_i^A 下，正如式（6）的表述所示。由于多元广义线性混合模型中的链接函数可能是非线性的，对（b_i^Y, b_i^M） $\sim P(b_i^Y, b_i^M | V, A_{0:h}, Y_{0:s}, M_{0:s}, b_i^A)$ 进行抽样可能较为复杂。我们考虑以下一般策略：首先，计算后验分布（$b_i^Y, b_i^M, b_i^A | V, A_{0:h}, Y_{0:s}, M_{0:s}$）的拉普拉斯近似，记为 $\hat{} MVN(b_i, V)$，然后在将 b_i^A 设为某个值的条件下，根据相应的条件分布（b_i^Y, b_i^M） $|b_i^A$ 对异质性进行抽样。该过程见附录B。附录C提供了基于式（6）从 $P(Y_{(h+1):(h+\tau)}(q)$ $M_{(h+1):(h+\tau)}(q) | V, A_{0:h}, Y_{0:h}, M_{0:h}, b_i^A)$ 生成反事实轨迹后验样本的伪代码。

在我们的讨论中，到目前为止一直按b_i^A进行分层。方程（5）中表示的目标估计量条件平均处理效应是边际亚组平均治疗效应，即对未观测异质性进行边际化。因此，

使用以下公式，我们根据给定亚组 T 的条件分布，对条件平均处理效应的每个组成部分进行积分，

$$\begin{aligned}&\mathbb{E}(Y_{h+\tau}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h},\overline{M}_{0:h})\\&=\int_w \mathbb{E}(Y_{h+\tau}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h},\overline{M}_{0:h},b_i^A=w)P(b_i^A=w|V,\overline{A}_{0:h},\overline{Y}_{0:h},\overline{M}_{0:h})dw.\end{aligned} \tag{7}$$

因此，条件平均处理效应是反事实联合分布 (4) 的一个泛函，因为 (7) 中的条件期望 $\mathbb{E}(Y_{h+\tau}(q)|V, A_{0:h}, Y_{0:h}, M_{0:h}, b_i^A)$ 可以表示为

$$\begin{aligned}&\mathbb{E}(Y_{h+\tau}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h},\overline{M}_{0:h},b_i^A)\\&=\int_{y_{h+\tau}}\int_{m_{h+\tau}}\cdots\int_{y_{h+1}}\int_{m_{h+1}}\\&\quad y_{h+\tau}P(\overline{Y}_{(h+1):(h+\tau)}(q)=\overline{y}_{(h+1):(h+\tau)},\overline{M}_{(h+1):(h+\tau)}(q)=\overline{m}_{(h+1):(h+\tau)}|V,\overline{A}_{0:h},\overline{Y}_{0:h},\overline{M}_{0:h},b_i^A)\\&\qquad dm_{h+1}dy_{h+1}\ldots dm_{h+\tau}dy_{h+\tau}.\end{aligned}$$

比较治疗方案 q_1 和 q_2时，我们通过对以下内容进行积分来估计因果对比

$$\mathbb{E}(Y_{h+\tau}(q_1)|V,\overline{A}_{0:h},\overline{Y}_{0:h},\overline{M}_{0:h},b_i^A)-\mathbb{E}(Y_{h+\tau}(q_2)|V,\overline{A}_{0:h},\overline{Y}_{0:h},\overline{M}_{0:h},b_i^A)$$

其积分范围为治疗分配异质性的亚组分布 $P(b_i^A|V, A_{0:h}, Y_{0:h}, M_{0:h})$。附录 C 介绍了计算条件平均处理效应的计算细节，示例应用则展示了如何使用所提出的方法进行亚组因果效应估计。

作为控制并进行边际化处理的 b_i^A 的亚组分布，是计算条件平均处理效应时使用的敏感性参数。回顾一下，多元广义线性混合模型假设 (b_i^Y, b_i^M, b_i^A) 服从$MVN(\mathbf{0}, G)$。令 v 表示 b_i^A 的方差，它代表假定的受试者间治疗分配异质性程度。当治疗分配如示例应用中那样是一个二元单调过程时，v 无法识别，因此在模型估计中需要指定其取值，因为治疗分配并非重复发生的事件。当亚组 T 包含历程长度不同的个体历程时，可针对每个个体在给定的 v 值下，推导亚组分布 $P(b_i^A|V, A_{0:h}, Y_{0:h}, M_{0:h})$。因果效能的评估可能随 v 的不同取值而变化。在这种情况下，不可识别的时不变量 b_i^A 的方差可作为因果效应估计中的敏感性参数。对于总体平均处理效应，目标量可推导为

$$\mathbb{E}(Y_{h+\tau}(q))=\int_w \mathbb{E}(Y_{h+\tau}(q)|,b_i^A=w)P(b_i^A=w)dw,$$

其中 $b_i^A \sim N(\mathbf{0}, v)$ 基于模型假设。附录 A 进一步详述了目标人群的混合平均处理效应（F. Li 等，2022）的计算方法$\widehat{\ }$，<style id='6>E($Y_{h+\tau}(q)$)，该方法以相应的经验分布替代目标人群分布。

当 MGLMM 中不存在治疗异质性，即设定 $v = 0$ 时，该方法便简化为标准 g-计算，仅利用结局和混杂因素的模型组成部分，因为分配机制不存在混杂（F. Li 等，2022）。满足 $v = 0$ 是不存在未测量混杂因素的充分但非必要条件。在 MGLMM 中，cov(b_i^A, b_i^M) = cov(b_i^A, b_i^Y) $= 0$ 意味着不存在未测量混杂因素。当不存在未测量混杂因素时，

MGLMM 仍允许未观测因素影响治疗分配，即 $v \neq 0$，只要b_i^A与生物标志物动态中的未观测异质性（b_i^Y, b_i^M）不相关；这类非混杂性治疗分配异质性的例子包括基于个人信念或社会污名而偏好某种治疗。另一方面，我们注意到，在给定 b_i^A, v 的方差时，协方差 cov(b_i^A, b_i^M) 和 cov(b_i^A, b_i^Y) 是可估计的。因此，我们的方法能够根据 MGLMM 估计的协方差，揭示未测量混杂因素可能存在的情况。

4 模拟

假设每个人有两次随访访视，$T_i = 2$，我们通过以下方式模拟连续型生物标志物 Y_{it} 和二元时变治疗 A_{it}

$$Y_{it} = 0.4 - 0.3V_i - 0.1t + \sum_{k=1}^{2} \frac{\nu_k}{2} \times \mathbb{1}\Big\{ \sum_{s=1}^{t} A_{is} = k \Big\} + 0.4Y_{i,t-1} + b_{i0}^Y + e_{it}^Y \text{ and}$$
$$\text{logit}\{P(A_{it}(s) = 1 | A_{i,t-1}(s) = 0)\} = -0.1V_i - 0.5t - 0.35Y_{i,t-1} + b_{i0}^A,$$

其中 $e_{it}^Y \sim N(0, 0.42)$，V_i是基线协变量，$V_i \sim$Bernoulli(0.5)，$Y_{i0} \sim N(0, 1)$。记$\rho = \text{Corr}(b_{i0}^A, b_{i0}^Y)$，$s_A = \sqrt{\text{Var}(b_{i0}^A)}$，$s_Y = \sqrt{\text{Var}(b_{i0}^Y)}$。我们假设随机效应服从 $(b_{i0}^A, b_{i0}^Y) \sim N(0, G)$，其中协方差矩阵 G 的元素为$G_{11} = s_A^2$、$G_{12} = G_{21} = \rho s_A s_Y$和$G_{22} = s_Y^2$。我们设定 $s_Y = 0.8$，并从参数组合 $(s_A, \rho) \in (0, 0) \cup \{(s_A, \rho);\ s_A \in \{0.1, \ldots, 0.9, 1\}, \rho \in \{0, 0.1, \ldots, 0.9\}\}$定义的 101 种设置中，每种随机抽取 100 个重复样本。当 $s_A > 0$ 且 $\rho < 1$时，由于 $\det(G) = (1 - \rho)s_A^2 > 0$，$G$ 保证为正定。当 $\rho = 0$时，矩阵 $G = \begin{pmatrix} s_A^2 & 0 \\ 0 & s_Y^2 \end{pmatrix}$ 表示不存在未测量混杂的情形。对于 $k = 1$，我们考虑两种情景：$\nu_k = k$ 和 $\nu_k = 0$，其中 $\nu_k = k$ 表示随时间推移治疗效应更强，而 $\nu_k = 0$ 表示没有治疗效应。对于每种 ν_k 情景以及 (s_A, ρ) 的每种设置，我们模拟样本量为 $n = 500$ 的数据重复样本 100 份。

将始终治疗方案定义为 q_1，未接受治疗定义为 q_0。对于每个模拟数据集，我们旨在比较在始终接受治疗方案 $a_{1:2}(q_1) = (1, 1)$与未接受治疗方案 $a_{1:2}(q_0) = (0, 0)$下，$t = 2$ 时的总体混合平均处理效应（F. Li 等人，2022）。根据一致性和条件序贯可交换性假设，对于任何感兴趣的治疗方案 q，即 $a_{1:2}(q) = (a_1, a_2)$，目标反事实量可通过 g公式表示为观测变量的函数，即 $\mathbb{E}[Y_{i2}(q)] = \mathbb{E}(Y_{i2} | A_{i1} = a_1, A_{i2} = a_2)$，推导见附录D。总体平均处理效应是反事实结局 $Y_{i2}(q_1)$与 $Y_{i2}(q_0)$ 之间的平均差，

$$\begin{aligned} g(q_0, q_1) =& \mathbb{E}[Y_{i2}(q_1)] - \mathbb{E}[Y_{i2}(q_0)] \\ =& \mathbb{E}(Y_{i2} | A_{i1} = a_1, A_{i2} = a_2) - \mathbb{E}(Y_{i2} | A_{i1} = 0, A_{i2} = 0) \\ =& \sum_{k=1}^{2} \frac{k}{2} \times 1\{a_1 + a_2 = k\} + 0.4 \times 0.5 a_1. \end{aligned} \tag{8}$$

我们用混合平均处理效应估计总体平均处理效应，$\widehat{g}(q_0, q_1) = \mathbb{E}[Y_{i2}(q_1)] - \mathbb{E}[Y_{i2}(q_0)]$, 其中，我们使用样本特定经验分布 $P(V = v)$，而非 $P(V = v)$，即推导总体平均处理效应时所涉及的总体分布。

根据公式（8），情景 $\nu_k = k$ 下真实总体平均处理效应为 $g(q_0, q_1) = 1.2$，情景 $\nu_k = 0$下则为 0。对于每个数据重复样本，我们通过在 $\widehat{s}_A \in \{0, 0.3, 1\}$下对 $b_i^A \sim N(0, \widehat{s}_A^2)$ 边际化，抽取混合平均处理效应的后验预测分布；其中，$\widehat{s}_A$ 表示对未解释的治疗分配异质性变化程度的假设。模拟真实值（s_A, ρ）与假设参数 $\widehat{s}_A$ 的不同组合用于考察以下三种情形：（1）不存在未测量混杂（$\rho = 0$）；（2）存在未测量混杂，且模型设定正确（$\rho \neq 0, \widehat{s}_A = s_A$）；（3）存在未测量混杂，且治疗分配异质性的程度设定错误（$\rho \neq 0, \widehat{s}_A \neq s_A$）。请注意，关于未测量混杂的表述基于因果有向无环图中编码的假设以及模型选择。令 N_{post} 为后验抽样的数量。给定某个模拟设定下第 r 个数据重复样本，我们用后验均值 $g^{(r)} = \sum_{\ell=1}^{N_{post}} g_\ell^{(r)} / N_{post}$ 和 95% 可信区间（$L^{(r)}, U^{(r)}$）概括混合平均处理效应的后验样本（$g_1^{(r)}, \ldots, g_{N_{post}}^{(r)}$）。我们通过均方误差（MSE）$\frac{1}{100} \sum_{r=1}^{100} [g^{(r)} - g(q_0, q_1)]^2$和覆盖率 $\frac{1}{100} \sum_{r=1}^{100} \mathbb{1}\{g(q_0, q_1) \in (L^{(r)}, U^{(r)})\}$，对模拟重复样本进行汇总。

对于情景 $\nu_k = k$，图3分别在第一行和第二行展示了后验混合平均处理效应的均方误差和覆盖率。从左到右的三列分别对应于 $\widehat{s}_A$为0、0.3和1时的估计结果。对于每个图，横轴和纵轴分别表示真实参数s_A 和 ρ，数据重复样本在这些参数条件下生成。绿色表示因果效应的估计效果更好，即均方误差更低且后验覆盖率更高。我们观察到，当 $\widehat{s}_A$ 不大于真实值 s_A 时，因果效应的估计相对更好。此外，当不存在或几乎不存在未测量混杂，即 ρ 接近于零时，估计得到的混合平均处理效应对于设定值 $\widehat{s}_A$具有稳健性。换言之，当治疗分配异质性中假定的变异与其真实值存在显著差异，且存在严重的未测量混杂时，估计和覆盖率较差。对于 $\widehat{s}_A = 1$，估计似乎始终更好，因此我们还在图4中将 $\widehat{s}_A$ 为0与为1、以及为0.3与为1时的均方误差之比分别进行了可视化。我们看到，当$\widehat{s}_A$ 不大于真实值时，估计效果更好；当不存在未测量混杂这一情况成立时，无论真实值 s_A为何，较小的 $\widehat{s}_A$ 值都更为有利。对于另一种情景 $\nu_k = 0$，处理在所有时间点都对结局没有影响，并且 $\mathbb{E}[Y_{it}(q)]$不取决于治疗路径。图5和图6总结了结果，并得出了与 $\nu_k = k$下相似的结论。当 $\widehat{s}_A$ 接近真实值（$\widehat{s}_A \approx s_A$）或不存在未测量混杂这一情况接近成立（$\rho \approx 0$）时，该方法能够较好地识别零效应估计值。

5 应用

本研究将所提出的方法应用于通过约翰斯·霍普金斯硬皮病中心研究登记库纵向收集的硬皮病患者临床数据。该应用旨在研究接受吗替麦考酚酯治疗的亚组中，启动吗替麦考酚酯治疗方案的因果效能。通过抽样并比较不同治疗方案下反事实结局轨迹在各时间间隔上的后验预测分布来进行推断，其中结局为连续的时变多变量生物标志物。疾病发病定义为症状出现，通常发生在入组就诊之前，并在入组就诊时询问。入组就诊时间在疾病发病后六年内且介于2010年和2020年之间的患者被纳入分析数据。分析使用了截至2022年2月28日的个体临床病史；该数据中观察到的最长随访时间为十年。具体而言，数据包括从未使用吗替麦考酚酯期间所有可用的随访访视，以及首次连续使用吗替麦考酚酯后最多两年的随访访视。研究纳入了506名此前未曾接受过吗替麦考酚酯治疗的硬皮病患者

这些患者在入组时尚未接受吗替麦考酚酯治疗，其中194人在随访期间开始使用吗替麦考酚酯。在这些个体中，80%为女性，20%为非裔美国人，40%患有弥漫性硬皮病；这一构成具有整体队列的代表性。疾病发病年龄的第一、第二和第三四分位数分别为38岁、48岁和58岁。由于预计患者每六个月就诊一次，我们以六个月为间隔来描述时变变量的进展。超过90%的观察到的吗替麦考酚酯治疗启动发生在入组后的前五个时间间隔内。

为研究在假定患者对药物耐受的情况下，连续使用吗替麦考酚酯（MMF）期间的疗效，我们重点关注接受MMF治疗的患者使用MMF的前两年。假设患者 i 在时间 $\tilde{s}_i$开始使用MMF，我们通过比较两年时间区间 $[\tilde{s}_i, \tilde{s}_i + 4)$ 内连续使用MMF与不使用MMF这两种治疗方案，评估MMF的疗效。本研究将结局 Y_{it} 定义为改良Rodnan皮肤评分（mRSS）测量值，以及通过用力肺活量预计值百分比（FVC）和一氧化碳弥散量预计值百分比（DLCO）评估的肺部评分。FVC和DLCO是连续评分，分析时进行了标准化。皮肤评分mRSS是取值范围为0至51的连续指标，我们将mRSS量化为标准正态分布。由于临床数据仅在患者就诊时才会观测到，因此生物标志物测量缺失十分常见。本研究纳入的患者中，分别有82.6%、83.6%和50%至少有一个时间区间未测量FVC、DLCO和mRSS。就诊模式可能会混杂MMF对生物标志物的疗效。我们将混杂因素M_{it} 定义为患者 i 在时间区间 t是否更新了FVC、DLCO和mRSS的指标，用以表示随时间变化的就诊模式。

我们将 V_i 定义为基线人口统计学变量向量，包括性别、种族和年龄。令 B_i 表示基线疾病类型，即弥漫性硬皮病的指示变量。根据领域知识，我们认为生物标志物会随疾病发病后的时间推移而进展，该时间记为 $S_{it} = t + O_i$，其中 O_i 表示从疾病发病到入组就诊的时长。将 Y_{it} 记为时间 t 时生物标志物的前移测量值。由于我们将研究范围限制为连续使用吗替麦考酚酯不超过两年，即四个时间间隔，因此时间 t 的剂量信息可由向量 $D(A_{i,\ 0:t})$ $= \{\mathbb{1}\big(\sum_{\ell=1}^{t} A_{i\ell} = 1\big), \ldots, \mathbb{1}\big(\sum_{\ell=1}^{t} A_{i\ell} = 4\big)\}$ 概括，其中 $A_{i,\ 0:t} = (A_{i1}, \ldots\ldots, A_{it})$ 是二元指示向量，表示是否观察到个体 i 在相应时间段使用吗替麦考酚酯。此外，疾病发病与吗替麦考酚酯治疗启动之间的时间记为 I_{it}；当 $A_{it} = 0$ 时，该值为零；当 $A_{it} = 1$ 时，该值为 $S_{it'}$；其中 t' 是治疗启动时间，满足 $A_{i,\ t'-1} = 0$、 $A_{it'} = 1$ 和 $t' \leq t$。假设结局、时变混杂因素和治疗分配满足以下联合模型，

$$
\begin{aligned}
Y_{it}|(M_{it} = 1) =& \phi_1(\mathcal{H}_{it})\beta_1^Y + \phi_2(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T\beta_2^Y + b_{i0}^Y + e_{it}^Y \\
\text{logit}\{P(M_{it} = 1)\} =& \phi_1(\mathcal{H}_{it})\beta_1^M + \phi_2(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T\beta_2^M + b_{i0}^M \\
\text{logit}\{P(A_{it} = 1|A_{i,t-1} = 0)\} =& \phi_1(\mathcal{H}_{it})\beta_1^A + b_{i0}^A
\end{aligned}
$$

其中

$$
\begin{aligned}
(b_{i0}^Y, b_{i0}^M, b_{i0}^A)^T &\sim N(0, G), \\
\phi_1(\mathcal{H}_{it}) &= \{1, \tilde{Y}_{i,t-1}, V_i, B_i, ns(S_{it}, \nu_s), B_i \times ns(S_{it}, \nu_s)\}, \\
\phi_2(\mathcal{H}_{it})\phi_A(\overline{A}_{i,0:t})^T &= \{D(\overline{A}_{i,0:t}), V_i \times D(\overline{A}_{i,0:t}), B_i \times D(\overline{A}_{i,0:t}), I_{it} \times D(\overline{A}_{i,0:t})\},
\end{aligned}
$$

并且我们假设 $\nu_s = 4$。请注意，结局 Y_{it} 和混杂因素 M_{it} 均为多变量，即Y_{it}和 $M_{it} \in \mathbb{R}^3$。具体而言，$(b_{i0}^Y,\ b_{i0}^M,\ b_{i0}^A) \in \mathbb{R}^7$ 以及协方差矩阵 $G \in \mathbb{R}^{7\times7}$。模型

验证结果汇总于图7。黑色三角形表示各时间点时变变量的观测均值。彩色曲线和区域表示在假定 b_{i0}^{A}的标准差取不同值时，各时变变量一步向前预测的后验均值和95%后验可信区间。

对于每位在随访期间观察到接受吗替麦考酚酯（MMF）治疗的患者 i，我们从时间 $\tilde{s}_i$ 开始比较两种治疗方案 q_1 和 q_2，并以该患者截至时间 $\tilde{s}_i - 1$ 的临床病史为条件，即 $(V_i, A_{i,0:(\tilde{s}_i-1)}$ $Y_{i,0:(\tilde{s}_i-1)}$ $M_{i,0:(\tilde{s}_i-1)})$。根据附录C中概述的算法，我们从 $(Y_{i,\tilde{s}_i:(\tilde{s}_i+3)}(q_z)\, M_{i,\tilde{s}_i:(\tilde{s}_i+3)}(q_z))$ 的后验预测分布中抽样；该分布对应于治疗方案 q_z, $z = 1, 2$ 下的反事实轨迹，并由此获得反事实轨迹的后验样本，记为 $\{Y^{(\ell)}_{i,\tilde{s}_i:(\tilde{s}_i+3)}(q_z)\ M^{(\ell)}_{i\,\tilde{s}_i:(\tilde{s}_i+3)}(q_z);\ \ell = 1, \ldots, N_{post}\}$。两种治疗方案之间的因果比较有效性通过生物标志物随时间变化的平均差异量化，$\mathbb{D}_j = \{\sum_{i \in T} d_{ij}^{(\ell)}/N_T;\ \ell = 1, \ldots, N_{post}\}$，其中 $j \in \{0, 1, 2, 3\}$ 表示自治疗方案实施以来的时间索引，$d_{ij}^{(\ell)} = Y_i^{(\ell)}{}_{,\tilde{s}_i+j}(q_1) - Y^{(\ell)}{}_{i,\tilde{s}_i+j}(q_2)$，而 T 是感兴趣的亚组。图 8 展示了接受治疗的个体中 MMF 的条件平均处理效应，并比较两种治疗方案：按观察到的情况开始使用 MMF与不使用 MMF。该图呈现了两年方案比较期间 $\mathbb{D}_j$ 的后验均值和95%可信区间，即 $j = 0, 1, 2, 3$；在设定的 $\widehat{s}_A \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}$ 取值下，随着 $\widehat{s}_A$ 增大，不透明度逐渐降低，并按弥漫性硬皮病状态分层。

图8显示，在假设患者能够耐受药物并连续使用的情况下，将吗替麦考酚酯纳入弥漫性硬皮病患者的治疗后，药物启动后的两年内对皮肤评分有显著影响。这与硬皮病肺部研究II的结果一致；该研究发现，吗替麦考酚酯显著改善了弥漫性硬皮病患者在24个月结束时的改良Rodnan皮肤评分。我们考察了将吗替麦考酚酯纳入硬皮病患者治疗的情况，而该临床试验比较的是单独使用吗替麦考酚酯24个月与使用环磷酰胺12个月后再使用安慰剂12个月。此外，结果表明，若患者可能同时接受其他治疗，则没有充分证据显示吗替麦考酚酯能改善弥漫性或局限型／无皮肤硬化型硬皮病个体亚组的平均用力肺活量。我们的研究结果还显示，在吗替麦考酚酯治疗启动后的两年内，对不患弥漫性硬皮病的患者而言，将吗替麦考酚酯加入治疗方案，对肺部或皮肤没有明显益处；这一点具有临床意义，因为吗替麦考酚酯是一种免疫抑制剂，可能增加严重感染和胃肠道副作用的风险。请注意，在实际临床中，含或不含吗替麦考酚酯的治疗方案，其药物联合用药和治疗实践可能有所不同；还需进一步研究，以考察这些实践差异以及使用或不使用吗替麦考酚酯所导致的治疗模式差异所带来的影响。

6 讨论

确定哪种治疗方案对特定亚组或病史模式的患者更有利，是临床机构治疗患者时面临的一个基本问题。因果推断是回答此类问题的自然工具，但在评估治疗路径的疗效时，必须纳入临床数据的特征，才能进行有效推断。观察性纵向临床数据集通常包括根据观测到的患者病史而进行的非随机治疗分配，以及可能因患者就诊模式而产生信息性缺失的不规则测量。比较治疗路径时，一个关键因素是治疗分配和生物标志物动态中超出可观测变量解释范围的自然异质性。这些是纵向临床数据集的典型特征。为某一类患者选择最有效的治疗方案，需要在考虑该患者特定生物标志物趋势的同时，总结

来自相似类型患者总体的证据。用于回答这一问题的统计模型本质上较为复杂。本文介绍了能够纳入这些特征的最简单模型，例如非随机治疗分配和患者异质性，并提供了一种比较治疗路径的工具。

本研究的主要贡献是开发一种用于观察性纵向数据因果推断的贝叶斯框架，在考虑时变混杂因素并允许存在时间不变的未测量混杂的同时，估计二元治疗路径对纵向结局的亚组疗效。我们提出使用多元广义线性混合模型同时对生物标志物动态和治疗分配建模；当模型设定与真实情况足够接近时，该方法在一定程度上仍能处理未测量混杂。由于混合效应模型不依赖于「不存在未测量混杂」这一假设，而现有的大多数 g 估计方法都要求满足该假设，因此，即使确实存在未测量混杂，或无法获得无混杂变量或工具变量，我们的方法仍有望一致地估计治疗路径的因果效应。需要指出的是，本文介绍的多元广义线性混合模型无法处理时变未测量混杂。当模型中为时变变量指定随机斜率时，该变量被视为一种时间不变的未观测特征，用以刻画动态进展中的患者异质性。此外，多元广义线性混合模型对未测量混杂有特定的表示方式，其程度由治疗分配中未解释的变异 s_A 以及通过相关参数 ρ起作用的治疗分配与生物标志物动态之间的相关性决定。较小的 ρ 和较大的 s_A，或较大的 ρ 和较小的 s_A，都可能导致严重的未测量混杂。该方法有望进一步扩展，以指导在贝叶斯因果推断中纳入其他潜变量模型。

请注意，我们的提议确实依赖于关于结局、时变混杂因素和治疗分配联合分布的参数假设。除非治疗分配的随机化得到保证或受到控制，否则我们无法检验不存在未测量混杂的假设。在不存在未测量混杂这一强假设下，有效的因果推断并不需要很强的参数模型假设。然而，在临床机构中对慢性罕见疾病患者进行治疗的例子中，未测量混杂因素不可避免地存在。我们认识到，因果推断中不存在免费的午餐；要放宽无混杂假设，需要在模型设定上作出额外假设，而这些假设同样在很大程度上无法检验。观察性数据中的完全非参数因果效应无法识别，而因果效应的识别始终需要无法检验的假设。参数模型假设使得即使可能存在未测量混杂因素，也能够识别治疗效应。MGLMM 所表示的结构化未测量混杂，有助于理解这种非常具体的未测量混杂如何影响治疗路径的因果效应。

在方法上，我们采用贝叶斯 g-计算算法（GCA），将多元广义线性混合模型作为随时间演变的生成组件，同时随着患者病史随时间累积，实时更新受试者特异的未观测稳定特征。我们的方案通过在 GCA 中纳入潜变量的时变估计，而非像总体平均处理效应那样将其边际化，从而能够评估亚组的治疗路径。此外，该方法为在贝叶斯因果推断中纳入倾向评分（PS）提供了一种途径。现有的倾向评分与结局模型结合方式包括：根据倾向评分指定结局分布、在倾向评分模型与结局模型之间设置共享参数或先验分布，或使用基于后验的逆概率加权或双重稳健估计量（F. Li 等，2022）。我们的方法属于在倾向评分模型与结局模型之间设置共享

参数或先验分布这一类，并通过潜变量之间的协方差，利用多元高斯潜在结构将二者联系起来。最后，我们的方法为评估因果推断中的敏感性分析提供了另一种途径。该方案不再假设不存在未测量混杂，也不再通过事后分析评估未观测混杂导致的偏倚，而是根据敏感性参数的不同假定取值估计因果效应；该参数即未观测治疗分配异质性的方差。未来对该方法的扩展应考虑分类变量和计数结局、多种处理或连续治疗，以及针对患者异质性更灵活的分布假设。

致谢

本研究部分得到了约翰斯·霍普金斯inHealth倡议、硬皮病研究基金会、Nancy and Joachim Bechtle硬皮病精准医学基金、Manugian家族学者项目、Donald B. and Dorothy L. Stabler基金会、Chresanthe Staurulakis纪念基金、Sara and Alex Othon研究基金、NIH P30AR070254以及NIH/美国国立关节炎、肌肉骨骼和皮肤病研究所 K24AR080217的支持。

感谢 Adrianne Woods 为数据库管理提供出色的支持，感谢约翰斯·霍普金斯硬皮病中心研究登记库的患者参与这项奠基性资源的建设，也感谢 Joseph Hogan 博士提出的专业意见。

参考文献

Achana, F., Gallacher, D., Oppong, R., Kim, S., Petrou, S., Mason, J., & Crowther, M. (2021). 基于临床试验的成本效益数据分析：多变量广义线性混合效应模型。《医学决策》, *41*(6), 667–684. Agresti, A., Booth*, J. G., Hobert*, J. P., & Caffo*, B. (2000). 分类响应数据的随机效应建模。《社会学方法论》, *30*(1), 27–80. Allison, P. D., Williams, R., & Moral-Benito, E. (2017). 带固定效应的交叉滞后面板模型的最大似然估计。《*Socius*》, *3*, 2378023117710578. Berger, M. P., & Tan, F. E. (2004). 线性混合效应模型的稳健设计。《皇家统计学会会刊：*C*辑（应用统计）》,*53*(4), 569–581. Bolger, N., & Laurenceau, J.-P. (2013). 密集纵向方法：日记法与经验抽样研究导论。Guilford出版社。 Coley, R. Y., Fisher, A. J., Mamawala, M., Carter, H. B., Pienta, K. J., & Zeger, S. L. (2017). 用于基于多数据源预测潜在健康状态的贝叶斯层级模型及其在前列腺癌主动监测中的应用。《生物统计学》, *73*(2), 625–634. Greenland, S., Pearl, J., & Robins, J. M. (1999). 因果推断中的混杂与可折叠性。《统计科学》, *14*(1), 29–46. Greenland, S., & Robins, J. M. (1986). 可识别性、可交换性与流行病学中的混杂。《国际流行病学杂志》, *15*(3), 413–419. Gunasekara, F. I., Richardson, K., Carter, K., & Blakely, T. (2014). 重复测量数据的固定效应分析。《国际流行病学杂志》,*43*(1), 264–269. He, J., Stephens-Shields, A., & Joffe, M. (2015). 用结构嵌套均值模型估计时变治疗对聚类结局的影响。《国际生物统计学杂志》,*11*(2), 203–222. Heagerty, P. J. (1999). 纵向二元数据的边际指定逻辑正态模型。《生物统计学》,*55*(3), 688–698. Heckman, J. J., & Willis, R. J. (1977). 用于分析已婚女性序贯劳动力参与的贝塔-逻辑斯蒂模型。《政治经济学杂志》, *85*(1), 27–58. Hern´an, M. A., & Robins, J. M. (2010). 因果推断。CRC博卡拉顿，佛罗里达州。 Imai, K., & Kim, I. S. (2019). 在纵向数据因果推断中，何时应使用单位固定效应回归模型？《美国政治科学杂志》,*63*(2), 467–490. Kaufman, J. S. (2008). 评论：我们为何对偏倚存有偏见？《国际流行病学杂志》,*37*(3), 624–626. Kosorok, M. R., & Laber, E. B. (2019). 精准医学。《统计学及其应用年度评论》, *6*, 263–286. Laird, N. M., & Ware, J. H. (1982). 纵向数据的随机效应模型。《生物统计学》, 963–974. Li, D., Iddi, S., Thompson, W. K., Donohue, M. C., & Initiative, A. D. N. (2019). 多队列纵向数据的贝叶斯潜在时间联合混合效应模型。《医学研究统计方法》,*28*(3), 835–845.

Li, F., Ding, P., & Mealli, F.（2022年）。贝叶斯因果推断：一篇批判性综述。英国皇家学会哲学汇刊*A*辑：数学、物理与工程科学。Luger, T. M., Suls, J., & Vander Weg, M. W.（2014年）。吸烟与成年抑郁症之间的关联有多稳健？一项采用线性混合效应模型的荟萃分析。成瘾行为，*39*(10), 1418–1429。Neugebauer, R., van der Laan, M. J., Joffe, M. M., & Tager, I. B.（2007年）。采用历史限制边际结构模型对纵向研究进行因果推断。《电子统计学期刊》，*1*，119。Omair, M. A., Alahmadi, A., & Johnson, S. R.（2015年）。吗替麦考酚酯治疗系统性硬化症的安全性和有效性：一项系统综述。*PLoS One*，*10*(5), e0124205。Qian, T., Klasnja, P., & Murphy, S. A.（2020年）。含内生协变量的线性混合模型：对序贯治疗效应进行建模，并应用于一项移动健康研究。《统计科学：数理统计学会评论期刊》，*35*(3), 375。Raudenbush, S. W., & Bryk, A. S.（2002年）。分层线性模型：应用与数据分析方法 (第1卷)。Sage。Richardson, T. S., & Robins, J. M.（2013年）。单世界干预图（SWIG）：统一反事实方法与因果图形方法。统计与社会科学中心，华盛顿大学系列工作论文，*128*(30), 2013。Robins, J.（1986年）。一种针对持续暴露期死亡率研究的因果推断新方法——应用于控制健康工人幸存者效应。《数学建模》，*7*(9-12), 1393–1512。Robins, J. M., Rotnitzky, A., & Scharfstein, D. O.（2000年）。缺失数据与因果推断模型中的选择偏倚和未测量混杂敏感性分析。载于《流行病学、环境与临床试验中的统计模型》（第1–94页）。施普林格。Rosen, A., Zeger, S. L., et al.（2019年）。精准医学：大规模发现临床相关且具有机制依据的疾病亚组。《临床研究杂志》，*129*(3), 944–945。Rubin, D. B.（1974年）。估计随机研究与非随机研究中的处理因果效应。《教育心理学杂志》，*66*(5), 688。Schnitzer, M. E., Sango, J., Ferreira Guerra, S., & Van der Laan, M. J.（2020年）。在因果推断中采用协作式靶向最小损失估计进行数据自适应纵向模型选择。《生物统计学》，*76*(1), 145–157。Schwartz, J. E., & Stone, A. A.（2007年）。实时瞬时数据分析：实用指南。《实时数据采集科学：健康研究中的自我报告》，76–113。Shardell, M., & Ferrucci, L.（2018年）。利用纵向数据进行因果推断的联合混合效应模型。《医学统计学》，*37*(5), 829–846。Sitlani, C. M., Heagerty, P. J., Blood, E. A., & Tosteson, T. D.（2012年）。用于分析不依从外科试验的纵向结构混合模型。《医学统计学》，*31*(16), 1738–1760。Tashkin, D. P., Roth, M. D., Clements, P. J., Furst, D. E., Khanna, D., Kleerup, E. C., . . . others（2016年）。吗替麦考酚酯与口服环磷酰胺治疗硬皮病相关间质性肺病（SLS II）的比较：一项随机、双盲、平行组对照试验。《柳叶刀·呼吸医学》，*4*(9), 708–719。Van der Laan, M. J., Rose, S., et al.（2011年）。靶向学习：观察性与实验数据的因果推断 （第10卷）。施普林格。

Wang, Z., Bowring, M. G., Rosen, A., Garibaldi, B., Zeger, S., & Nishimura, A.（2022年）。基于动态模型进行学习和预测，以监测 COVID-19 患者。《统计科学》，*37*(2), 251–265。Yang, S., & Lok, J. J.（2018年）。粗略结构嵌套均值模型中未测量混杂的敏感性分析。华人统计学，*28*(4), 1703。Zamora, A. C., Wolters, P. J., Collard, H. R., Connolly, M. K., Elicker, B. M., Webb, W. R., . . . Golden, J. A.（2008年）。使用吗替麦考酚酯治疗硬皮病相关间质性肺疾病。呼吸医学，*102*(1), 150–155。Zeger, S. L., & Karim, M. R.（1991年）。含随机效应的广义线性模型：一种吉布斯抽样方法。美国统计学会杂志，*86*(413), 79–86。Zeger, S. L., Liang, K.-Y., & Albert, P. S.（1988年）。纵向数据模型：一种广义估计方程方法。《生物统计学》，1049–1060。Zhou, T., Elliott, M. R., & Little, R. J.（2019年）。用于治疗比较的倾向评分惩罚样条法。美国统计学会杂志，*114*(525), 1–19。

图

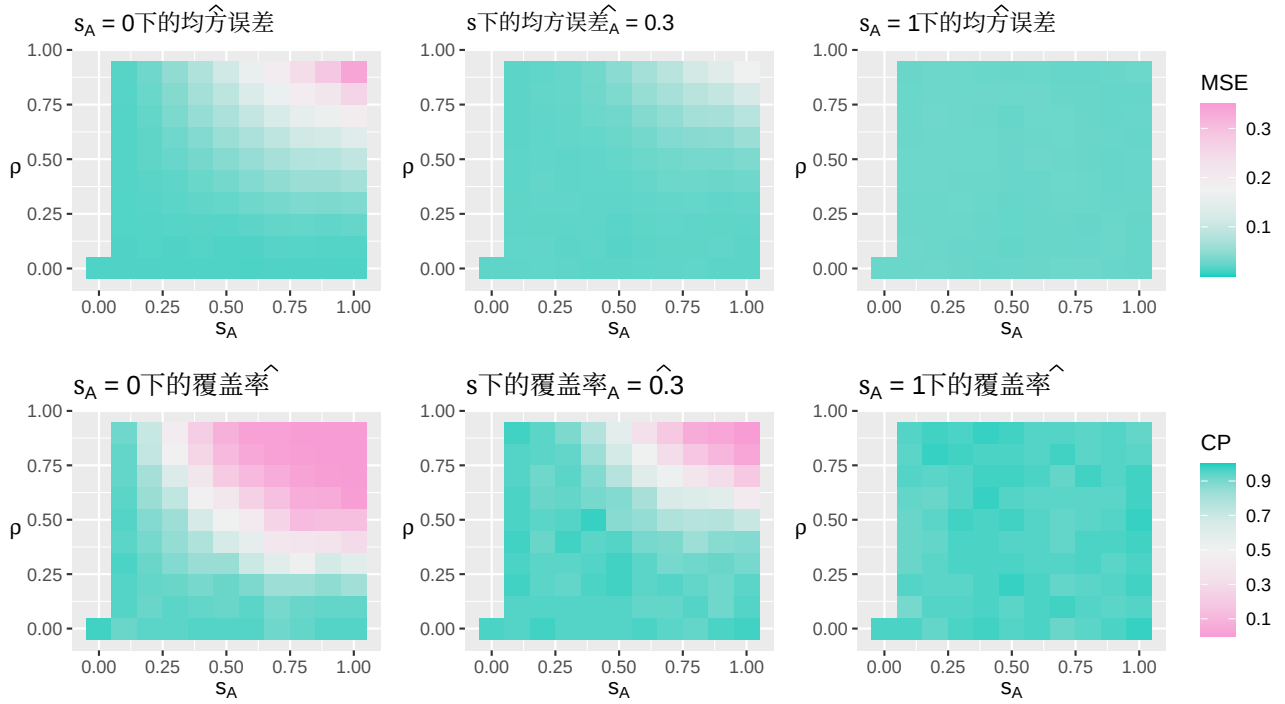

图3：在第二个时间点的真实治疗效应为 1.2 时，该图展示了不同模拟真实值（s_A，ρ）和假设模型参数 $\hat{s}_A$ 下混合平均处理效应的均方误差（MSE）和后验覆盖率。绿色表示估计效果较好，例如均方误差较低、覆盖概率较高。

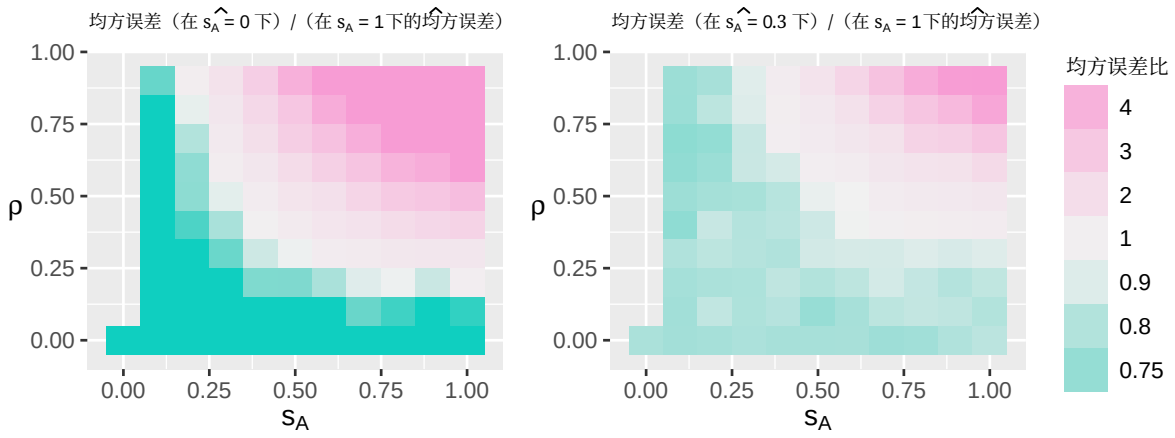

图4：第二个时间点的真实治疗效应为1.2时的均方误差比。

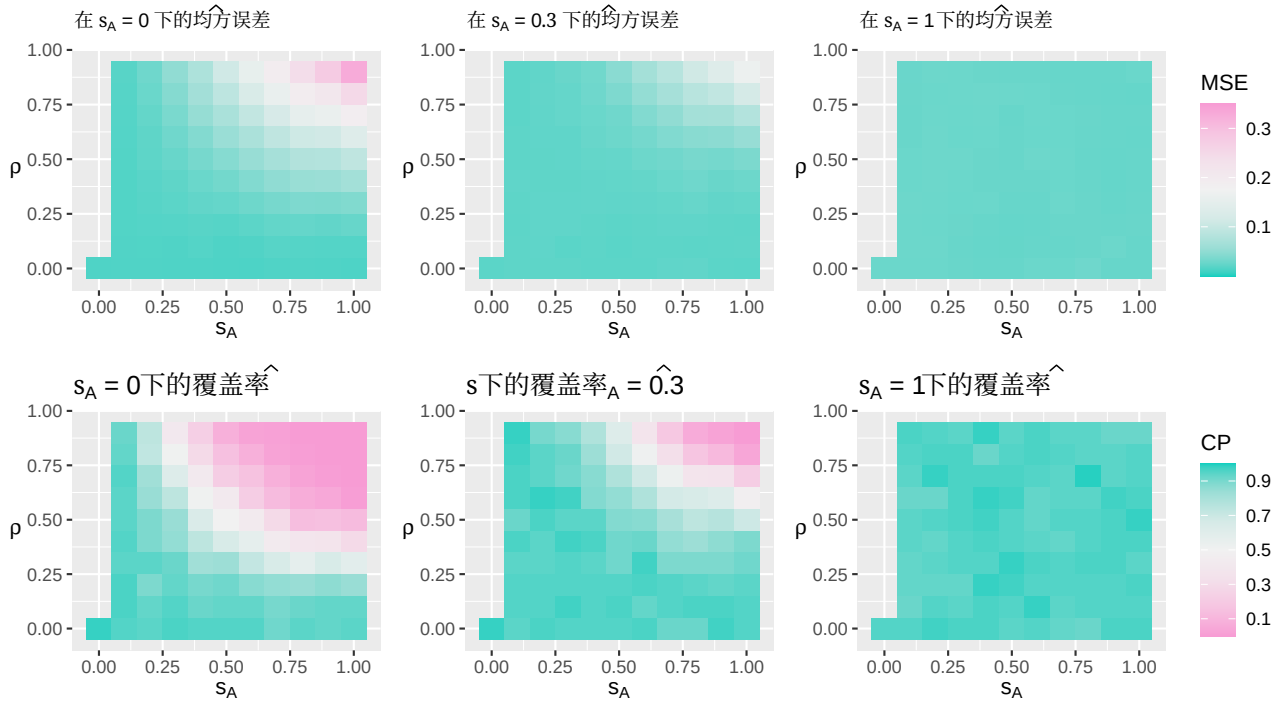

图5：第二个时间点的真实治疗效应为0时，该图展示了不同模拟真实值（s_A，ρ）和假设模型参数 $\hat{s}_A$下混合平均处理效应的均方误差（MSE）和后验覆盖率。绿色表示估计效果更好，例如均方误差更低、覆盖概率更高。

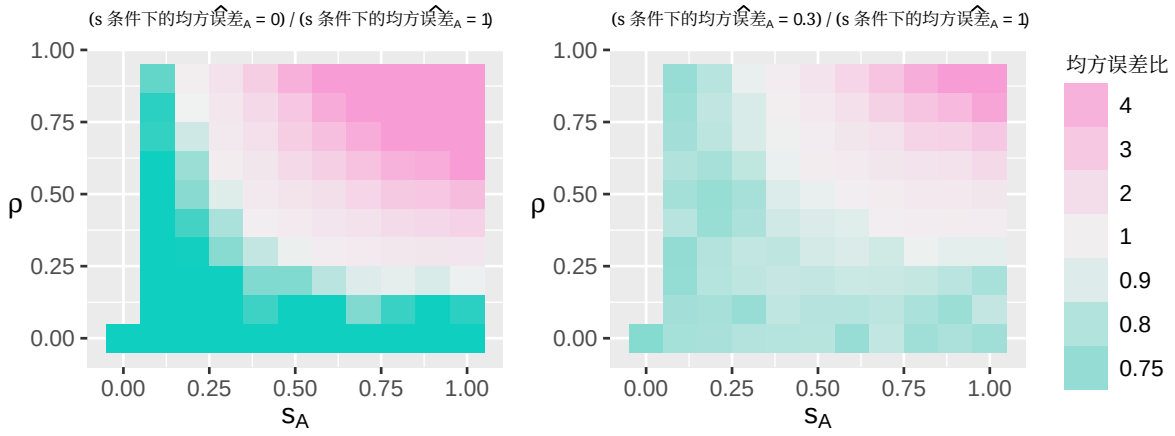

图6：无治疗效应时的均方误差比。

附录

A G公式的识别

为简化起见，我们忽略用于受试者索引的下标 i。假设 $A_{0:h} = a_{0:h}(q)$ 和时间不变的潜在治疗异质性 $b_i^A = b_i^A$，则在截至时间 h 的已观测信息条件下，未来 τ 个时间间隔内反事实轨迹的分布可按如下方式处理。

$$P(\overline{Y}_{(h+1):(h+\tau)}(q), \overline{M}_{(h+1):(h+\tau)}(q) | V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A)$$

by positivity and exchangeability,

$$= P(\overline{Y}_{(h+1):(h+\tau)}(q), \overline{M}_{(h+1):(h+\tau)}(q) | V, \overline{A}_{0:h}, A_{h+1} = a_{h+1}(q), \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A)$$

by consistency,

$$= P(Y_{h+1}, M_{h+1} | V, \overline{A}_{0:(h+1)} = \overline{a}_{0:(h+1)}(q), \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A)$$

$$P(\overline{Y}_{(h+2):(h+\tau)}(q), \overline{M}_{(h+2):(h+\tau)}(q) | V, \overline{A}_{0:(h+2)} = \overline{a}_{0:(h+2)}(q), \overline{Y}_{0:(h+1)}, \overline{M}_{0:(h+1)}, b_i^A)$$

by induction,

$$= \prod_{s=h}^{h+\tau-1} P(Y_{s+1}, M_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A)$$

account for and marginalize over patient heterogeneity,

$$= \prod_{s=h}^{h+\tau-1} \int_{u_s} \int_{v_s} P(Y_{s+1}, M_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^Y = u_s, b^M = v_s, b_i^A)$$
$$P(b^Y = u_s, b^M = v_s | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A) du_s dv_s$$

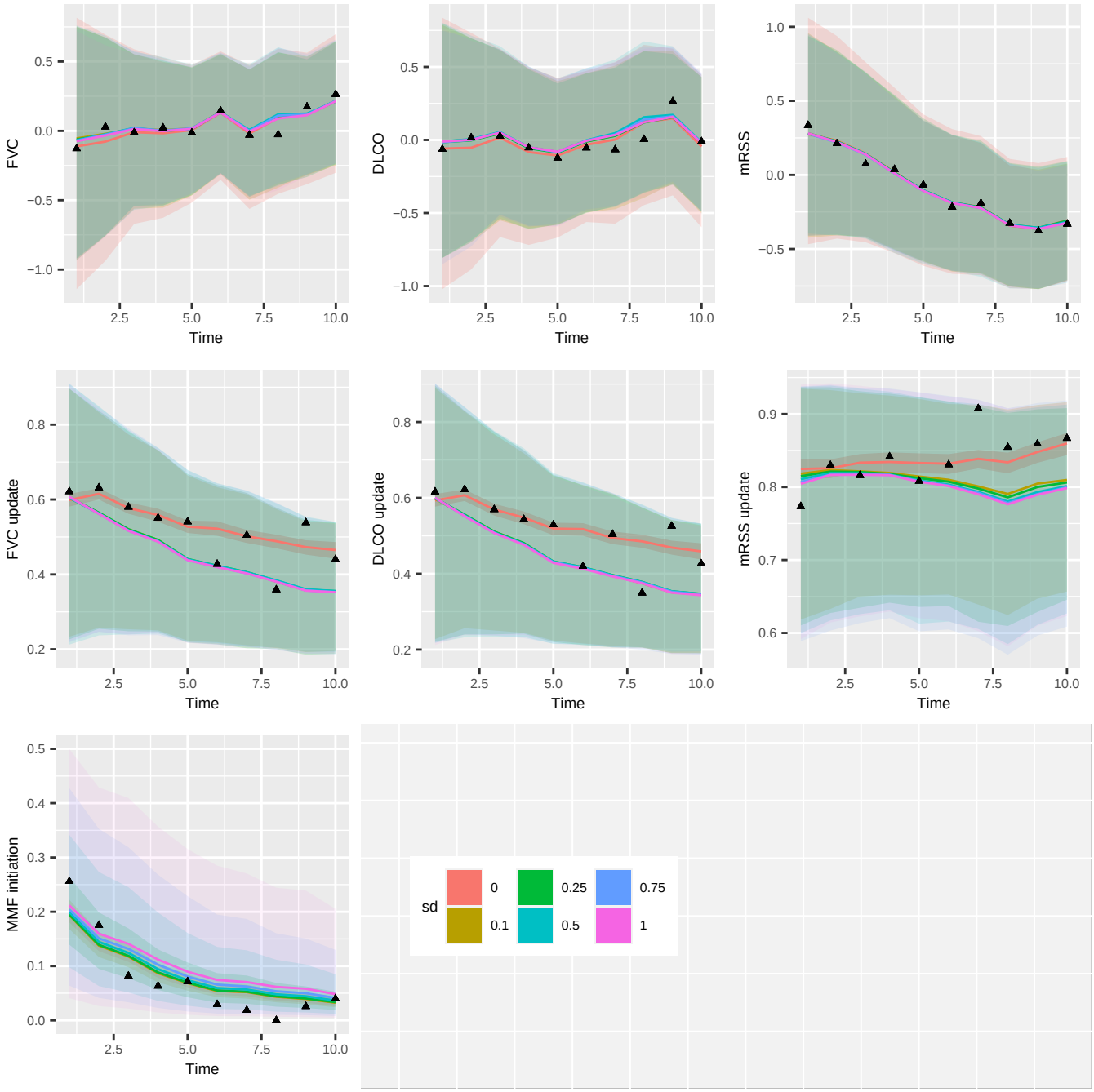
FVC
DLCO
mRSS
FVC update
DLCO update
mRSS update
MMF initiation
Time
sd
0
0.1
0.25
0.5
0.75
1

图7：应用准确率图。

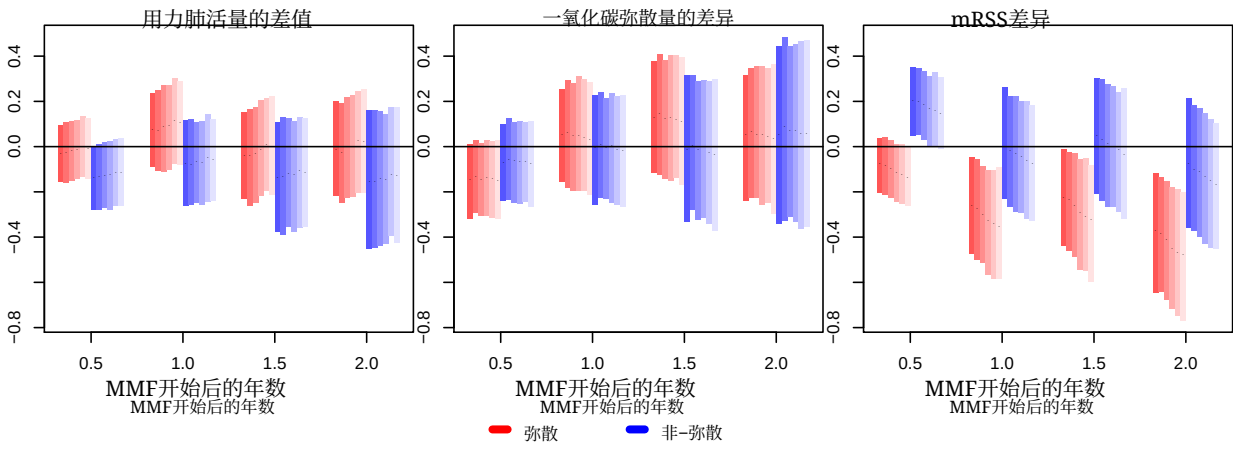

图8：按弥散型与非弥散型亚组应用因果估计。

be导致反事实治疗路径无法为异质性估计提供信息，

$$= \prod_{s=h}^{h+\tau-1} \int_{u_s} \int_{v_s} P(Y_{s+1}, M_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^Y = u_s, b^M = v_s, b_i^A)$$
$$P(b^Y = u_s, b^M = v_s | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A) du_s dv_s$$

根据图1所示的分布假设，

$$= \prod_{s=h}^{h+\tau-1} \int_{u_s} \int_{v_s} P(Y_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^Y = u_s)$$
$$P(M_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^M = v_s)$$
$$P(b^Y = u_s, b^M = v_s | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A) du_s dv_s$$

将多元广义线性混合模型参数化为线性模型，我们得到

$$= \prod_{s=h}^{h+\tau-1} \int_{u_s} \int_{v_s} P(Y_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^Y = u_s; \beta^Y, \sigma^2)$$
$$P(M_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^M = v_s; \beta^M)$$
$$P(b^Y = u_s, b^M = v_s | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A; G) du_s dv_s$$

给定 b_i^A，条件亚组平均处理效应的 g公式定义为：在用户指定的方案 q 下，随访结束时（时间 $h+\tau$）潜在结果的条件均值。由此可推导如下，

$$\mathbb{E}(Y_{h+\tau}(q) | V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A)$$
$$= \int_{y_\tau} y_\tau P(Y_{h+\tau}(q) = y_\tau | V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A) dy_\tau$$

$$
\begin{aligned}
&= \int_{y_{h+\tau}} \int_{m_{h+\tau}} \cdots \int_{y_{h+1}} \int_{m_{h+1}} \\
&\quad y_\tau P(\overline{Y}_{(h+1):(h+\tau)}(q) = \overline{y}_{(h+1):(h+\tau)}, \overline{M}_{(h+1):(h+\tau)}(q) = \overline{m}_{(h+1):(h+\tau)} | V, \overline{A}_{0:h}, \overline{Y}_{0:h}, \overline{M}_{0:h}, b_i^A) \\
&\qquad\qquad dm_{h+1} dy_{h+1} \ldots dm_{h+\tau} dy_{h+\tau} \\
&= \int_{y_{h+\tau}} \int_{m_{h+\tau}} \cdots \int_{y_{h+1}} \int_{m_{h+1}} y_\tau \Big\{ \prod_{s=0}^{\tau-1} \int_{u_s} \int_{v_s} \\
&\qquad P(Y_{s+1} = y_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^Y = u_s; \beta^Y, \sigma^2) \\
&\qquad P(M_{s+1} = m_{s+1} | V, \overline{A}_{0:(s+1)} = \overline{a}_{0:(s+1)}(q), \overline{Y}_{0:s}, \overline{M}_{0:s}, b^M = v_s; \beta^M) \\
&\qquad P(b^Y = u_s, b^M = v_s | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A; G) du_s dv_s \Big\} \\
&\qquad\qquad dm_{h+1} dy_{h+1} \ldots dm_{h+\tau} dy_{h+\tau} \\
&= \int_{y_{h+\tau}} \int_{m_{h+\tau}} \int_{u_{\tau-1}} \int_{v_{\tau-1}} \cdots \int_{y_{h+1}} \int_{m_{h+1}} \int_{u_0} \int_{v_0} \\
&\qquad y_\tau \Big\{ \prod_{s=0}^{\tau-1} P(Y_{s+1} = y_{s+1} | V, \overline{A}_{s+1} = \overline{a}_{s+1}(q), \overline{Y}_s = \overline{y}_s, \overline{M}_s = \overline{m}_s, b^Y = u_s; \beta^Y, \sigma^2) \\
&\qquad P(M_{s+1} = m_{s+1} | V, \overline{A}_{s+1} = \overline{a}_{s+1}(q), \overline{Y}_s = \overline{y}_s, \overline{M}_s = \overline{m}_s, b^M = v_s; \beta^M) \\
&\qquad P(b^Y = u_s, b^M = v_s | V, \overline{A}_{0:h}, \overline{Y}_{0:s}, \overline{M}_{0:s}, b_i^A; G) \Big\} \\
&\qquad\qquad du_0 dv_0 dm_{h+1} dy_{h+1} \ldots du_{\tau-1} dv_{\tau-1} dm_{h+\tau} dy_{h+\tau}
\end{aligned}
$$

对目标人群中观测到的临床病史分布进一步积分，即可得到以 b_i^A 为条件的总体平均处理效应，

$$
\begin{aligned}
\mathbb{E}(Y_{h+\tau}(q) | b_i^A) = \int_v \int_{y_h} \int_{m_h} \cdots \int_{y_0} \int_{m_0} & \mathbb{E}(Y_{h+\tau}(q) | V = v, \overline{A}_{0:h} = \overline{a}_{0:h}, \overline{Y}_{0:h} = \overline{y}_{0:h}, \overline{M}_{0:h} = \overline{m}_{0:h}, b_i^A) \\
& P(V = v, \overline{A}_{0:h} = \overline{a}_{0:h}, \overline{Y}_{0:h} = \overline{y}_{0:h}, \overline{M}_{0:h} = \overline{m}_{0:h}) dm_0 dy_0 \ldots dm_h dy_h dv
\end{aligned}
$$

条件平均处理效应的计算方法如下：将可观测变量的目标人群分布替换为相应的经验分布，$P(V = v,\ A_{0:h} = a_{0:h},\ Y_{0:\ h} = y_{0:h},\ M_{0:h} = m_{0:h})$。

$$
\begin{aligned}
\widehat{\mathbb{E}}(Y_{h+\tau}(q) | b_i^A) = \int_v \int_{y_h} \int_{m_h} \cdots \int_{y_0} \int_{m_0} & \mathbb{E}(Y_{h+\tau}(q) | V = v, \overline{A}_{0:h} = \overline{a}_{0:h}, \overline{Y}_{0:h} = \overline{y}_{0:h}, \overline{M}_{0:h} = \overline{m}_{0:h}, b_i^A) \\
& \widehat{P}(V = v, \overline{A}_{0:h} = \overline{a}_{0:h}, \overline{Y}_{0:h} = \overline{y}_{0:h}, \overline{M}_{0:h} = \overline{m}_{0:h}) dm_0 dy_0 \ldots dm_h dy_h dv.
\end{aligned}
$$

假设目标总体中的治疗分配异质性 b_i^A 在边际上服从 $N(0, v)$。因此，边际混合总体平均处理效应可通过对 b_i^A 按其分布积分 $P(b_i^A = w)$ 得到，表示为 $\widehat{\mathbb{E}}(Y_{h+\tau}(q)) = \int_w \widehat{\mathbb{E}}(Y_{h+\tau}(q) | b_i^A = w) P(b_i^A = w) dw$。

B 随机效应的序贯更新

不失一般性，假设连续变量和二元变量分别服从高斯分布并采用 logit 模型，则结构模型可写为:

$$
\begin{aligned}
&Y_{it}|M_{it}=1 \sim \eta_{it}^{Y}+\sigma\psi^{Y},\\
&P(M_{it}=1)=\frac{\exp(\eta_{it}^{M})}{1+\exp(\eta_{it}^{M})},\\
&P(A_{it}=1|A_{i,t-1}=0)=\frac{\exp(\eta_{it}^{A})}{1+\exp(\eta_{it}^{A})},
\end{aligned}
$$

其中 $\psi^{Y}\sim N(0,1)$、 $\eta_{it}^{Y}=\eta^{Y}(\mathcal{F}_{it},b_i^{Y};\,\theta^{Y})$、 $\eta_{it}^{M}=\eta^{M}(\mathcal{F}_{it},b_i^{M};\,\theta^{M})$ 和 $\eta_{it}^{A}=\eta^{A}(\mathcal{F}_{it}^{A},b_i^{A};\,\theta^{A})$。

对于每个个体，在给定截至时间 t 的生物标志物动态以及截至时间 h观测到的治疗序列的条件下，实施随机效应的序贯更新，其中 $h\le t$。对于受试者 i的观测轨迹，联合似然为

$$
\begin{aligned}
&P(\overline{Y}_{i,0:t},\overline{M}_{i,0:t},\overline{A}_{i,0:h}|b_i,\beta,\sigma)\\
&\propto\prod_{j=1}^{t}\left[\left(\frac{1}{\sigma}\exp\{-\frac{1}{2\sigma^2}(Y_{ij}-\eta_{ij}^{Y})^2\}\right)^{M_{ij}}\frac{\exp\{\eta_{ij}^{M}M_{ij}\}}{1+\exp(\eta_{ij}^{M})}\right]\times\prod_{j'=1}^{h}\left[\frac{\exp\{\eta_{ij'}^{A}A_{ij'}\}}{1+\exp(\eta_{ij'}^{A})}\right]^{\mathbb{1}(j'\le s_i)},
\end{aligned}
$$

其中 s_i 是受试者 i的观测治疗启动时间，随机效应 b_i 的先验分布为

$$
P(b_i|G)\propto|G|^{-1/2}\exp(-\frac{1}{2}b_i^{T}G^{-1}b_i).
$$

b_i 的对数后验可写为

$$
\begin{aligned}
&\log P(b_i|\overline{Y}_{i,0:t},\overline{M}_{i,0:t},\overline{A}_{i,0:h},\beta,\sigma,G)\\
&\propto-\frac{1}{2}b_i^{T}G^{-1}b_i+\sum_{j=1}^{t}\left\{-\frac{M_{ij}}{2\sigma^2}(Y_{ij}-\eta_{ij}^{Y})^2+\eta_{ij}^{M}M_{ij}-\log[1+\exp(\eta_{ij}^{M})]\right\}\\
&\quad+\sum_{j'=1}^{min(h,s_i)}\left\{\eta_{ij'}^{A}A_{ij'}-\log[1+\exp(\eta_{ij'}^{A})]\right\}.
\end{aligned}
$$

使用构造抽样链的算法（如马尔可夫链蒙特卡罗）对 b_i 进行抽样，由于计算个体反事实轨迹的复杂性，将消耗大量计算资源。为便于后验抽样，我们考虑 b_i 的后验分布的拉普拉斯近似。近似分布的均值通过求解下列关于后验众数 $\hat{}\,b_i=(\hat{}\,b_i^{Y},\ \hat{}\,b_i^{M},\ \hat{}\,b_i^{A})$ 的方程得到，

$$
\frac{\partial}{\partial b_i}\log P(b_i|\overline{Y}_{i,0:t},\overline{M}_{i,0:t},\overline{A}_{i,0:h},\beta,\sigma,G)\Big|_{b_i=\hat{b}_i}=0,
$$

其中

$$\frac{\partial}{\partial b_i}\log P(b_i|\overline{Y}_{i,0:t},\overline{M}_{i,0:t},\overline{A}_{i,0:h},\beta,\sigma,G)=-G^{-1}b_i+\begin{pmatrix}\sum_{j=1}^{t}\frac{M_{ij}}{\sigma_1^2}(Y_{ij}-\eta_{ij}^{Y})\\ \sum_{j=1}^{t}M_{ij}-\frac{\exp(\eta_{ij}^{M})}{1+\exp(\eta_{ij}^{M})}\\ \sum_{j=1}^{min(h,s_i)}A_{ij}-\frac{\exp(\eta_{ij}^{A})}{1+\exp(\eta_{ij}^{A})}\end{pmatrix}.$$

$\hat{}$ 近似分布的方差是 b_i 的渐近方差，它是观测费舍尔信息矩阵的逆，定义如下

$$V=\Big[-\frac{\partial^2}{\partial b_i\partial b_i^T}\log P(b_i|\overline{Y}_{i,0:t},\overline{M}_{i,0:t},\overline{A}_{i,0:h},\beta,\sigma,G)\Big|_{b_i=\hat{b}_i}\Big]^{-1},$$

其中

$$\begin{aligned}&\frac{\partial^2}{\partial b_i\partial b_i^T}\log P(b_i|\overline{Y}_{i,0:t},\overline{M}_{i,0:t},\overline{A}_{i,0:h},\beta,\sigma,G)\\ =&-G^{-1}-\text{diag}\Big\{\frac{1}{\sigma^2}\sum_{j=1}^{t}M_{ij},\sum_{j=1}^{t}\frac{\exp(\eta_{ij}^{M})}{[1+\exp(\eta_{ij}^{M})]^2},\sum_{j=1}^{min(h,s_i)}\frac{\exp(\eta_{ij}^{A})}{[1+\exp(\eta_{ij}^{A})]^2}\Big\}.\end{aligned}$$

因此，后验分布 $P(b_i|Y_{i,0:t},M_{i,0:t},A_{i,0:h},\beta,\sigma,G)$ 的近似为多元高斯分布 $MNV(b_i,V)$。

反事实轨迹的序贯更新也以 b_i^A 为常数为条件，即 $b_i^A=c$。我们按如下方式，在历史 $(Y_{i,0:t},M_{i,0:t},A_{i,0:h})$、总体水平估计 (β,σ,G) 和 $b_i^A=c$ 的条件下，对生物标志物动态的异质性进行序贯更新

$$(b_i^Y,b_i^M|b_i^A=c)\sim MNV(b_{\cdot|A},V_{\cdot|A})$$

使得

$$b_{\cdot|A}=\begin{pmatrix}\hat{b}^Y\\ \hat{b}^M\end{pmatrix}+\begin{pmatrix}V^{Y,A}\\ V^{M,A}\end{pmatrix}(V^A)^{-1}(c-\hat{b}^A)\tag{9}$$

$$V_{\cdot|A}=\begin{pmatrix}V^Y & V^{Y,M}\\ & V^M\end{pmatrix}-\begin{pmatrix}V^{Y,A}\\ V^{M,A}\end{pmatrix}(V^A)^{-1}(V^{Y,A},V^{M,A}).\tag{10}$$

假设我们要模拟患者纵向测量的反事实进展，并将治疗序列固定为 $a_{0:t}^q$，且采用治疗方案 q；截至时间 h 的序列为观察到的治疗，即 $A_{i,\,0:h}=a_{0:h}^q$。若将 G^{-1} 的第三行写为（C_1，C_2，C_3），则相对于 b_i^A 的导数项为

$$\sum_{j=1}^{min(h,s_i)}a_j-\frac{\exp(X_{ij}\beta^A+b_{i0}^A)}{1+\exp(X_{ij}\beta^A+b_{i0}^A)}=C_1b_i^A+C_2b_i^M+C_3b_i^Y,\tag{11}$$

我们可以看到，反事实治疗序列 $a_{(h+1):t}$ 的设定不会影响 $\hat{}\, b_i$ 的估计。请注意， $\sum^{min(h,s_{i_}}$ $_{j=1}$ a_j 只能是 0 或 1，因为求和会在治疗启动时停止。在应用中，我们着重研究此前未接受治疗者在时间 h 开始治疗的效果，即 $h < s_i$ 和 $\sum^{min(h,s_{i_}}$ $_{j=1}$ $a_j = 0$。因此，对于 $\hat{}b_i$ 的估计，公式（11）施加条件 $- \exp(X_{ij}\beta^A + b_{i0}^A)$ $1+\exp(X_{ij}\beta^A + b_{i0}^A) = C_1 b_i^A + C_2 b_i^M + C_3 b_i^Y$，仅使用治疗启动前的治疗信息。

C 生成反事实轨迹的伪代码

Algorithm for Dynamic Projection of Counterfactual Trajectories under MGLMM

Conditional on:

(a) observed history up to time h, $(V_i, \overline{Y}_{i,0:h}, \overline{M}_{i,0:h}, \overline{A}_{i,0:h})$

(b) posteriors of $(\theta^Y, \theta^M, \theta^A, G)$

(c) $var(b_i^A) = v$,

Goal: make posterior predictive inference of $(\overline{Y}_{(h+1):T}(q), \overline{M}_{(h+1):T}(q))$ under regime q.

Step 0: Initialization

(a) draw subject-specific stochastic matrices $\psi^Y, \psi^M \in \mathbb{R}^{N_{post}\times(T-h)}$, $\psi^Y \sim \mathcal{N}(0,1)$ and $\psi^M \sim \mathcal{U}(0,1)$

(b) $\mathcal{F}_{i,h+1}^{(\ell)}(q) = (V_i, \overline{Y}_{i,0:h}, \overline{M}_{i,0:h}, \overline{A}_{i,0:h}, a_{h+1}(q))$ for all ℓ

(c) for each ℓ, draw $b_i^{A(\ell)} \sim f(b^A|V_i, \overline{Y}_{i,0:h}, \overline{M}_{i,0:h}, \overline{A}_{i,0:h}; \theta^{Y(\ell)}, \theta^{M(\ell)}, \theta^{A(\ell)}, G^{(\ell)})$ if $h > 0$,
otherwise draw $b_i^{A(\ell)} \sim N(0, v)$

(d) $l = 0$

while $\ell < N_{post}$ **do**

for $t \in h+1, \ldots, T$ **do**

Step 1: Calculate $(\hat{b}_i^{(\ell)}(q), V_i^{(\ell)}(q))$ conditional on $(V_i, \overline{Y}_{i,0:(t-1)}^{(\ell)}(q), \overline{M}_{i,0:(t-1)}^{(\ell)}(q), \overline{A}_{i,0:h})$

Step 2: Draw $(b_i^{Y(\ell)}(q), b_i^{M(\ell)}(q))|b_i^{A(\ell)} \sim MVN(b_{t|A}^{(\ell)}(q), V_{t|A}^{(\ell)}(q))$, where
$b_{t|A}^{(\ell)}(q)$ and $V_{t|A}^{(\ell)}(q)$ are obtained by (9) and (10), respectively.

Step 3: Update $M_{it}^{(\ell)}(q)$
let $p_{it}^{(\ell)}(q) = \text{logit}^{-1}\eta^M(\mathcal{F}_{it}^{(\ell)}(q), b_i^{M(\ell)}(q); \theta^{M(\ell)})$
draw $M_{it}^{(\ell)}(q) \sim \text{Bernoulli}(p_{it}^{(\ell)}(q))$ by setting $M_{it}^{(\ell)}(q) = \mathbb{1}\{\psi_{\ell,t-h}^M \le p_{it}^{(\ell)}(q)\}$

Step 4: Update $Y_{it}^{(\ell)}(q)$
draw $Y_{it}^{(\ell)}(q) \sim f_Y(\eta_{it}^{Y(\ell)}(q), (\sigma^{(\ell)})^2)$ by
setting $\eta_{it}^{Y(\ell)}(q) = \eta^Y(\mathcal{F}_{it}^{(\ell)}(q), b_i^{Y(\ell)}(q); \theta^{Y(\ell)})$ and $Y_{it}^{(\ell)}(q) = \eta_{it}^{Y(\ell)}(q) + \sigma^{(\ell)}\psi_{\ell,t-h}^Y$

Step 5: Define

$$\mathcal{F}_{i,t+1}^{(\ell)}(q) = (V_i, \overline{Y}_{i,0:t}^{(\ell)}(q), \overline{M}_{i,0:t}^{(\ell)}(q), \overline{A}_{i,0:(t+1)}(q)),$$

where

$$\begin{aligned}\overline{Y}_{i,0:t}^{(\ell)}(q) &= (\overline{Y}_{i,0:h}, \overline{Y}_{i,(h+1):t}^{(\ell)}(q))\\ \overline{M}_{i,0:t}^{(\ell)}(q) &= (\overline{M}_{i,0:h}, \overline{M}_{i,(h+1):t}^{(\ell)}(q))\\ \overline{A}_{i,0:(t+1)}(q) &= \overline{a}_{0:(t+1)}(q)\end{aligned}$$

and the observed equals the counterfactual during the given history, i.e. $\overline{A}_{i,0:h} = \overline{a}_{0:h}(q)$.

end for

end while

Step 6: $\{\overline{Y}_{i,(h+1):T}^{(\ell)}(q), \overline{M}_{i,(h+1):T}^{(\ell)}(q); \ell = 1, \ldots, N_{post})\}$ are samples from the posterior predictive distribution of $(\overline{Y}_{i,(h+1):T}(q), \overline{M}_{i,(h+1):T}(q))$ under regime q.

D G-formulain 模拟

$$
\begin{aligned}
\mathbb{E}[Y_{i2}(q)] &= \iint y_2 P(Y_{i1}(q)=y_1, Y_{i2}(q)=y_2|V=v)P(V=v)dvdy_1dy_2 \\
&= \iint y_2 P(Y_{i1}(q)=y_1, Y_{i2}(q)=y_2|V=v, b_i^A=w)P(b_i^A=w)P(V=v)dwdvdy_1dy_2 \\
&= \iint y_2 P(Y_{i1}(q)=y_1, Y_{i2}(q)=y_2|A_{i1}=a_1, V=v, b_i^A=w)P(b_i^A=w)P(V=v)dwdvdy_1dy_2 \\
&= \iint y_2 P(Y_{i1}=y_1, Y_{i2}(q)=y_2|A_{i1}=a_1, V=v, b_i^A=w)P(b_i^A=w)P(V=v)dwdvdy_1dy_2 \\
&= \iint y_2 P(Y_{i2}(q)=y_2|A_{i1}=a_1, Y_{i1}=y_1, V=v, b_i^A=w) \\
&\qquad P(Y_{i1}=y_1|A_{i1}=a_1, V=v, b_i^A=w)P(b_i^A=w)P(V=v)dwdvdy_1dy_2 \\
&= \iint y_2 P(Y_{i2}(q)=y_2|A_{i1}=a_1, A_{i2}=a_2, Y_{i1}=y_1, V=v, b_i^A=w) \\
&\qquad P(Y_{i1}=y_1|A_{i1}=a_1, V=v, b_i^A=w)P(b_i^A=w)P(V=v)dwdvdy_1dy_2 \\
&= \iint y_2 P(Y_{i2}=y_2|A_{i1}=a_1, A_{i2}=a_2, Y_{i1}=y_1, V=v, b_i^A=w) \\
&\qquad P(Y_{i1}=y_1|A_{i1}=a_1, V=v, b_i^A=w)P(b_i^A=w)P(V=v)dwdvdy_1dy_2 \\
&= \mathbb{E}(Y_{i2}|A_{i1}=a_1, A_{i2}=a_2).
\end{aligned}
$$

补充材料

与结构嵌套模型的联系

Sitlani 等（2012）和 Qian 等（2020）将结构嵌套模型（SNM）中的瞬时治疗效应视为“瞬时效应”，采用线性混合模型作为结构模型，并比较仅在特定时间 m 的治疗状态不同的治疗路径，即在二元单调治疗的情况下，比较 $A_m = 1$ 与 $A_m = 0$。另一方面，我们提出的方法比较不同治疗方案下治疗路径的效应，即比较 $A_{0:t}$ 为 $a_{0:t}$（q_1）与 $a_{0:t}$（q_2），其中 q_1 和 q_2 是感兴趣的治疗方案。示例应用研究的是持续随时间服用药物的治疗效应；由于因果效应会随时间累积，因此需要采用与结构模型方法根本不同的表征方式。在我们的框架下，瞬时治疗效应（即瞬时效应）可表征为比较 q_1 与 q_2 的平均因果效应，其中 a_t（q_1）$= a_t$（q_2），对于 $t \neq m$，a_m（q_1）$= 1$，且 a_m（q_2）$= 0$。具体而言，假设 ϕ_A（$A_{i,\ 0:t}$）$= A_{it}$，ϕ_4^Y（$\mathcal{H}_{it}$）$= 0$，$\tau = 1$，并针对连续型结局采用线性混合模型，就得到 Qian 等（2020）中的一个特例；此时，我们将得到在给定截至时间 h 的信息条件下、时间 $h+1$ 的亚组瞬时治疗效应。

—

$$\mathbb{E}(Y_{i,h+1}|V_i, A_{i,h+1} = 1, \overline{A}_{i,0:h}, \mathcal{H}_{ih}) - \mathbb{E}(Y_{i,h+1}|V_i, A_{i,h+1} = 0, \overline{A}_{i,0:h}, \mathcal{H}_{ih}) = \phi_2^Y(H_{it})\beta_2^Y. \quad (12)$$

因此，在本例中，模型参数 β_2^Y 对由（V_i，$A_{i,\ 0:h}$，$\mathcal{H}_{ih}$）定义的亚组而言，在边际层面具有因果解释，而 MGLMM 则简化为线性结构混合模型。然而，当 $\phi_4^Y(\mathcal{H}_{it}) \neq 0$时，等式 (12) 不再成立，因为 β_2^Y 仅在给定b_i^Y的条件下仍具有因果解释，如下文的条件亚组因果效应所示，

$$\begin{aligned}\mathbb{E}(Y_{i,h+1}|V_i, A_{i,h+1} = 1, \overline{A}_{i,0:h}, \mathcal{H}_{ih}, b_i^Y) - \mathbb{E}(Y_{i,h+1}|V_i, A_{i,h+1} =& 0, \overline{A}_{i,0:h}, \mathcal{H}_{ih}, b_i^Y) \\ =& \phi_2^Y(H_{it})\beta_2^Y + \phi_4^Y(H_{it})b_{i1}^Y, \quad (13)\end{aligned}$$

且条件期望 $\mathbb{E}(b_i^Y|V_i,\ A_{i,\ 0:h},\ \mathcal{H}_{ih})$ 不一定为零。

B 与 Shardell 和 Ferrucci（2018年）的联系

Shardell 和 Ferrucci（2018年）通过联合混合效应模型展示了纵向因果推断，假设结局、混杂因素和治疗分配各模型组件之间共享随机效应，

他们的模型设定与第2节中的多元广义线性混合模型相似，即采用$b_{i0}^A = b_{i0}^M = (b_{i0}^Y,\ b_{i1}^Y)$和 $\phi_4^M \equiv 0$，但有一个显著区别： ϕ_2^A和 ϕ_3^M是总体层面系数，而非观测变量。Shardell 和 Ferrucci（2018年）假设，在结局进程中未观测异质性的条件下满足序贯可交换性，即$(b_{i0}^Y,\ b_{i1}^Y)$；并假设该异质性与混杂因素及治疗分配中的异质性成比例。相比之下，我们通过随机效应 b_i^A，在治疗分配中考虑未观测的时不变特征，并假设该随机效应与$(b_{i0}^Y,\ b_{i1}^Y)$相关，且在b_i^A的条件下满足序贯可交换性，而非在$(b_{i0}^Y,\ b_{i1}^Y)$的条件下。

Shardell 和 Ferrucci（2018年）模型中不存在未测量混杂因素这一假设意味着不存在治疗分配异质性。虽然在多元广义线性混合模型下假设不存在治疗异质性，相当于将 $v = 0$ 设为零；这是不存在未测量混杂因素的充分但非必要条件。在多元广义线性混合模型中，$\text{cov}(b_i^A,\ b_i^M) = \text{cov}(b_i^A,\ b_i^Y) = 0$ 意味着不存在未测量混杂因素。也就是说，即使不存在未测量混杂因素这一条件成立，只要治疗分配异质性与生物标志物动态中未观测到的异质性 ($b_i^Y,\ b_i^M$) 不相关，多元广义线性混合模型仍允许存在治疗分配异质性；这类不造成混杂的治疗分配异质性例如包括患者基于个人信念或社会污名而对某种治疗的偏好。另一方面，我们注意到，在给定 v（b_i^A 的假定方差）的情况下，协方差 $\text{cov}(b_i^A,\ b_i^M)$ 和 $\text{cov}(b_i^A,\ b_i^Y)$ 可以估计。因此，我们的方法确实能根据多元广义线性混合模型中的估计协方差，部分揭示未测量混杂因素可能存在与否。

当不存在治疗分配异质性时，Shardell 和 Ferrucci（2018）的方法与我们的方法都会简化为标准 g-计算：仅使用广义线性混合效应模型拟合结局与混杂因素模型，因为此时分配机制变得可忽略（F. Li 等，2022）。考虑一个简化情景：在假设不存在时变混杂因素和治疗分配异质性的情况下，考察给定截至时间 h 的历史信息时，时间$h+1$ 的亚组平均处理效应。Shardell 和 Ferrucci（2018）的方法无法识别该亚组平均处理效应。原因如下：根据他们的条件序贯可交换性假设

$$Y_{h+1}(q) \perp A_{h+1} | V, \overline{A}_{0:h}, \overline{Y}_{0:h}, b_i^Y,$$

我们可以直接识别条件反事实分布，如下所示：

$$P(Y_{h+1}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h},b_i^Y)=P(Y_{h+1}|V,\overline{A}_{0:(h+1)}=\overline{a}_{0:(h+1)}(q),\overline{Y}_{0:h},b_i^Y).$$

然而，目标数　　量 $P(Y_{h+1}(q)|V,\ A_{0:h},\ Y_{0:\ h})$ 无法计算，因为

$$P(Y_{h+1}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h})=\int P(Y_{h+1}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h},b_i^Y)P(b_i^Y|V,\overline{A}_{0:h},\overline{Y}_{0:h})db_i^Y$$

而亚组异质性分布 $P(b_i^Y|V,\ A_{0:h},\ Y_{0:\ h})$ 未知。相比之下，我们的方案假设一种不同的条件序贯可交换性假设

$$Y_{h+1}(q)\perp A_{h+1}|V,\overline{A}_{0:h},\overline{Y}_{0:h},b_i^A.$$

根据不存在治疗分配异质性的假设，我们知道 $\text{var}(b_i^A)=0$，因此 $\text{cov}(b_i^A,\ b_i^Y)=0$，从而得到 $P(b_i^Y|V,\ A_{0:h},\ Y_{0:\ h},\ b_i^A)=P(b_i^Y|V,\ A_{0:h},\ Y_{0:\ h})$。因此,一可以通过式（6）识别目标量：

$$P(Y_{h+1}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h})=\int P(Y_{h+1}(q)|V,\overline{A}_{0:h},\overline{Y}_{0:h},b_i^Y)P(b_i^Y|V,\overline{A}_{0:h},\overline{Y}_{0:h})db_i^Y.$$

我们的方案可视为对 Shardell 和 Ferrucci（Shardell & Ferrucci, 2018年）工作的以下扩展：（1）对条件序贯可交换性采用较宽松的假设，按 b_i^A 分层，而不是按结局、混杂因素和治疗模型间共享的随机效应分层；（2）将模型设定为多元广义线性混合模型（MGLMM），其更具一般性，并有望将他们的联合混合效应模型纳为特例；（3）在假设不存在治疗分配异质性的情况下，可识别亚组因果效应。这项扩展的优点伴随着引入 b_i^A 的方差作为敏感性参数这一代价；在存在治疗分配异质性的情况下识别亚组因果效应，则需要对 b_i^A 的亚组分布作出额外假设，这可能需要专家知识。